The *Discussion Forum* provides a medium for airing your views on any issues related to the pharmaceutical industry and obtaining feedback and discussion on these views from others in the field. You can discuss issues that get you hot under the collar, practical problems at the bench, recently published literature, or just something bizarre or humorous that you wish to share. Publication of letters in this section is subject to editorial discretion and company-promotional letters will be rejected immediately. Furthermore, the views provided are those of the authors and are not intended to represent the views of the companies they work for. Moreover, these views do not reflect those of Elsevier, *Drug Discovery Today* or its editorial team. Please submit all letters to Steve Carney, Editor, *Drug Discovery Today*, e-mail: S.Carney@elsevier.com

Understanding chemoinformatics: a unifying approach

The term chemo- (chem-, chemi-) informatics is currently used widely in the literature, and often in a fairly loose manner. Thus, it might be appropriate and timely to put this discipline into scientific context and question what chemoinformatics really is, or perhaps should be.

In 1998, Brown first introduced this term and essentially defined it as the combination of ‘all the information resources that a scientist needs to optimize the properties of a ligand to become a drug’ [1]. In Brown’s definition of this field, both decision support by computer and drug discovery relevance are fundamentally important aspects of chemoinformatics. However, as pointed out by Goodman, the term ‘chemical informatics’ was already used much earlier and more generally defined as the ‘application of information technology to chemistry’ [2]. Thus, chemical informatics lacks a specific focus on drug discovery, in contrast to chemoinformatics. In addition, there is ‘chemometrics’, which is generally understood as the ‘application of statistical methods to chemical data’.

In drug discovery, it is of course recognized that gaining knowledge from chemical data alone is not sufficient to be ultimately successful and that boundaries between informatics disciplines in chemical and life sciences are fluid [3]. Such insights are also reflected by the introduction of terms such as ‘discovery informatics’ [4].

However, if we wish to focus on chemoinformatics in a more narrow sense, what should we really refer to? In light of the many and, in part, conceptually overlapping methodological developments in this area, I would propose that it has become increasingly difficult to distinguish between chemical informatics, chemometrics and chemoinformatics. Therefore, I would suggest combining the following types of computational methods and infrastructures under the chemoinformatics umbrella, regardless of whether or not they are applied in the context of drug discovery:

- chemical data collection, analysis and management;
- data representation and communication;
- database design and organization;
- chemical structure and property prediction (including drug-likeness);
- molecular similarity and diversity analysis;
- compound or library design and optimization;
- database searching and virtual screening;
- compound classification and selection;
- qualitative and quantitative structure–activity or –property relationships;
- information theory applied to chemical problems;
- statistical models and descriptors in chemistry;
- prediction of *in vivo* compound characteristics.

This scheme takes into account that the field is still evolving and implies that approaches that are long disciplines in their own right (e.g. QSAR) are also an integral part of the chemoinformatics spectrum. Thus, research in this area should be capable of adopting established scientific concepts and putting them into a novel context (e.g. the development of QSAR models as virtual screening tools [5]). Furthermore, in its extended definition, chemoinformatics includes all concepts and methods that are designed to interface theoretical and experimental efforts involving small molecules (such as virtual and high-throughput screening [6]), thus emphasizing the experimental relevance of many developments in the chemoinformatics arena.

References

1 Brown, F. K. (1998) Chemoinformatics: what is it and how does it impact drug discovery. *Ann. Rep. Med. Chem.* 33, 375–384

2 Goodman, J. M. (2003) Chemical informatics. *Chem. Inf. Lett.* 6 (2); (http://www.ch.cam.ac.uk/MMRG/CIL/cil_v6n2.html#14)

3 Bajorath, J. (2001) Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. *Drug Discov. Today* 6, 989–995

4 Claus, B. L. and Underwood, D. J. (2002) Discovery informatics: its evolving role in drug discovery. *Drug Discov. Today* 7, 957–966

5 Hopfinger, A. J. *et al.* (1999) Construction of a virtual high throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase *b*. *J. Chem. Inf. Comput. Sci.* 39, 1151–1160

1359-6446/04/$ – see front matter ©2004 Elsevier Ltd. All rights reserved. PII: S1359-6446(03)02916-7

6 Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nature Rev. Drug Discov.* 1, 882–894

Jürgen Bajorath
Senior Director
Computer-Aided Drug Discovery
Albany Molecular Research
Bothell Research Center
and University of Washington
Seattle
Washington, USA
e-mail: jurgen.bajorath@albmolecular.com

Positioning ADMET *in silico* tools in drug discovery

Recently there has been a dramatic increase in the size of compound collections in pharmaceutical companies and also, because of ultra high-throughput screening, an increase in the rate at which biological activity data can be obtained. The generation of pharmacokinetic and safety data at the lead optimization stage has therefore struggled to keep up with the screening of compounds. Virtual screening of compound libraries has also become a component of many lead generation and optimization programmes. Large pharmaceutical and many smaller chemistry-oriented companies support substantial efforts in this area. The hope is that *in silico* screening and analysis will aid significantly in addressing the balance between drug potency and various ADMET properties.

Although ADMET *in silico* tools can be used for high-throughput screening, their main benefit is in predicting properties of compounds before they are synthesised and also in understanding the relationship between chemical structure and ADMET properties. In my personal experience I have seen a number of examples that show that blending measurements together with information generated by these assays can offer a better chance of success than just by testing more compounds alone. However, ADMET *in silico* models are frequently complex and it is perhaps not surprising that many experimentalists perceive this field as an 'algorithmic jungle'. Consequently, the benefits and limitations of virtual screening are sometimes misunderstood.

In silico predictions are probably no less predictive of what occurs *in vivo* than are *in vitro* tests. They have the decisive advantage of being cheap and they enable predictions to be made on virtual compounds. Although these *in silico* models do undoubtedly work, in many cases experimentalists often prefer to generate 'wet' data on all the compounds, irrespective of what the odds of success might be. If we are to fully capitalize on the opportunities presented by *in silico* tools, implementation and integration of these tools into drug discovery processes needs to be carried out in a rational and systematic manner. We need to better understand the relationship between the physicochemical properties and structure of a molecule and its likely fate in the body. For example, for intestinal absorption there are at least five different processes that can affect the absorption of a molecule. These are passive paracellular, passive transcellular, efflux and/or influx transporters, solubility and dissolution rate. As with any complex problem, the task of building an understanding becomes less daunting if it can be broken into different, simple processes.

We also have to resolve the issue of whether global or local models should be used. Global models that are based on data from several programmes are usually good in modeling general phenomena and trends but are less good at enabling an understanding of the effect of small structural differences. By contrast, local models, which are built on a particular chemical series, might work well within a closely related series but will quickly cease to be applicable as the synthetic direction of a programme changes. *In silico* modelers should therefore check the validity of their models by selecting and obtaining experimental data from new compound sets with structures different from those in the original model training sets. It is also important that *in silico* modelers choose the right data to model and use descriptors that convey a simple message to experimentalists. This is why Lipinski's 'Rule of Five' [1] is widely used and has gained acceptance with chemists and biologists. As these tools become more user-friendly, and as more examples of successful applications are shown, it seems highly probable that *in silico* approaches will evolve rapidly, as was the case with *in vitro* methods during the last two decades. It should also be acknowledged that *in silico* tools have been in existence for a relatively short time and it is therefore unrealistic to expect good predictions in every application. However, *in silico* modelers must adhere to best practice, performing adequate validation and testing the reliability of the predictions to ensure that these tools are used appropriately. It is also equally important for model developers to work with programme teams to explore and understand the reasons why these tools demonstrate limitations in certain cases.

Drug discovery has always been a competitive industry and now the stakes are even higher than ever. In terms of sales and profits, only one in five new products launched since 1997 has been 'significant' [2]. A combination of technologies such as HTS and *in silico* models can offer great advantages in improving the odds of success in a discovery programme. In the early stages this can be done by eliminating the guesswork and decreasing the experimental load. What needs to be avoided is using *in silico* models simply in addition to *in vitro* and *in vivo* experiments, rather than being used in

1359-6446/04/$ – see front matter ©2004 Elsevier Ltd. All rights reserved. PII: S1359-6446(03)02956-8