



## **The drug discovery process – space and time in the pharmaceutical universe**

**Nick Miller (Beremans Limited) and Mohammad Afshar  
(Ariana Pharmaceuticals)**

### ***The three degrees***

Basically, drug discovery involves throwing chemicals at a target and seeing which ones stick. On this level, there are three degrees of freedom in the drug discovery process, namely, how good are your targets (are you aiming at the right proteins), how good are your chemical compounds (how diverse, how drug-like), and how well can you throw (how many compounds can you test in unit time and how accurate are your measurements). In this article we present a broad overview of pharmaceutical drug discovery, and outline some of the main types of technologies associated with the above three elements of the process. Limitations of space necessitate a shallow treatment of these topics; however, future articles will examine particular aspects of the process in more detail.

### ***Biological space – targets for drugs***

The first requirement in conventional drug discovery is identification of a valid target, ie a molecule which has a link with the disease of interest such that pharmacological intervention would be expected to cure the disease or ameliorate its symptoms. There are two aspects to this process, ie identification of potential targets, and their subsequent validation as actual targets.

Identification of potential targets requires exploration of biological space, and is exemplified by *grands projets* such as the sequencing of the human genome, with their reliance on high throughput sequencing technologies and informatics algorithms to handle and analyse the large volumes of data being generated. The data from such projects is useful as a starting point; for example, the knowledge that the human genome contains about 30,000 genes provides a framework within which drug discovery must operate. Similarly, the use of technologies such as gene chips to determine the level of expression of genes in particular diseases may suggest a subset of proteins to be of particular interest. The identification of novel targets in this way is relatively straightforward; however the validation of a target as a suitable point for pharmacological intervention in a given disease is more difficult. Furthermore, validation is critical to successful drug discovery; if an inappropriate target is picked at the start of the process, all subsequent efforts are wasted. (The definition of 'validated' is somewhat contentious; here we use it to refer to instances in which manipulation of a target results in dose-dependent changes that are (i) consistent with an expected therapeutic effect and/or with the hypothesis that the target is involved in a biological pathway linked to the disease state; and (ii) inducible in an animal model).

Target validation may be effected by various more-or-less labour-intensive methods, which usually are directed at experimental modulation of gene activity. Technologies commonly employed include methods to knock out or knock down the product of a gene, or to induce gene expression. Gene product knock-down

methods include RNAi (see previous Beremans article), antisense, inducible gene knockouts in cells or in transgenic animals (where an external stimulus can be applied to nullify expression of a particular gene at a particular point in time), and intrabodies (antibodies designed to knock-down intracellular target proteins). Methods to upregulate or confer expression of a gene include various gene transfer and inducible expression technologies in cells and transgenic animals.

Both the target identification and target validation phases of the process generate large quantities of data presented in different formats. Capture and storage of this data, and subsequent manipulation and analysis of the data to extract complex relationships between data points, requires sophisticated database and informatics technologies. Indeed, informatics technology is almost ubiquitous in modern drug discovery.

### ***Chemical space – drugs for targets***

Having identified and validated a biological target, the next step is to identify an entity which can specifically interact with that target in such a way as to produce a therapeutic effect. In classical drug discovery the 'entity' is a small molecule chemical compound. Identification of a chemical which specifically binds to the right part of a target protein is not trivial. The chances of success are increased by exploring the chemical space as fully as possible. The total theoretical chemical space, ie the number of different compounds that could in theory be synthesised, is said to be in the region of  $10^{40}$ . This is a number so large that it is not readily comprehensible; in the state of our current capabilities, it might as well be infinity. Although there are efforts to make very large libraries, complete coverage of chemical space appears to be beyond current capabilities, and it is unusual for any given pharmaceutical group to have a 'library' of more than one or two million. Even completely 'virtual' libraries (existing only as data in a computer programme) do not span the chemical universe. Consequently, the focus has shifted from the absolute size of the chemical library to library quality.

The definition of 'quality' depends on the precise object that the library is intended to achieve. Large libraries of random, unrelated molecules are intended to cover as much chemical space as possible, and in this case a measure of library quality is the diversity of structures contained within it. Obviously, one has to bear in mind that even one million diverse compounds ( $10^6$ ) would cover only a tiny fraction of the entire space. So, in practice, pure diversity-based approaches are often limited to exploring space around known active compounds. Traditionally, screening libraries have been the major component of the war chest of big pharmaceutical companies, and are heavily biased towards their respective histories in pursuing discovery projects against particular target/disease areas. Thus, successive lead optimisation campaigns over tens of years often have led to an over-representation in the library of certain types of "scaffolds" or shapes, such as GPCR-I modulators, nucleoside mimetics, etc.

Hence, using cheminformatics methods, the industry is moving towards identifying gaps (areas where chemical space is under-represented) in compound collections, and plugging these gaps either through acquisition of compounds from third parties or through internal synthesis. The advent of a large number of chemistry companies that provide drug-like molecules as a commodity has opened up this traditional big pharma approach. Thus biotech companies now can assemble, at reasonable cost, screening sets that are as rich as those of any pharmaceutical company.

Indeed, library design may have a rational element to it in that the compounds may be based around particular designs that are thought most likely to interact

with a target, or most likely to have favourable drug-like characteristics (eg high solubility, low toxicity). Similarly, the libraries may be 'lead-based', in that their constituents are consistent with the general properties of existing safe drugs. In these cases the notion of quality is determined by the extent to which library contents are biased towards those types of compound that are *a priori* thought to have more chance of producing a useful drug.

As with biological space, informatics and associated software has a critical role to play in the exploration of chemical space. This *in silico* exploration seeks to identify an acceptable compromise between the diversity of the collection, its chemical tractability (ie if the compound is identified as a hit, its chemical modification would be straightforward), and its *a priori* adverse properties such as low solubility, poor predicted absorption or high risk of toxicity issues.

### ***Space/time is money – fitting chemical space to biological space, at speed***

After identification of a valid target and construction of chemical libraries of appropriate quality, the chemical compounds must be screened against the target to identify promising interactions. There are two elements to this 'primary screening' stage of the process, namely the rate at which compounds can be tested (the throughput) and the method for identification of an appropriate interaction (the assay).

Throughput often has been seen as a critical bottleneck in drug development, and usually has been addressed by miniaturisation technologies. Miniaturisation both enables more compounds to be tested per unit space/time, and also decreases the cost per compound tested, eg due to the use of smaller volumes of expensive reagents. Hence various high throughput screening (HTS) systems have been developed which allow testing of many compounds in parallel through reproducible manipulation of small volumes of liquid.

A common means of achieving HTS involves the use of high density plates for sample analysis. HTS normally uses plastic plates containing 96 or 384 wells, each well containing a different small molecule compound. The sample plate has been pushed to ever-denser formats, with ultra-HTS plates containing 1536 or 3456 wells, allowing testing of 100k samples in 24 h. Some reports suggest that microplates with 20,000 wells may be feasible. Similarly, labchips contain arrayed networks of channels for the manipulation of liquids and parallel assay of reactions in submicrolitre volumes.

The use of smaller volumes does however bring its own problems; the sample errors produced by evaporation and reagent addition mechanisms become increasingly large as the absolute volume is reduced. Hence the miniaturisation of sample volumes has had to proceed hand-in-hand with advances in small volume liquid handling robotics. Many of these systems and other HTS innovations, such as barcoding of samples, have built on robotic instrumentation and procedures taken from the world of industrial manufacture. Co-ordination and control of the various robotic functions and collection of assay data are performed by computers. Again, informatics is essential for the analysis of the data that would lead to the identification of "hits". This can be quite daunting as the "signal" has to be extracted from a noisy background (ie at some high concentration, almost all molecules are seen as active, and the key is to find an appropriate threshold).

The assay itself should be as far as possible a surrogate for clinical effect, and therefore should measure clinically relevant aspects of target function in an

environment that is as close as possible to the clinical situation. If the assay measures irrelevant aspects of the interaction between chemical and target, significant resources may be wasted in taking doomed compounds through later, more expensive phases of the process. However, the assay also needs to be simple, reproducible and suitable for automation, in order to keep costs down and to allow high-throughput testing. For example, testing the activity of an isolated enzyme by spectrophotometric methods is a robust and cheap assay that satisfies the above requirements.

Unfortunately, many targets require the presence of multiple factors to allow relevant assays to be performed, and hence may require testing in whole, living cells. Whole cell assays, also known as 'high-content screens' or HCS, usually take place in 96 well plates using HT microscopy to analyse, for example, morphology and intracellular distribution of fluorescence over time. Such assays may require multiple steps (eg of addition of compounds, washing the cells, addition of fresh culture medium, etc) over extended periods of time. Thus there is a trade-off between the number of clinically relevant parameters that can be included in a primary screen and the cost/throughput of that screen. Therefore pragmatic considerations may force the screens to be done on isolated proteins rather than on proteins expressed in a given cell type.

HTS can execute about 10,000 assays per day. If the assay is kinetic, in which a measurement is tracked over time, the HTS output can produce millions of data points per day. High content screening with image capture also generates data rich files. Large pharmaceutical companies may undertake 35 screening campaigns against 100-500k compounds per year giving up to 18 million primary screening data points; this is predicted to increase to 100 campaigns using 1.5 million compounds. This situation, together with the issues relating to the size of combichem libraries (which may contain millions of compounds), leads to particular problems of data capture, storage and analysis.

Hence the discipline of informatics is now fundamental to drug discovery. For example, the compounds that are put into the screen may be dictated by *in silico* methodologies; eg docking algorithms allow pre-screening of libraries and identification of those elements most likely to bind a given target. Such *in silico* pre-screening activities allow one to limit the numbers of molecules put through the real screening process, with the result that screening results will be rich in 'quality' data that provide positive guidance for the next steps in the process. In addition the various data types produced by primary assays, eg images of cells, gene expression data, colorimetric data, all must be captured and related to the targets and chemical structures being examined, and correlations extracted.

### ***Space and time in drug discovery***

The primary screening stage referred to above will identify a number of compounds which appear to affect a target in a drug-like way. These compounds are known as hits, and are subjected to further screening and analysis steps. In particular, the chemical structures of the hits are examined, and they are prioritised in terms of their likely drug-like nature. Initial structure-activity relationships (SAR) are described, for example correlations between activity and the presence of particular chemical motifs. Identified correlations may be tested by the study of related compounds. Preliminary studies are carried out to determine parameters such as potency, specificity and toxicity, and preliminary ADME (absorption of the drug from the blood, distribution of the drug within the body, metabolism of the drug over time, and excretion of the drug in urine/faeces) tests are carried out. The best compounds identified by this evaluation process are known as leads.

Even with good secondary screens, a HTS programme may produce hundreds of leads. The cost of maintaining sufficient numbers of rodents for screening of all such leads is significant, hence efforts to use smaller and less demanding animal models. Fruitflies, nematodes and zebrafish embryos all have been used in this context, with fish obviously being the most physiologically relevant. (There are some claims that 50k or more fish embryo screens can be carried out per day, putting the zebrafish model into the primary screening category).

Lead compounds are passed to medicinal chemists for iterative rounds of optimisation, in which undesirable characteristics are removed and desirable ones added, eg to improve solubility. Thorough screens of potency, toxicity, ADME, etc, and analysis of effect in animal models of disease, may follow. Compound pharmacodynamics (the time course of a biological response to a compound in different tissues), administration route and potential drug interactions also may be assessed at this point. The mechanism of action of the drug may be established and quantitative SAR (QSAR; which is intended to predict pharmacokinetic and pharmacodynamic features from structural/chemical features) and pharmacophore modelling (ie computer-assisted determination of the combination of functional groups and hydrogen bond donors and acceptors that is required for activity) undertaken. Data obtained from this phase of the process are likely to be used in support of any future applications to carry out clinical trials of the lead compounds.

The accumulated expertise and tacit knowledge of medicinal chemists is likely to be very important at this stage, for example to remove molecular domains that have toxic or oncogenic potential. In addition, computational techniques are of increasing importance for rational design of small molecules (eg by generation of *in silico* libraries of molecules and *in silico* screens to detect probable interactions with the target), and for the identification of pharmacophores and QSAR models.

It has been said that less than 25% of lead compounds that enter clinical trials ever reach the market, and of those that do reach the market, few are blockbusters. This high failure rate is at least partly due to suboptimal ADME or toxicity characteristics becoming apparent in the clinic. In addition, many hits are screened out at the preclinical stage due to undesirable ADME/Tox characteristics, and these also represent substantial wasted investment and opportunity cost.

This situation has led to pressure to push the ADME/Tox testing to the first stages of the discovery process, and indeed the throughput of many ADME/Tox assays has been substantially increased, allowing them to be used earlier. A complementary approach is the use of computational tools to generate predictive models of the likely characteristics of a given set of small molecules. Hence the field of 'chemogenomics', where chemical space meets biological target space *in silico*. This allows virtual screening of virtual libraries, with the result that the libraries can be stratified into subsets of molecules with particular predicted characteristics of solubility, ADME, toxicity and so forth. (However, accurate *in silico* modelling of ADME has so far been difficult due to the lack of sufficiently large training datasets of sufficiently high quality). In fact, these types of computational methods can be used at various stages in the drug discovery process from *in silico* screening of compound libraries to clinical development.

There is also some effort directed at the early identification of metabolites of candidate drugs, eg to draw early attention to toxic or otherwise active metabolites, and to predict the effect of inhibition or induction of metabolising enzymes. Such information could support lead optimisation by suggesting the need to enhance or block features of the metabolism of the drug candidate. In

addition, the use of high content primary screens where multiple parameters are measured in whole cells may enable some parameters such as toxicity to become apparent earlier in the process.

### ***Time travel – future trends in drug discovery***

Drug discovery trends have, fundamentally, been aimed at increasing the number of drugs discovered per unit time while containing the associated costs. Hence there has been huge investment in the 'space and time' aspects of drug discovery alluded to above, with combinatorial chemistry, genomics technologies, proteomics technologies, target validation methodologies, miniaturisation and high throughput screening technologies all having been claimed at one point or another as heralding a step change in drug discovery. Whether they have created step changes or not is debatable; what is certain is that they have generated huge and increasing amounts of data. Further '-omics' and other techy catchphrases no doubt will be coined, and indeed those who follow the field may have noticed a recent vogue for applying high throughput technologies to drug metabolite identification. To some extent, any over-emphasis of a particular data generation technology with a narrow field of application in drug discovery is missing the point, which is that data is of limited value until it has been turned into information.

In brief, and at the risk of gross oversimplification, future trends are likely to be (i) to push critical screening steps to as early in the drug discovery process as possible ('fast to fail'), which may mean carrying out some steps of the drug discovery process in parallel rather than in series; and (ii) to enhance the quality of HTS, by improving the quality of the compound collections and the quality of the assays, as well as the quality of the data analysis methods. Underpinning both of these trends is, and will remain, a continuing expansion in the use of sophisticated *in silico* technologies. One of the most important aspects of this trend probably will be the development of multiparametric analysis software that, for example, allows many assay endpoints of small molecules (such as ADME) to be predicted, so as to allow development of a balanced portfolio of characteristics, rather than the emphasis of one feature at a time. It seems likely that it will not be the method of getting data that is a source of competitive advantage (no doubt we will see a number of more or less indistinguishable HT metabolite identification technologies); rather, competitive advantage will stem from flexible software that can manage a large variety of data types, and analyse them to extract non-obvious correlations and predictive rules. We will return to this topic in our next article. Comments may be addressed to the author at [nm01@beremans.com](mailto:nm01@beremans.com).

Copyright Beremans Limited 2004