In silico approaches to predicting drug metabolism, toxicology and beyond

S. Ekins
Concurrent Pharmaceuticals Inc., 502 West Office Center Drive, Fort Washington, PA 19034, U.S.A.

Abstract
The discovery and optimization of new drug candidates is becoming increasingly reliant upon the combination of experimental and computational approaches related to drug metabolism, toxicology and general biopharmaceutical properties. With the considerable output of high-throughput assays for cytochrome-P450-mediated drug-drug interactions, metabolic stability and assays for toxicology, we have orders of magnitude more data that will facilitate model building. A recursive partitioning model for human liver microsomal metabolic stability based on over 800 structurally diverse molecules was used to predict molecules with known log in vitro clearance data (Spearman's rho $-0.64$, $P < 0.0001$). In addition, with solely published data, a quantitative structure–activity relationship for 66 inhibitors of the potassium channel human ether-a-gogo (hERG) that has been implicated in the failure of a number of recent drugs has been generated. This model has been validated with further published data for 25 molecules (Spearman’s rho $0.83$, $P < 0.0001$). If continued value is to be realized from these types of computational models, there needs to be some applied research on their validation and optimization with new data. Some relatively simple approaches may have value when it comes to combining data from multiple models in order to improve and focus drug discovery on the molecules most likely to succeed.

Introduction
With the increased cost to bring a molecule to market, the pharmaceutical industry has little choice but to improve the success rate of drug discovery, both in terms of overall efficiency and ability to pick molecules that will successfully pass the various stages of drug development. Since the late 1990s, there has been a focus on decreasing the attrition of molecules in the clinic by finding as much out about their biopharmaceutical properties as early as possible. The properties relating to absorption, distribution, metabolism, excretion and toxicology (ADME/Tox) have been studied in some detail and, in numerous cases, attempts to increase the throughputs of selected in vitro assays have yielded large corporate databases. The next logical step is to use this data to generate predictive computational algorithms, which can be used to improve the process of drug discovery. Ultimately, only leads with preferable predicted properties should be synthesized. There have been numerous original papers and reviews that have dealt with various aspects of these early in silico ADME/Tox models that, in most cases, used small numbers of molecules, raising the question of their general applicability [1,2]. Recently, there have also been some timely discussions of the appropriate validation techniques that should be used for quantitative structure–activity relationships (QSARs), although these have perhaps not approached this problem from the industrial perspective of having larger quantities of data [3]. Considering that most published QSARs appear to be retrospective in nature rather than truly prospective (as drug discovery is), the value of these validations may be questionable. However, occasionally a model is published using thousands of molecules, such as the recent model for Ames mutagenicity using 2018 molecules [4] and these, in themselves, raise new questions about how sense is made of so many divergent models for different ADME/Tox properties and how they should be used.

The process of model building
Although there has been discussion of specific model technologies, there has been little consideration for the overall process of model development for ADME/Tox. In most major pharmaceutical companies, each week a bolus of new data is incorporated into proprietary databases, and this becomes available for the computational modeller. This raises the issue that, rather than which molecules need to be tested to develop a good generalizable model, instead, at which point does one start using the data for generating computational models, and what is the quality of the data? If the level of noise in the data is high, it is likely that computational model building will be difficult, if not impossible, regardless of the algorithms used. If the data are biased towards a single core structure, any model produced is likely to be of limited value to other discovery programs, unless they are using the same structure type. Once the database is populated by a number of distinct core structures, the generation of a more general model for that property is possible. Depending

Key words: cytochrome P450, human ether-a-gogo, in silico, quantitative structure–activity relationship (QSAR).

Abbreviations used: ADME/Tox, absorption, distribution, metabolism, excretion and toxicology; $Cl_{in vitro}$, in vitro clearance; CYP, cytochrome P450; hERG, human ether-a-gogo; QSAR, quantitative structure–activity relationship; RUM, removes undesirable molecules.
on the type of data (continuous versus classification), the selection of the appropriate molecular descriptors, algorithm and software to use may not be obvious or unanimous, and this could entail some time and investment in using multiple approaches [5]. In companies where data is continually available, there should be no shortage of new data to test the model and to assess whether the model is predictive by a simple comparison of the predicted and observed data. Depending on the requirements of the model, it may be adequate to generate a simple ranking of the property in question (using the Spearman rho coefficient, for example) [6], allowing the rapid triage of very large databases of virtual molecules and picking the predicted best. Once used as a test set, new data from the same experimental system can be used to rebuild the model, regardless of how well it performed. The frequency of model rebuilding may depend on the quality of the model, based on these types of predictions, the amount of structural diversity inherent in the model, the cost of data generation, the number of other different models to build, the time required to build the model, the ability to automate model rebuilding and the frequency of utilization of the model. With these considerations, the following examples represent some of the types of models that can be built using different sized data sets for drug metabolism and toxicology-related properties.

Figure 1 | Comparison of log human Clu and predicted human metabolic stability

The data represent 41 molecules derived with observed human Clu data [21] and predicted human metabolic stability data obtained from a recursive partitioning model derived from 875 molecules. The straight line represents the linear regression, whereas the outer line represents the 95% confidence interval, ($r^2 = 0.34$).

Role of computational approaches for drug metabolism

To date, research predicting drug metabolism has been limited to a number of technologies, such as rule-based tools and algorithms for sites of metabolism [7], electronic models [8,9], homology models [10–12], as well as pharmacophores and QSARs for $K_m$ values [13–15]. In general, the data sets that the QSAR models use are severely limited in both size and structural diversity. In terms of predicting drug–drug interactions, studies published to date are far more advanced, perhaps due to the early development of pharmacophores and models to understand the individual cytochrome P450s (CYPs). There are also the high-throughput-type assays for CYPs [16], which readily enable generation of $IC_{50}$ values, which ultimately could be used in statistical modelling [17]. A recent pharmacophore analysis of CYP3A4, CYP3A5 and CYP3A7 data proved to be quite enlightening, as it suggested that the pharmacophores for CYP3A5 and CYP3A7 inhibitors were more compact than CYP3A4 inhibitors [18].

There is a need for a general ‘filter’ for metabolic stability that is rapid and perhaps less reductionist than using the models derived for individual CYPs. Pharmacophore and statistical models for predicting human in vitro intrinsic clearance (Clu) have been attempted [19,20] using small numbers of molecules. Building a larger model requires perhaps going beyond literature data, and generating a very large diverse model with human liver microsomal data. A new recursive partitioning model was generated using ChemTree (HelixTree; Bozeman, MT, U.S.A.) with 875 molecules and human metabolic stability percentage-inhibition data (Cerep, Seattle, WA, U.S.A.). The model possessed an observed-versus-predicted correlation $r^2 = 0.71$. Owing to the scarcity of literature, human metabolic stability data for diverse structures not in the model, the model was used to predict 41 literature molecules with Clu data [21] not present in the training set, and known to be cleared by many hepatic, renal and non-metabolic processes [21]. Although perhaps not expected to correlate directly because of the relationship of metabolic stability complexity and the metabolic Clu parameter [22], remarkably, the model is able to generate a statistically significant ranking for the 41 molecules not in the training set (Figure 1; Spearman’s rho $\sim 0.64$, $P < 0.0001$, $r^2 = 0.34$), with only phenobarbital and tolcapone outside of the 95% confidence interval. This suggests that a model like this may be useful for ranking the clearance of structurally diverse new molecules based on human microsomal metabolic stability data. Future tests with new metabolic stability data are planned as a more compatible test set. It is hoped that at some point metabolic properties may be automatically considered during de novo design, particularly if a general oxidation or clearance potential could be combined into the scoring functions currently used [23]. New algorithms could then be used to reject molecules with poor metabolic stability and, instead, substitute less labile functional groups or blocking groups.

Role of computational approaches for toxicology

Toxicology modelling has become an urgent focus in many companies [24]. The adoption of human ether-a-gogo
Validation of models and ‘beyond’

As we develop larger empirical data sets, the tools commonly used for model building become less interpretable. With thousands of data points, the modelling approaches change to using many hundreds of descriptors and machine-learning algorithms that can perform well with non-linear data to enable the discovery of a QSAR. The most valuable form of validation of any computational model is therefore the prediction of a new test set, which is rarely published. The applicability of a model to a new set of molecules should also be considered simultaneously. Although most groups will use ‘Leave one out’ as a form of model validation, this is increasingly recognized as a weak form of model validation [3]; there should be some attempts at classifying the extent of validation, as follows:

Class 1: Leave one out ($q^2$) (weakest)
Class 2: Leave out a large percentage of data altogether (50%), and scrambling of the activity and descriptors (adequate)
Class 3: Generate a new diverse and large data set after model building, keep observed data and predictions blinded, (best).

With very large data sets for both model building and testing, and possibly commercially available data, we could utilize ‘pre-filtering’, which removes undesirable molecules (RUM), based on structural or physicochemical limits for molecular mass, rotatable bonds, cLogP and particular undesirable substructures. In some respects, this limits the training and test sets with RUM, enabling the avoidance of extremes in molecular characteristics and the increasing likelihood of capturing a similar chemical space.

The combination of computational models will be valuable as a way to present a more readily understandable output for chemists, such that for one data set there may be multiple models with different modelling software or descriptors. For example, there may be multiple metabolic stability models that can be combined to give an overall score for this property. Alternatively, different data sets for the same property, descriptors and algorithms may be created and these may be combined, or data sets for different types of properties may be combined. For example, an overall ADME/Tox score could be created by summing the scores derived for all the ADME models and all the toxicity models. In a rather simplistic manner, the stratification of models may depend on importance, weighting, confidence and size, amongst other factors that relate specifically to the therapeutic programs. A summation of such models could be desirable as a more holistic approach, in a move away from our prior modelling reductionism. Recognition that a molecule fails may not be due to a single model (‘serial filtering’), but is dependent on the balance of other models (‘parallel filtering’). There may be no clear optimal way to combine the data for multiple predictions, and this represents an area in urgent need of study. Ease of use and visualization of any combination of models is key: a simple scale used for all models (or the output of a combination), ordinal number- or colour-based, may work best in expressing data derived from diverse experimental units (e.g. $-\log IC_{50}$, percentage inhibition, etc.).

The late 1990s saw the start of computational ADME/Tox approaches on a reductionist paradigm, e.g. a model for

**Figure 2** | Test set analysis for hERG log observed $IC_{50}$ ($\mu M$) and log predicted $IC_{50}$ ($\mu M$)

The data represent 25 molecules derived from the literature with log observed hERG $IC_{50}$ data and log predicted hERG $IC_{50}$ data obtained from a recursive partitioning model derived from 66 molecules. The straight line represents the linear regression, whereas the outer line represents the 95% confidence interval (Spearman rho 0.83, $P < 0.0001$, $r^2$ 0.67).
one enzyme, transporter, receptor or channel. Development of certain models has almost been on an ad hoc basis. If a certain therapeutic program requires a single property optimized, a QSAR model is developed and then probably discarded. Models generated in this way may have some value for the future too. Computational models based on anti-arrhythmics could have helped prevent some companies from selecting non-cardiovascular drugs with hERG inhibitory activity, and ultimately saved them money. This decade should see the beginning of holistic approaches for optimizing multiple ADME/TOX properties simultaneously, which will be central to the development of this field.

I am grateful for the support and suggestions of Dr Maggie A.Z. Hupcey and colleagues at Concurrent Pharmaceuticals Inc.

References


Received 16 January 2003

©2003 Biochemical Society