Industrial-scale, genomics-based drug design and discovery

Philip M. Dean, Edward D. Zanders and David S. Bailey

The demands on drug discovery organizations have increased dramatically in recent years, partly because of the need to identify novel targets that are both relevant to disease and chemically tractable. This is leading to an industrial approach to traditional biology and chemistry, inspired in part by the revolution in genomics. The purpose of this article is to highlight the flow of investigation from gene sequence of potential therapeutic targets, through mRNA and protein expression, to protein structure and drug design. To deal with this scale of activity, many commercial and public organizations have been established and some of the key players will be listed in this article.

Philip M. Dean*, Edward D. Zanders and David S. Bailey
De Novo Pharmaceuticals Ltd, St Andrew's House, 59 St Andrew's Street, Cambridge, UK CB2 3DD.
*e-mail: philip.dean@denovopharma.com

The extraordinary pace and scale of developments in the field of genomics has forced a paradigm shift in the manner with which the pharmaceutical industry approaches the discovery and development of new drug compounds. The completion of the first draft of the human genome has made it possible to foresee major steps forward in our understanding of the molecular basis of disease, both from attack by external pathogens and internally from variations within the human genome resulting in a plethora of new molecular therapeutic targets for drug design and discovery. New technologies have sprung up to cope with the avalanche of genomic data. As a consequence of these exciting developments, the research process of drug discovery is becoming industrialised with the research impetus being taken out of academic institutes and put into cutting-edge niche research companies that have better funding and significant commercial opportunities.

It is salutary to look at the numerical scale of the problems that face the pharmaceutical industry. The human genome probably contains 35 000–50 000 genes[1]. Currently, there are probably 100 bacterial genomes sequenced each containing ~4000 genes. Three million single nucleotide polymorphisms (SNPs) are expected within the human genome and many might be linked directly with disease conditions or affect the pharmacokinetic profiles of drug treatment. Within each cell type in the human body, perhaps 100 genes show significant differential expression during drug treatment. It is estimated that the number of therapeutic molecular targets will increase from ~1000 currently used by the pharmaceutical industry to perhaps as many as 10 000.

A pharmacophore is defined as a subset of ligand atoms (such as hydrogen-bonding atoms, charged atoms and hydrophobic residues) that are principally involved in binding a ligand to a target. Within a potential drug-binding site, the number of possible pharmacophores can be expressed combinatorially; if there are 30 site points and a selection of any five site points form a pharmacophore then the number of subsets of pharmacophores is 140 000 per site. It has been estimated that the number of possible molecules with a molecular weight of less than 500 Da is 10^{200} of which perhaps 10^{60} might possess drug-like properties[2]. Thus, the pharmaceutical industry is presented with a universe of opportunities from which to pick commercial winners. Major numerical optimization problems have to be resolved if the industry is to efficiently explore the drug discovery process.

'It is estimated that the number of therapeutic molecular targets will increase from ~1000 currently used by the pharmaceutical industry to perhaps as many as 10 000.'

This article pinpoints the major components of this new wave of drug discovery and looks forward to the benefits of current progress.

Genomics and target identification

Genomics is the study of linear gene sequences and has been used to estimate the number of different mRNA species that could be expressed from the genomes of living organisms. It has also prompted a more global approach to biological problems, superseding the 'one gene at a time' approach that cannot reveal the full diversity and complexity of gene expression.

The field has been driven by the development of high-throughput DNA sequence analysis (and increased computing power) resulting in a proliferation of public and private databases containing full and partial sequences of mammals, plants and microrganisms. The ultimate objective of any gene-sequence database is to be able to store and retrieve information on full-length genes (and by implication, full-length proteins and regulatory regions for transcription factor binding etc.) but this has been a slow process. Rapid improvements in sequencing technology, and subsequently reduction in costs and increased throughput, has resulted in the expansion of public databases and the emergence of private databases for commercial purposes. The impetus provided by the latter has improved efficiencies in the overall strategic approach as well as invigorated the public genomics enterprises. Since

 0167-7799/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0167-7799(01)01696-1

1992, the ability to sequence only cDNA copies of mRNA (expressed sequence tags, ESTs) as opposed to waiting for the entire genomic sequence of an organism, has led to the creation of thousands of gene fragments for use with microarray and related technologies. In addition, many new genes have been discovered on the basis of the biological activity of their encoded proteins. For some organisms, including humans, the peak of EST sequencing has been reached and the way ahead is clear for revisiting the original objective of creating a catalogue of full-length genes. This is also the time for the companies that sell databases of genomic information to adapt to the necessity for adding significant extra value to sequence information, for example through detailed annotation, expression analysis and proteomics.

The recent landmark publications on the human genome sequence suggest that the total number of human genes is likely to be in the region of 35 000 (Ref. 3). The expected diversity of gene expression is therefore likely to derive from alternative splicing of the mRNA; this phenomenon is well known to create functionally different proteins from a single gene, one such example being the dopamine receptor[4].

Several companies have recognised the importance of this alternative splicing issue in biological and pharmaceutical research and offer the ability to search for known and putative splice variants within their databases of human genes.

Genetics and pharmacogenetics

With the recognition that many diseases involve an interplay between multiple genes and environmental factors, genetic association studies at high throughput have been required to tease out these complex traits. Luckily, the discovery SNPs, which are distributed throughout the human genome, offers a way forward. Furthermore, SNPs allow the application of genetics to efficacy of drug action at the level of the individual patient, thus revitalising the field of pharmacogenetics. Thus, there is a considerable effort being undertaken in big pharmaceutical and biotechnology companies to locate a sufficiently high density of SNPs along the human genome to map disease-related genes in a matter of months rather than years. It would be naïve to expect that every disease locus will provide a drug target in its own right, but these studies will provide some insight into complex pathological processes that are still poorly understood for major diseases such as asthma.

There is considerable interest in pharmacogenetics, that is the concept of 'the right drug for the right patient' using genetics to identify responders or non-responders to a particular medicine[5]. SNPs are being used to classify patients whose target protein for their prescribed medicine shows sequence variation or differences in expression level. The results of these trials will have obvious implications for clinical practice, as well as future scientific and commercial aspects of drug discovery.

Although some private companies have been SNP mapping on an individual basis, substantial resources are required to do so. This consideration has encouraged a group of large pharmaceutical companies to acquire and share data as a consortium to contain overall costs. Thus, the most significant genetic data on major Western diseases will probably emerge from the SNP consortium working in close association with academic laboratories over the next few years.

Gene expression and array technology

Early attempts to systematically catalogue the full complement of mRNA molecules expressed by a cell or tissue, relied on large-scale sequencing of cDNA libraries. The relative abundance of a particular mRNA is related to the proportion of sequences obtained[6]. This approach has been used successfully by Incyte (Palo Alto, CA, USA) to produce electronic northern blots of cells and tissues, and by public resources such as the IMAGE consortium. Improvements in differential gene expression technology based on PCR [differential display, serial analysis of gene expression (SAGE)] or array hybridization (such as nylon, glass microarrays and oligonucleotide chips) have changed the landscape completely[7]. Few aspects of biology have been untouched by the application of array technologies in one form or another. Several themes have emerged from published reports over the past few years that involve the expression analysis of mRNAs from yeast and humans in particular. In broad terms, these themes consist of evaluations of gene expression changes in systems undergoing biological responses, for example during the cell cycle[8]. In addition, statistical analyses of gene expression in tissue samples have been employed in the search for pathological markers of disease[9]. All these studies find a place in drug discovery, both as an aid to target identification and also for evaluating the detailed efficacy and side-effect profiles of experimental compounds, as well as the identification of surrogate markers of drug action.

Now that gene expression technology and data analysis methodology has matured, there has been a call for data sharing and standardisation. These are clearly difficult issues, with the need to accommodate different experimental platforms and analytical systems, but the first step forward has been taken with the establishment of the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress). As the input of quality data into these resources becomes significant, it will be easier to compare gene expression under widely differing conditions *in vitro* and *in vivo*, thus establishing the rules under which living systems operate at the transcriptional level.

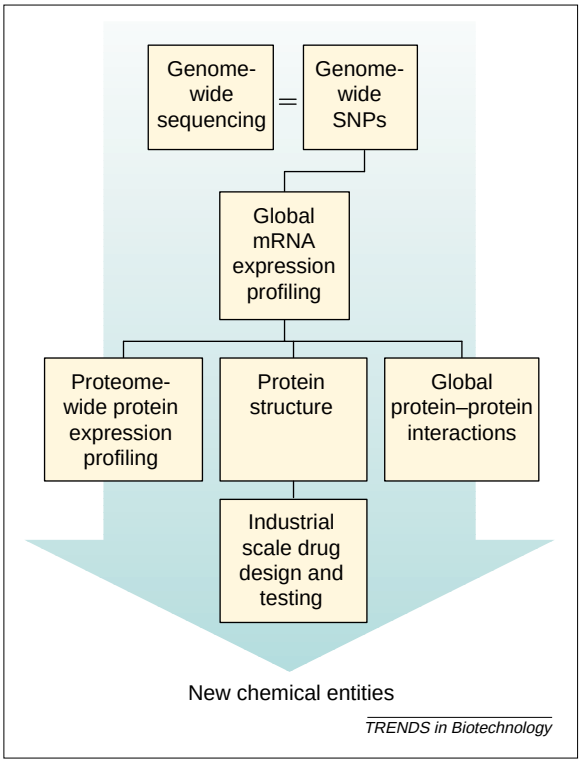

Fig. 1. The biotechnology-based drug discovery assembly line. The figure shows the components of the genomics-based drug discovery process and associated organisations. Initial nucleotide sequence analysis provides the basis of gene identification and genetic analysis using single nucleotide polymorphisms (SNPs). This is followed by mRNA and protein expression analyses to determine the cell and tissue distribution of the gene of interest in health and disease. The structure of the protein of interest is determined, to provide a framework for small molecule discovery through *in silico* design. In addition, the association of the target with other proteins in functional complexes might add value to the analysis.

Proteomics – analysing the protein complement of the genome

High-throughput analysis of protein sequences and structures has lagged behind that of gene sequences because of the greater technical challenges involved. This information, however, is fundamentally important for drug discovery because proteins are the physical targets of most drugs (small molecules or proteins). Not surprisingly, there is substantial effort and progress in this area, particularly after the introduction of specialised mass spectrometry techniques to analyse small amounts of protein with high precision using information from sequence databases to aid in their identification. Initial applications were concerned with the identification of proteins that had been resolved using one or two-dimensional gel electrophoresis. This has proven highly successful in the characterisation of members of protein complexes (for example from signalling complexes or organelles) as a complementary approach to the yeast two-hybrid system. (In the latter case, it is possible to identify interactions between individual pairs of proteins using genetic methods.) Matthias Mann, one of the pioneers of the application of mass spectrometry to proteomics, has now commercialized his research to form Protana (now acquired by MDS Proteomics, Calgary, Canada) to scale up this technology.

An exciting further application of mass spectrometry is the detection of low molecular weight compounds that are bound to proteins by affinity adsorption. This could dramatically reduce the time required to identify and characterize lead compounds for drug discovery. This is also possible with a logical extension of cDNA microarrays, in which proteins are immobilized at high density on solid supports. In this case, these targets might be interrogated using other proteins or small molecules in a screening format. The standard fluors used for nucleic acid arrays have been employed to measure protein–protein and protein–drug interactions on glass slides thus avoiding the use of expensive mass spectrometers[10].

These novel assay systems will begin to take over from high-throughput screening (HTS) as a primary method for identifying lead compounds. A more efficient process than either method would be to directly identify potential leads *in silico* by exploiting the 3D structures of proteins and advances in computer-aided drug design. This requirement is, in part, currently driving the intense activity in structural genomics and bioinformatics; made possible both through the increased number of 3D protein structures being solved using X-ray crystallography and NMR, or being inferred through homology modelling. Recently, several new companies have been formed to generate revenues from partnerships with drug discovery organisations. A structural genomics consortium (SGC) is planned for late 2001 by major pharmaceutical companies for the same reasons that the SNP consortium was founded – that is to benefit from sharing information without bearing a disproportionate amount of cost. Likewise, major national efforts are underway to achieve the same result in recognition of the fundamental importance of protein structure in exploiting the genome for drug discovery. Figure 1 summarizes the flow of activity from sequence identification to exploiting the proteome to provide targets for drug discovery. The main companies involved in these processes are listed in Box 1.

Small molecule discovery

The modern pharmaceutical industry was founded on the manufacture of low molecular weight compounds (<500 Da) for clinical use. Despite the successful introduction of protein therapeutics and the promise of gene therapy, the discovery of small molecule drugs remains the dominant activity of the industry. As the following sections will make clear, this activity is no longer confined to traditional large pharmaceutical companies, it is also being undertaken by biotechnology companies that are using a new generation of discovery tools based on computer-aided design and combinatorial chemistry. Many reviews of genomics and proteomics stop short of discussing how target information can be efficiently converted to small molecule drug candidates.

The most commonly used method of identifying a lead molecule for a protein target is HTS of compound databases. This promised much for delivering hits from a company's own resources, with numerous contract screening organizations being established to provide this service. Two types of

Box 1. Companies involved in genomics-based drug design and discovery

Genome wide sequencing
Celera Genomics (Rockville, MD, USA); http://www.celera.com
European Bioinformatics Institute (EBI, Hinxton, UK); http://www.ebi.ac.uk
Human Genome Sciences (Rockville, MD, USA); http://www.hgsi.com
Incyte Genomics (Palo Alto, CA, USA); http://www.incyte.com
National Center for Biotechnology information (NCBI, Bethesda, MD, USA); http://www.ncbi.nlm.nih.gov

Genetics and SNP analysis
DeCode Genetics (Reykjavic, Iceland); http://www.decode.com/
Genset (Paris, France); http://www.genxy.com
SNP consortium; http://snp.well.ox.ac.uk
Genaissance Pharmaceuticals (New Haven, CT, USA); http://www.genaissance.com

mRNA expression profiling and altenative splicing
Compugen (Jamesburg, NJ, USA); http://www.cgen.com/science/splicing.htm
I.M.A.G.E. consortium (Livermore, CA, USA); http://image.llnl.gov
Incyte Genomics (Palo Alto, CA, USA); http://www.incyte.com
Rosetta Inpharmatics (Kirkland, WA, USA); http://www.rii.com
Affymetrix (Santa Clara, CA, USA); http://www.affymetrix.com
GeneLogic (Rockville, MD, USA); http://www.genelogic.com

Proteome-wide expression profiling
OGS (Oxford, UK); http://www.ogs.com/AboutOGS/auoverview
Swiss 2D-PAGE; http://www.expasy.ch/ch2d
Ciphergen Biosystems (Freemont, CA, USA); http://www.ciphergen.com
LSB (Vacaville, USA); http://www.lsbc.com

Protein structure
Astex (Cambridge, UK); http://www.astex-technology.com
Inpharmatica (London, UK); http://www.inpharmatica.com
Integrative Proteomics (Toronto, Canada); http://www.integrativeproteomics.com
NIGMS Structural genomics initiative (Bethesda, MD, USA); http://www.nih.gov/nigms/fundingpsi.html
Protein Data Bank (PDB); http://www.rcsb.org/pdb
Protein Structure Factory (Berlin, Germany); http://userpage.chemie.fu-berlin.de/~psf
Prospect Genomics Inc. (San Francisco, CA, USA); http://www.prospectgenomics.com
RIKEN (Saitama, Japan); http://www.riken.go.jp
Structural Bioinformatics (San Diego, CA, USA); http://www.strubix.com
Structural Genomix (San Diego, CA, USA); http://www.stromix.com
Syrrx (San Diego, CA, USA); http://www.syrrx.com

Global protein–protein interactions
Caprion Proteomics (Montreal, Canada); http://www.caprion.com
Curagen (New Haven, CT, USA); http://www.curagen.com
Hybrigenics (Paris, France); http://www.hybrigenics.com
MDS Proteomics (Calgary, Canada); http://www.mdsintl.com

Industrial scale drug design and testing
De Novo Pharmaceuticals (Cambridge, UK); http:/www.denovopharma.com
Prospect Genomics Inc. (San Francisco, CA, USA); http://www.prospectgenomics.com
CEREP (Rueil-Malmaison, France); http://www.cerep.fr/Cerep
Evotec (Hamburg, Germay); http://www.evotec.de
Neogenesis (Cambridge, MA, USA); http://www.neogenesis.com
Protherics (Macclesfield, UK); http://www.protherics.com

compound database have been used, large company compound collections, and optimally diverse subsets of the compounds. However, in general, the results obtained using HTS have not been as attractive as hoped[11] for two reasons: (1) the numbers of compounds that can be economically screened is small compared with the chemical space available, and (2) the theoretical coverage of molecular diversity within the screening set is limited. Thus, there might be no compound in the collection that has an appropriate pharmacophore for the target binding-site. Furthermore, the hits obtained might not be amenable for further elaboration into lead compounds through medicinal chemistry.

There will, however, be several cases in which HTS will provide a number of hits that can then be analysed by molecular similarity studies. These identify the different pharmacophore subsets present in the screening data and provide clues for different possible lead series. For these molecular similarity approaches to be useful, they must be able to identify partial molecular similarity within a set of diverse hits, something that is not always obvious on visual inspection of the structures. The SLATE algorithm for example[12], is capable of extracting this information from structurally distinct molecules that bind to the histamine H3 receptor allowing the design of a novel chemical series with potent activity at H3.

Virtual screening

As implied previously, *in silico* screening is an important complementary technology to HTS using virtual compounds rather than in-house collections. The technique requires a 3D molecular structure of the target molecule, either determined using X-ray crystallography and/or NMR, or by homology modelling. The molecular structures to be screened might be an existing compound collection or a virtual collection of molecular structures obtained from a preferred set of combinatorial chemistries. The advantage of the virtual set is that it is not necessary for the structures to have been synthesised before the *in silico* docking experiments. *In silico* screening relies on the generation of a representation of the ligand binding site on a protein that can be used for docking, the aim is to identify molecules that have

the correct geometric and electronic features to fit the designated sites. As might be expected, the development of computer algorithms that have these capabilities is a large field of study in itself and has been ongoing for many years.

'...industrialisation of the gene-to-drug discovery process is essential...'

Although *in silico* virtual screening has been successful, there is always room for improvement in identifying structures that have a high chance of being pharmacologically active and these issues are discussed below. First, the rapid versions of *in silico* screening that employ flexible docking only deal with flexible compounds and ignore the computationally demanding problem of including flexibility within the sites. Recently, Carlson and McCammon[13] reviewed these problems of flexible docking and speculated that loop fluctuation and domain movement models need to be developed to tackle flexibility within sites. Second, the generic weakness of current docking algorithms is that they require an accurate scoring function to prioritise the virtual hits. Scoring function technology is not particularly advanced and does not usually take into account explicit water molecules in the site. However, there are some interesting successes; a notable one being the discovery that some steroids bind to FK506 binding protein (FKBP; Ref.14).

Further evidence of activity in the field of virtual docking is provided by the large-scale docking collaboration (Dockcrunch, Macclesfield, UK) that has been established to combine novel software with special hardware systems (http://www.protherics.com/crunch) to overcome some of the problems highlighted above.

Drug design

Whereas virtual screening is a powerful adjunct to real HTS, true *de novo* design of drug candidates is in a class of its own with promises and challenges to match. A new generation of *de novo* design algorithms can produce large numbers of molecular structural types for a particular protein site. Furthermore, the structures can be provided in combinatorial chemistry formats and tailored to the preferences of client companies by enabling synthesis to be performed in-house using proprietary chemistries. The algorithms work by exhaustively building virtual structures in the site and can be geared to create novel diverse scaffolds that can be subsequently 'decorated' with substituents as a combinatorial library. The use of sophisticated optimisation algorithms ensures that large chemical spaces can be explored, thus increasing by many orders of magnitude the numbers of structures that can be considered for design.

Coupling these algorithms to large-scale structural genomics and high-throughput chemistry will provide novel and patentable compounds within a fast timeframe. This industrialisation of the gene-to-drug discovery process is essential if the plethora of new targets spawned from genomics is to be capitalised upon. Because of the sheer number of targets, it will be necessary to identify families of functional protein classes (e.g. hydrolases and kinases) as targets for drug design. This has the advantage of creating the chemical strategies that are appropriate for a particular family and also helps in the all-important issue of selectivity. Large pharmaceutical companies have not traditionally invested as much in this area as in HTS and therefore it is likely that niche companies capable of synergizing with complementary companies will be the main beneficiaries of this approach to drug discovery. Exploitation of data from the public and private domains has been a significant catalyst in this commercialisation process, despite the patenting issues on privately generated gene and protein sequences. This new biotechnology-driven aspect of the drug discovery process will benefit from organisational strategies that are commonplace in other high technology industries such as semiconductors. Thus, the implementation of an assembly line process via the seamless integration of target discovery, validation and drug design modules and the use of automated procedures, will result in dramatic increases in efficiency to the benefit of all concerned.

References

1 Lawrence, R. (2001) Craig Venter discusses life after the Human Genome Project. *Drug Discov. Today* 6, 10–12

2 Bohacek, R.S. *et al.* (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* 16, 3–50

3 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 862–921

4 Usiello, A. *et al.* (2000) Distinct functions of the two isoforms of dopamine D2 receptors. *Nature* 408, 199–203

5 Drazen, J.M. *et al.* (1999) Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. *Nat. Genet.* 22, 168–170

6 Audic, S. and Claverie, J.M. (1997) The significance of digital gene expression profiles. *Genome Res.* 7, 986–995

7 Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature* 40, 827–836

8 Spellman, P.T. *et al.* (1998) A comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridisation. *Mol. Biol. Cell* 9, 3273–3297

9 Golub, T.R. *et al.* (1999) Molecular classification of cancer; class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537

10 MacBeath, G. and Schreiber, S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science* 289, 1760–1763

11 Bailey, D.S. and Brown, D. (2001) High-throughput chemistry and structure-based design: survival of the smartest. *Drug Discov. Today* 6, 57–59

12 Mills, J.E.J. *et al.* (2001) SLATE: a method for the superposition of flexible ligands. *J. Comput. Mol. Design* 15, 81–96

13 Carlson, H.A. and McCammon, J.A. (2000) Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.* 57, 213–218

14 Burkhard, P. *et al.* (1999) The discovery of steroids and other novel FKBP inhibitors using a molecular docking program. *J. Mol. Biol.* 16, 853–858