



## **Computational challenges for drug discovery: transforming data into knowledge**

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

The pharmaceutical industry's difficulty in finding novel drugs is well documented. Over the last decade, its response has been massive investment in high throughput methods (genomics, high throughput screening and combinatorial chemistry). However, efficient identification and optimisation of potent lead molecules is still the highest and riskiest hurdle in current drug discovery and development. The only clear outcome of high throughput methods has been an unparalleled production of large quantities of data. This in turn has shifted the focus to increased investment in the computational tools needed to analyse vast and disparate data collections.

In this article, we will be outlining the main sources of drug discovery data, the approaches currently being used to bridge the gap between data and knowledge, and the limitations of these approaches. In addition, we touch on some of the emerging computational technologies in this field.

### **Generating even more data**

Computational infrastructures must be designed to cope with two types of problem, namely the huge amounts of data that are being generated, and the great diversity of data types that must be captured, stored, annotated and analysed. Each of the various types of data handled in the drug discovery process poses its own specific challenges.

#### *Target identification - Microarrays*

Gene expression microarrays, or 'genechips', conventionally are composed of many short sequences of DNA immobilized in a regular fashion (arrayed) on a 2-dimensional surface. Each point in the array represents a unique spatial 'address'. By arraying many thousands of sequences, each specific to a different gene, one can two-dimensionally represent a large proportion of the genes of relevance to drug discovery. Such chips can be mass-produced. Their utility lies in the specific affinity of each DNA sequence for RNA with a complementary sequence. Since RNA is produced by transcription from active genes, one can use these chips to identify which genes in a cell are active, by extracting the cell's RNA and allowing it to react with the sequences on the chip. Thus chips can provide information on which genes are active in normal vs diseased cells, and thereby point out differences which may lead to identification of new drug targets. Equally, the chips may be used to identify tell-tale transcription signatures typical of, for example, drug toxicity, and thereby screen out toxic compounds at an early stage in the discovery process. Clearly the measuring of many thousands of gene expression profiles against many thousands of compounds will result in a significant data burden over time, particularly if there are kinetic (time-course) elements to the experiment.

#### *Combinatorial chemistry and library design*

The advent of new combinatorial chemistry techniques has dramatically increased the size of compound collections, the idea being that if one could make and test enough random molecules, one would find the right “hit”. Obviously, the key question here is how big is “enough” and current thinking is that a few million molecules is still short of covering an adequate “diversity space” (see previous TranScript article). Computational techniques are essential at this stage for ensuring adequate diversity in relatively focussed sets of compounds that are thought to have “drug-like” or “lead-like” properties.

#### *High throughput screening*

The ability to produce large numbers of compounds goes hand in hand with high throughput screening (HTS) capacity, ie the ability to test >500,000 compounds in an assay. If one disregards the actual technical difficulties of obtaining reliable data from single point measurements in HTS, one is still left with the daunting task of setting an arbitrary threshold and choosing the active molecules. Indeed, often, if one sorts the molecules according to their activity, the activity measurement decreases in a continuous manner, making it difficult to identify any “natural” break points. This means that very small differences in the threshold chosen (which can be set at a given activity limit or a more complicated “profile”) can have a huge impact on the compounds selected. Furthermore, the most useful compounds are not necessarily the most potent ones in HTS.

#### *Virtual screening*

Virtual screening avoids many of the shortcomings of traditional HTS, using an initial step of computational filtering that reduces the number of compounds that subsequently need to be tested experimentally. Although this approach can improve the reliability of the measurements, one is still confronted with the same issue of setting a threshold to select compounds that are tested experimentally.

#### *Lead optimisation methods and early ADME & T*

Once a “hit” has been identified, and confirmed as a genuine lead by further testing, variants of the lead molecule are designed, synthesized and tested in order to improve the potency and pharmacokinetic properties of the drug candidate, while reducing its toxicity. This process is known as lead optimisation. In a typical lead optimisation process, tens of assays are run in parallel to evaluate the potency of each candidate molecule, its specificity, its good Absorption and Distribution, good Metabolism and Excretion profiles and limited Toxicity (**ADMET**). Some assays are run only on a few compounds, while others are run for all the compound series involved in the lead optimisation. Numbers of compounds can rapidly get in the hundreds; in this multi-parametric space, identifying the next compound to be synthesised in order to get closer to a candidate molecule which would display the optimal combination of properties is a true challenge.

#### *Impact on drug discovery: decisions, decisions!*

All in all, far from making the drug discovery process a “no-brainer” activity, high throughput methods have made the process even more daunting where even simple manipulation and visualisation of the data requires sophisticated computational tools. Decision support has become a key concept in modern drug discovery.

#### **Tools for transforming data into knowledge**

A computational infrastructure including hardware, relational databases, etc, is required to store and handle the underlying data. In addition, integration issues, linking diverse sources of chemical, biological and virtual data can be challenging. We will not address

these issues in detail. Instead, we focus below on computational methods that rest on these platforms.

### *Predictive modelling*

Using experimental data (a 'learning set'), the goal of predictive modelling is to develop *in silico* models that would mimic as far possible an experimental response and that would, in time, allow one to use the model instead of performing a 'real' experiment. Ideally, the model would also explain the response, allowing one to design compounds with the desired outcome. An example of a predictive model would be one that predicts the activity of a compound against a particular target, given the chemical structure of the compound. Although few models approach the ideal of accurate prediction of parameters of interest, nevertheless some have been proven to be useful, at least as first approximations or filters prior to more extensive investigation.

### *Descriptors*

In order to allow the methods to predict the activity of molecules that are not included in the learning set, a number of "descriptors" are used to characterize each compound. These can be physico-chemical properties such as molecular weight, the vdW surface, number of hydrogen bond donors, etc. The presence or absence of particular fragments (such as benzene ring, primary amine, pyridazole, etc) can also be used; these profiles sometimes are called finger-prints. Many of the above descriptors are relatively straightforward; however, some are harder to interpret, for example the measure of the topological complexity of the bonded connections within a molecule.

Predictive methods try to identify the relevant descriptors as well as a relation that would link them to the observed property to be predicted. For example, one could calculate a descriptor that reflects the hydrophobicity for each molecule (such as Log P) and "learn" that highly hydrophobic compounds are unlikely to be absorbed. This rule could then be used to predict the absorption property of a compound that would be structurally different to the ones in the learning set.

The descriptor's relevance to the mechanism that is linked to the measured response, as well as its interpretability, are important measures of its quality (real-life utility) for a given problem.

### *The methods*

Broadly speaking, the methods can be divided into two families: numerical methods (statistical analysis, principal component analysis, neural networks, PLS, SVM, etc) and logical methods (such as Inductive Logic Programming). Hybrid methods, such as decision trees, combine some aspects of both approaches.

Numerical methods have been extensively used to identify Quantitative Structure Activity Relationships (QSAR). These methods can handle very large datasets and can be trained to accurately predict compounds that are similar to the ones in the learning set. However their ability to "explain" the result is often low. The models tend to amalgamate different types of properties in an additive way and present rules in a manner that is difficult to interpret (e.g. what does it mean to have activity=  $0.03 * \text{vdW surface} - 0.4 * \text{nHydrogen\_bond\_donor}$ ?). Models generated by numerical methods such as neural nets tend to be even more difficult to interpret, and therefore suggestions for better compounds tend to be limited to trial and error against the model.

Logical (rule-based) methods, on the other hand, usually involve rules that are easily interpreted and have the advantage of being able to explain their decisions. The initial approach to development of so called "expert systems", was strongly limited by the set of rules that had been input "by hand" into the system. The DEREK software (LHASA) is an early example of a rule based expert system applied to predicting toxicity end

points. Current approaches tend to automate the learning process in order to enable the system to automatically generate its set of rules from the learning set.

### *Datamining*

A number of techniques that are similar to the ones used for predictive modeling have been applied to datamining, where the data is searched for hidden relations. Numerical methods such as naïve Bayes (probabilistic model) as well as several rule-based methods have been successfully applied to both the mining of relationships in experimental data and to text-mining of scientific literature data. It is important to note that many data-mining activities require access to disparate data and a common “understanding” of the meaning of each data record. ‘Ontology-based’ technologies help build this common framework.

### **Current issues**

Despite substantial progress, the current state of the art in predictive modelling is inadequate, as summarised by H Waterbeemd (Pfizer)<sup>1</sup>: (1) “In general, reliable predictions are only possible for molecules similar to those in the training set”; (2) “Most models [...] use descriptors that are not easily understood by the chemist and not easy to translate into better molecular structures” (and hence have little impact in drug discovery). We would add a third issue: (3) An overwhelming majority of the existing methods focus on predicting a single activity whereas in many phases of the drug discovery process one is confronted with multi-parametric problems, where many activities must be predicted.

### **New trends - Connecting it all together**

Issue (1) has heavily relied on more generalised descriptors that would help extend the predictability outside of the learning set. Addressing issue (2), several groups, such as Mike Abraham at UCL, have focused on developing descriptors that have strong physico-chemical meanings and that are hopefully more understandable. This has also been achieved through the use of hierarchical fragment libraries that help describe a molecule.

Multi-parametric problems (issue 3) have been most difficult to handle. Most convincing approaches use data visualisation techniques (such as Spotfire) combined with simple statistical analysis. These methods have been applied to analysis of data from a number of sources such as proteomics, HTS, virtual screening, etc. They allow the user to consider multiple viewpoints and extract useful knowledge. However, the results produced are highly dependent on the ability of the user to adequately build the best views to extract the information.

A recent breakthrough in this field has come from the application of machine learning techniques that include rule-based methods. Although this method has been successfully applied to problems ranging from fault detection in nuclear plants to optimisation of production processes in industrial environments to financial portfolio optimisations, this is its first application in drug discovery. The Ariana Pharmaceuticals KEM (Knowledge Extraction and Management) methodology automatically generates a set of logical rules from a multi-parametric learning set. These predictive rules can be examined by chemists and are meant to be directly understandable by the user. In order to broaden the system’s knowledge base, an environment allows the knowledge extracted from databases to be *consistently* combined with knowledge extracted through interaction with the user. The end results are a set of predictive rules as well as an ontology that can be used to describe and organise the initial data. The interactive environment allows the user to challenge the validity of each of the rules (and / or the

---

<sup>1</sup> Nature Drug Discovery Reviews, March 2003

experimental data that has lead to generating that rule). The user can mine existing knowledge and develop better predictive models by overcoming lack of data, or poor data, for a particular molecular series. Finally, the user can interactively identify modifications that would improve the overall quality of the molecule in the studied multi-parametric space. This type of system may be the vanguard of the next generation of computational decision support tools for drug discovery.

© Beremans Limited 2004