Chemoinformatics: are we exploiting this new science?

'We need to make chemoinformatics tools more accessible to the bench chemist...'

Chemoinformatics is one of our newest sciences. Frank Brown[1] deserves credit for introducing this term as 'the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization'. Who can argue with the combined logics of converting data to knowledge for identifying potential drugs faster? If we embrace this premise, then we are obliged to examine and track our progress with chemoinformatics.

Chemoinformatics has been confused with bioinformatics. Bioinformatics is an older science, which was genomics-inspired, and which covers the computational aspects of genomics. Had chemoinformatics preceded genomics as an established science, one could argue that bioinformatics should logically be a subset, admittedly a large subset, of chemoinformatics. In terms of nomenclature, bioinformatics can manage the structural information using the four letters of the genetic code (ATGC) or the 20 letters coding amino acids. Chemoinformatics also has to handle the large variety of chemical structures, including stereochemistry.

Chemoinformatic techniques

Important facets of chemoinformatics are included in Box 1. Compound registration, with links to synthesis records (in an ideal situation, electronic lab notebooks), is a crucial first step without which subsequent events could not occur. Entry of chemical structures with correct stereochemistry, as well as links to spectral data (e.g. NMR) and purity determinations (e.g. HPLC) that were recorded during or before the compound registration process is a required first step. Complete incorporation of all biological activity (or inactivity) data that is being gathered is a dynamic and crucially important process for developing profiles for registered compounds.

Tools for relating structure to biological activity would include comparative molecular field analysis (CoMFA), which is especially useful when a target protein conformation is unknown, and the shape of the ligand docking site can be deduced only by examining ligand sets. CoMFA can determine receptor surface shape by defining volumes and excluding volumes occupied by ligands with a broad range of affinities. The calculation of physiochemical properties such as cLogP or polar surface area (PSA), a process referred to as chemometrics, has become relatively standard for developing correlation with projected bioavailability.

Virtual database assembly is a crucial activity as it enables access to the large number of drug-like molecules that could theoretically be made, and is estimated to be 10^{40} (Ref. 2). This can serve several purposes: for example, to generate a maximally diverse virtual library for lead generation, a biased library aimed at a specific target or target family, or a lead optimization library.

Database mining for analogues of screening hits can be performed using substructure, 2D- or 3D-similarity, molecular shape, framework, or (if known) pharmacophore searching. The molecules in a 3D database are usually represented by a set of low energy, accessible conformers suitably encoded to minimize storage space.

At a time when the cost of screening vast libraries is a serious issue, virtual screening of the theoretical (or real proprietary) databases has become a more crucial function. As theoretical databases are not proprietary, everyone will ultimately have access. Although several virtual screening efforts have been reported, not all have been applicable to all targets[3]. Only recently has a virtual screening method been developed that accommodates zinc in the active site of a metalloprotein, which has been used successfully in identifying micromolar inhibitors of the important oncology target,

*Nicholas J. Hrib, Aventis Pharmaceuticals, Bridgewater, NJ 08807, USA and Norton P. Peet, ArQule, Woburn, MA 01801, USA.
*tel: +1 908 231 2761, fax: +1 908 231 3632, e-mail: nicholas.hrib@aventis.com

1359-6446/00/$ – see front matter ©2000 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(00)01560-9

Box 1. Essential components of chemoinformatics

- Compound registration
- Chemical structure, spectral data, purity determinations
- Biological activities
- Tools for relating structure with biological activity
- Calculation of physiochemical properties
- Virtual database assembly
- Virtual screening
- Database mining (real/virtual)
- Structure-based design
- Statistical methods
- Visualization

farnesyl transferase[4]. Obviously, the success rate of the virtual screening will depend on both the quality of information about the target (crystal structure of receptor binding site or enzyme active site, well-defined 3D pharmacophore, etc.) and the quality of 'fitting' software (docking of molecules to the target). The speed of the screen (time required to evaluate conformation sets of each molecule against the target) is also a major factor.

Statistical methods such as principal components analysis or factor analysis are used to reduce the large number of molecular descriptors (e.g. molecular weight, LogP, number of H-bond donors or acceptors, rotatable bonds, PSA) for a set of compounds to a smaller set of crucial descriptors that can be more readily handled computationally[5]. Genetic algorithms enable optimized solutions to develop by challenging a 'population' of solutions. This technique can be applied to find those crucial descriptors that most reliably predict the properties of a set of analogues[6].

The visualization of large sets of data must also be facilitated. Manual assembly of a table of structure–activity relationship (SAR) data might have been acceptable in evaluation of a dozen compounds, but not in the analysis of hundreds (or thousands) of compounds using multiple criteria (activity *in vitro*, *in vivo*, transport, etc.). Programs such as Spotfire™ (Spotfire, Cambridge, MA, USA) and LeadScope™ (LeadScope, Columbus, OH, USA) enable graphical plotting of activities, filtering of data to facilitate analysis, and linking with chemical structure displays to assist in pharmacophore detection[7].

The traditional WIMP (Windows, icons, mouse and point-and-click) visualization techniques are currently still very useful. However, new futuristic techniques are emerging such as immersive virtual reality (IVR). IVR enables the user to literally become a part of his or her data and to use additional senses[8]. Although IVR has not yet enjoyed widespread use in scientific disciplines, it has been cost-effective in architectural design.

Challenges and chemoinformatics

Chemoinformatics is an integral part of the discipline of knowledge management. Distribution of knowledge is a measure of how well we are performing as an industry with knowledge management (and chemoinformatics, by association). It has been estimated that only 58% of our organizational knowledge is documented and that only 12% of the documented knowledge is accessible to others[9]. These are abysmal statistics, which point to our need for better capture and display techniques for knowledge. With chemoinformatics being a principal method for converting data to knowledge or creating knowledge in the pharmaceutical industry, we need to focus on how to better handle this knowledge once it is generated.

Monitoring progress of chemoinformatics

To be competitive, a research organization must be able to rapidly assemble all historical biological data on a set of HTS hits, along with all of their calculated/measured physicochemical properties, to enable a rapid prioritization decision to be made on which series are 'druggable', and should therefore receive precious chemical synthesis resources. We must have the capability to take a rationally designed chemical structure and generate a library of virtual analogues, along with calculated metrics (cLogP, PSA) and bioavailability, metabolic stability and toxicity predictions, so that chemistry efforts can focus on the molecules most likely to produce drug candidates[10]. We must also be able to filter and multidimensionally analyze a large set of biological data (correlated with displayable chemical structures) to disclose key trends and generate histograms relating structural features (including user-defined structural categories) to activity/bioavailability/safety.

Our industry predicts that we will triple the success rate of products by 2008 (Ref. 11). If we are truly to achieve this goal, our efforts must be focused from the very beginning on those structures most likely to lead to drugs and on the rapid culling of false leads (those unlikely to survive downstream tests of bioavailability, metabolic profile and safety).

The future

Clearly, much needs to be done, as chemoinformatics is a fledgling science and one that the scientific community should have initiated ten years earlier. We are on the verge of identifying 40,000–120,000 human gene sequences from the ongoing genome mapping projects, and the fallout from target identification/validation will place an unbelievable strain on our understaffed chemoinformatics initiatives. We need to make chemoinformatics tools more accessible to the bench chemist (balanced with the appropriate security protocols) so that the elimination of time wasted pursuing poor candidates and the focus of efforts on 'druggable' molecules can begin as early as possible. Web-based, intuitive interfaces enabling access to various programs can help[12], but tighter integration of these applications is still needed to rapidly assemble the information required for faster, more reliable decisions. We also need to develop an increasing awareness of new and more predictive molecular descriptors and pharmacophore techniques[13]. Furthermore, we need to foster and embrace futuristic visualization techniques for human–computer interactions, such as IVR, that could be applied to drug discovery. Only if we act now on these initiatives will we have tools in place in time to handle the floods of data on our horizon.

REFERENCES

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

2 Martin, Y.C. (1997) Challenges and prospects for computational aids to molecular diversity. *Perspect. Drug Des. Discovery* 7/8, 159–172

3 Walters, W.P. *et al.* (1998) Overview of virtual screening. *Drug Discov. Today* 3, 160–178

4 Perola, E. (2000) Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem*. 43, 401–408

5 Walters, W.P. *et al.* (1999) Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* 3, 384–387

6 Gillet, V.J. *et al.* (1998) Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* 38, 165–179

7 Ahlberg, C. (1999) Visual exploration of HTS databases: bridging the gap between chemistry and biology. *Drug Discov. Today* 4, 370–376

8 Ihlenfeldt, W-D. (1997) Virtual reality in chemistry. *J. Mol. Model.* 3, 386–402

9 Lawrence, R.N. (1999) Enhancing information sharing. *Drug Discov. Today* 4, 494–495

10 Brennan, M.R. (2000) Drug discovery: filtering out failures early in the game. *Chem. Eng. News* 78, 63–73

11 Anderson Consulting (2000) *Speed to Value: Delivering on the Quest for Better Medicines*, March 10, Reuters

12 Guner, O.F. and Casher, O. (1999) Role of the Internet in chemo-informatics: recent developments. *Curr. Opin. Drug Disc. Dev.* 2, 204–208

13 Hahn, M. and Green, R. (1999) Chemoinformatics – a new name for an old problem? *Curr. Opin. Chem. Biol.* 3, 379–383

Nicholas J. Hrib and Norton P. Peet