Increasing the efficiency of small-molecule drug discovery

Barry A. Bunin, President & CSO, Libraria, 2372-D Qume Drive, San Jose, CA 95050, USA; tel: +1 650 219 4153, fax: +1 408 383 0799, e-mail: bunin@libraria.com. Please note, Libraria has recently been acquired by Sertanty.

A recent interview with Chris Lipinski in *Drug Discovery Today* highlighted the challenges of improving efficiency within the drug discovery industry [1]. At the heart of the challenge of making the drug discovery process more efficient is the fact that it is fundamentally an experimental and iterative process. The drug discovery process will become more efficient as a direct function of how well we can: (i) take both technical and human considerations into account, (ii) profile and prioritize molecules using all available relevant information, (iii) develop multiple chemotypes to compensate for attrition, (iv) take chemical feasibility into account upfront, and (v) better integrate computational with experimental approaches to drug discovery. Potential solutions to the challenges associated with increasing the efficiency of drug discovery are discussed here, with specific suggestions of the best methods to implement the proposed solutions and what problems are likely to be encountered in the process. This analysis does not address antibody therapies, biological therapeutics, nor the complex field of target validation, which are all, of course, important considerations, but defines the most efficient strategies for the discovery and development of small molecule drug therapies to treat human disease.

Take both technical and human considerations into account

Technology must take into account the way in which scientists react to data and the ramifications of such reactions. When a scientist obtains information, it can quickly become a 'truth', therefore the quantity, quality, and timing of the presentation of data is important in influencing the ultimate decisions. Computational scientists often fault experimentalists for not taking computational approaches into account initially. Conversely, empirical scientists often bemoan the quality of the predictive algorithms. The key is a complete understanding of the scope and limitations of computational approaches; with this, new models or paradigms can be placed in an appropriate context that best leverages the logic, creativity, and desire of the empirical scientists, who will, ultimately, produce and evaluate the drug candidates. The challenge is to shape the technology to complement human intuition, rather than compete with it. One of the reasons that the Rule of Five [2] was so widely adopted, while so many other computational methods were not, is that it was based on statistical analyses of experimental data and framed in simple terms that scientists of different backgrounds could easily remember and adopt it. Perhaps the most important factor in the judicious use of new technologies, like predictive modeling, is to understand in which circumstances a meaningful association is best identified by automated algorithms and in which, by contrast, a scientist can best identify the relevant association. Each time a process becomes automated, scientists are able to focus their limited attention on better defining the most relevant problems. The trend towards the automation of routine molecular analysis inspires much optimism.

Profile and prioritize molecules using all available relevant information

One approach to making the drug discovery process more efficient is to create statistically robust theoretical models that simultaneously mimic each experiment, synthesis, assay, animal model and clinical trial as closely as possible. Physical chemical profiling is convenient because it is an absolute measure that is a function of the molecular structure and can be directly calculated without any experimentation. In contrast, QSAR and modeling technologies can predict which individual molecules are more likely to be active against a range of targets, thereby enabling significant acceleration of the identification of novel chemotypes.

Much effort has been placed on increasing the sophistication of predictive algorithms. In contrast, less effort has gone into creating appropriate training sets of molecules with associated data to build up better models that more accurately mimic actual experiments. There is no fast and convenient way to capture scientific data because the data are, by nature, complex and need to be evaluated within their biological context. Nevertheless, when captured in an intelligent fashion, the data can be preserved in a format that will be

1359-6446/03/$ – see front matter PII: S1359-6446(03)02747-8

useful not only for today's problems but also for tomorrow's challenges – resulting in greater long-term efficiency.

Various QSAR predictive methods have been developed to help direct the synthesis and screening process [3–7]. In general, the quality or relevance of any predictive model is a function of the quality of algorithm, the quality of data and the contextualization of the predicted property. Whereas the 'quality of algorithm' component has been widely studied (as evidenced by the number of journal articles, books, and even entire conferences dedicated to the subject) [8–10], the 'quality of data' component is far less studied.

'Quality of data' can be further broken down by evaluating the signal:noise ratio of the data (quality of the assays), the diversity of the assays (appropriate contextualization), the number of compounds evaluated per assay and the corresponding diversity of different chemotypes. The quality of the data depends on the number of different compounds, the structural diversity of different compounds, the quality of the assays, and the similarity of assay conditions (this approaches 100% when all the assays are run under precisely the same conditions).

It could be argued that a QSAR model based on external, industry-wide experimentation, is better than a model built on internal experimentation because a diverse set of different chemotypes are obtained from different laboratories. However, it could also be argued that external data are worse than internal experimentation for building models because assay conditions are not completely uniform and data might not be the best and most recent. A model that takes into account the assay results of internal and external experimentation ultimately represents the best of both worlds. Furthermore, this is technically feasible, given the current database technologies, and would not require disclosure of any intellectual property in the combining of public data with a proprietary SAR knowledge base. The appropriate 'contextualization of the predicted property' is a complex process to automate; 'contextualization' refers to the exact assay conditions and how similar they are in interpreting the meaning of the very terms potency and selectivity. The difficulty is that the best method of analyzing the data is often project-, target(s)-, or disease-specific. Nonetheless, decisions must always take into account this complexity. Ultimately, we will need better meta-models using larger sets of related data to define the relative importance of molecular properties that are relevant to particular targets and gene families, and to the overall drug discovery process.

Develop multiple chemotypes to compensate for attrition

It is human nature (and often necessary) to focus on the first, or best, chemical lead exclusively. There is tremendous pressure to advance the lead compound as far as possible. Within smaller companies, in particular, there are usually sufficient resources to focus only on the one or two most promising candidates. The issue is that the majority of compounds are ultimately dropped from development, sometimes for efficacy or selectivity reasons, but equally often for reasons associated with the specific molecule or molecular class (PK, ADME or toxicity). It is particularly important to have multiple options, in terms of distinct chemotypes, early in a medicinal chemistry project life cycle and to have a variety of different scaffolds with varied synthetic methodologies to allow medicinal chemists to design out any functionality responsible for attrition. Novel chemotypes are also desirable because of the prospect of entering unclaimed patent space. By-passing the optimization of side chains around existing scaffolds to design novel chemotypes (perhaps with better properties such as ease of synthesis or various physical chemical properties) has been referred to as 'scaffold jumping', a term that accentuates the difference and greater challenge involved [11–13]. The development of this technology is extremely beneficial for medicinal chemists because it enables them to pick and choose the more tractable chemotypes on which to focus their efforts.

Another maturing technology is an outgrowth of combinatorial chemistry that will enable scientists to define the majority of virtual chemistry space that is actually synthetically feasible. The number of commercially available discreet molecules has been estimated as ~10^8. The number of potential completely theoretical molecules has been estimated as ~10^{60}. The number of molecules in the accessible 'actual–virtual' libraries, based on known synthetic pathways and commercially available precursors, is currently unknown but it is much greater than 10^8 and much less than 10^{60}. It is this universe of 'actual–virtual' synthesizable molecules that is relevant in all computational models for prioritization when using scaffold-jumping technologies. These technologies are now maturing to the level where a good training set is valuable as intellectual property and synonymous with a novel chemotype after a rapid screening exercise when using the universe of synthesizable (versus just 'purchasable') molecules as a source pool.

Take chemical feasibility into account upfront

Lipinski notes that much more effort has been put into the development of methods to predict the bioactivity and drug-likeness of compounds (to address what should be produced), than has been put into the development of

predictive methods to account for chemical feasibility (what can be produced) [1]. Why is this?

First, it is human nature to focus on determining what should be made, as that is the ultimate answer to the fundamental problem and the fewer molecules that need to be synthesized to get there, the better. However, predictive models that enhance the probability of accurately prioritizing what should be made will be successful relative to the range of truly different drug-like molecules that are accessible through diverse, robust chemistries. Thus, problems of compound synthesis prioritization and feasibility are completely interrelated. Integrating synthetic feasibility models into small molecule drug discovery strategies is complex but very important, if we aim to produce non-linear improvements in efficiency. Those scientists who have optimized chemical reactions for combinatorial libraries or process chemistry realize the extent of experimentation required to identify optimal conditions.

Although the prediction of synthesis is a difficult technical challenge, some progress has been made. Johann Gasteiger's research group (http://www2.chemie.uni-erlangen.de/presentations/symposium/ torvs_e.pdf) has developed computational programs that simulate chemical reactions in a mechanistic manner. These include EROS (Elaboration of Reactions for Organic Synthesis), WODCA (Workbench for the Organization of Data for Chemical Applications) for the simulation and prediction of organic reactions and CORA (Classification of Organic Reactions for Applications), which uses reaction databases to derive knowledge on chemical reactions. At Libraria (http://www.libraria.com, recently acquired by Sertanty), a similar knowledge-based approach has been used; a program called LUCIA (Libraria's Unique Chemically Intelligent Archive) accurately defines the most synthetically accessible chemical spaces by capturing robust and reliable generic reaction transformations. With such an array of chemical reactions (e.g. acylations, reductive aminations and Suzuki couplings), there is the potential to create chemical reaction training sets appropriate for building predictive chemical compatibility algorithms. LUCIA combines a reaction-chemistry archive with a robust enumeration engine as the first step towards addressing the question of chemical feasibility. For example, one could imagine collating all the examples in the published literature on Suzuki coupling reactions with diverse aryl halides and boronic acids to glean the inherent reactivity trends and side chain reaction incompatibilities that define the scope and limitations of chemical reactions as a function of the actual reactions being performed throughout the chemical community.

Traditionally, chemical feasibility was considered in terms of retrosynthetic analysis from final products; the focus has now shifted towards consideration of chemical feasibility in terms of a 'forward' synthetic analysis from commercially available compounds. The LHASA (Logic and Heuristics Applied to Synthetic Analysis) program from Harvard, which has long been used as a tool for retrosynthetic analysis, now includes a forward analysis module for virtual library enumeration (http://lhasa.harvard.edu/LHASApublications.htm).

Chemical feasibility is at the heart of small molecule drug discovery and chemotype prioritization; essentially, it defines what can and cannot be analoged. Analogability is not the only factor, but is a major factor that is often overlooked. When screening commercial compounds or complex natural products, leads are often dropped because they may be difficult to rapidly analog during lead optimization. The desirability of a chemotype is a function of the drug-likeness, potency, novelty, and analogability. If this process of exploring truly synthetically accessible chemical spaces can be automated, then the exciting possibility of modeling the iterative synthesis and screening cycle would emerge. Predicting, or even just mapping, synthetic feasibility is a sleeping giant; few people are looking into it but the ramifications of a breakthrough would be revolutionary for both chemistry and drug discovery.

Better integrate computational with experimental approaches

Time available for experimentation is limited. The only way to achieve a quantum jump in medicinal chemistry productivity is through the creation of better models and tools that complement, rather than replace, human intuition. We will always be limited to a finite number of molecules that can be economically synthesized and evaluated. Despite the advances in automation technologies, combinatorial chemistry, and higher-throughput screens that improve our ability to rapidly confirm or disprove hypotheses, the synthesis and screening cycle remains the rate-determining process in preclinical R&D. Fortunately, we continue to make great strides in the quality and refinement of predictive algorithms and in the breadth of the training sets amassed. These advances will significantly improve the effectiveness of the drug discovery cycle.

There might come a day in the not-so-distant future, when large sets of molecular files of hypothetical molecules for synthesis are intelligently profiled and electronically screened against meaningful models for entire gene families (see Figure 1). Statistically significant enhancements would allow much smaller sets of compounds to be tested against a range of judiciously

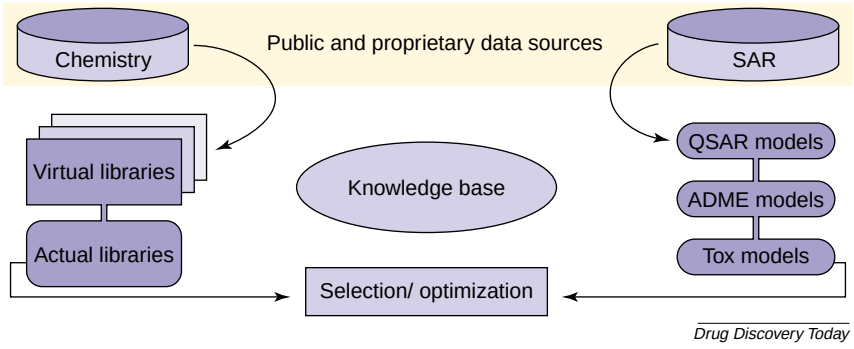

Figure 1. Knowledge-based synthesis and screening models of the iterative drug discovery process.

selected higher-quality bioassays of relevance to disease states and thus, the iterative synthesis and screening drug discovery process would become more cost effective. Each experiment could be optimally selected through the integration of human intelligence and the intelligence of all the historical data generated. A general truth about new technologies is that their short-term impact is often overestimated…and their long-term impact is often underestimated.

References

1 Owens, J. (2003) Chris Lipinski Discusses Life and Chemistry After the Rule of Five. *Drug Discov. Today* 8, 12–16

2 Lipinski, C. A. *et al*. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv. Rev* 46, 3–26

3 Barnum, D. *et al*. (1996) Identification of Common Functional Configuration Among Molecules. *J. Chem. Inf. Comput. Sci.* 36, 563–571

4 Greene, J. et al (1994) Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* 34, 1297–1308

5 Cramer, R.D. III et al. (1998) Comparative Molecular Field Analysis (CoMFA), 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* 110, 5959–5967

6 McGregor, M.J. and Muskal, S.M. (1999) Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* 39, 569–74

7 Matter, H. (1997) Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* 40, 1219

8 Bohm, H.J. and Stahl, M. (2000) Structure-Based Library Design: Molecular Modelling Merges with Combinatorial Chemistry *Curr Opin Chem Biol* 4, 283–286

9 Joseph-McCarthy D. (1999) Computational Approaches to Structure-Based Ligand Design. *Pharmacol Ther* 84, 179–191

10 Diller, D.J. and Merz, K.M. Jr. (2001) High Throughput Docking for Library Design and Library Prioritization. *Proteins* 43, 113-124

11 Schneider, G. *et al*. (1999) 'Scaffold-Hopping' by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem. Int. Ed.* 38, 2894–2896

12 Naerum, L. *et al*. (2002) Scaffold Hopping and Optimization towards Libraries of Glycogen Synthase Kinase-3 Inhibitors *Bioorg. Med. Chem. Lett.* 12, 1525–1528

13 Schneider, G. *et al* (2001) Integrating Virtual Screening Methods to the Quest for Novel Membrane Protein Ligands. *Curr. Med. Chem. – Central Nervous System Agents* 1, 99–112