Pharm Res (2010) 27:2035–2039
DOI 10.1007/s11095-010-0229-0

PERSPECTIVE

Chemical Space: Missing Pieces in Cheminformatics

Sean Ekins • Rishi R. Gupta • Eric Gifford • Barry A. Bunin • Chris L. Waller

Received: 22 April 2010 / Accepted: 23 July 2010 / Published online: 4 August 2010

ABSTRACT Cheminformatics is at a turning point, the pharmaceutical industry benefits from using the various methods developed over the last twenty years, but in our opinion we need to see greater development of novel approaches that non-experts can use. This will be achieved by more collaborations between software companies, academics and the evolving pharmaceutical industry. We suggest that cheminformatics should also be looking to other industries that use high performance computing technologies for inspiration. We describe the needs and opportunities which may benefit from the development of open cheminformatics technologies, mobile computing, the movement of software to the cloud and precompetitive initiatives.

S. Ekins (✉)
Collaborations in Chemistry
601 Runnymede Avenue
Jenkintown, Pennsylvania 19046, USA
e-mail: ekinssean@yahoo.com

S. Ekins • B. A. Bunin
Collaborative Drug Discovery
1633 Bayshore Highway, Suite 342
Burlingame, California 94010, USA

S. Ekins
Department of Pharmaceutical Sciences
University of Maryland
College Park, Maryland 21201, USA

S. Ekins
Department of Pharmacology University of Medicine & Dentistry of New Jersey (UMDNJ)—Robert Wood Johnson Medical School
675 Hoes Lane
Piscataway, New Jersey 08854, USA

R. R. Gupta • E. Gifford • C. L. Waller
Pfizer Global Research and Development
Eastern Point Road
Groton, Connecticut 06340, USA

KEY WORDS ADME/Tox · cheminformatics · open chemistry development kit · pharmacophores · QSAR

INTRODUCTION

Measure what is measurable, and make measurable what is not so. -Galileo Galilei

2009 was the international year of astronomy celebrated in books, exhibitions and beyond all around the world, all because Galileo had the good sense 400 years ago to apply a technology invented by others (two lenses at opposite ends of a tube to form a telescope) that allowed him to see the stars beyond our solar system and forever change our understanding of it. Drug discovery needs a few Galileos to bring in technologies from other fields, and the timing is right to be provocative and provide a long overdue paradigm shift (1). Are there technologies that we could bring together in pharmaceutical research that may seem rather simplistic yet if combined could lead to new insights? A case of the parts being greater than the whole, we are sure Aristotle would approve. We are motivated to suggest this because our particular field of interest, cheminformatics, is still mired in addressing fundamental problems that have existed for decades. We believe that this may not be unique to us and our rather narrow perspective and could be of much broader interest, as other facets of pharmaceutical research and development could benefit from the opportunity to reassess the situation (motivated by company mergers, discussion of new R&D models, precompetitive approaches, more collaborations, etc.) (2–4). As we know, it is also important to access your progress as you participate in a research project, whether developing a drug or a technology, take stock and change direction as necessary.

THE COMET

In our humble opinion, as users of cheminformatics software for anywhere from 13 to more than 20 years, we have seen the field plagued with several unresolved issues and apparent unnecessary repetition, like a comet that revisits us periodically. First, the major innovative cheminformatics developments could be recognized to have largely been initiated in the 1980s, and early 1990s, e.g. the following represents examples of technologies that includes comparative molecular field analysis (CoMFA) (5), docking (6), pharmacophores (7–9), 3D database searches (10,11), molecular descriptors (12) and similarity searching (13). Without wishing to offend our colleagues, we think they would agree that many improvements have been relatively cosmetic compared with the development of whole new transformational or disruptive technologies. We have also seen the few software companies that control this space essentially try to mimic the same technologies of their competitors, so we are in the position of having several platforms for storing and mining molecules, many docking, QSAR or other searching methods, etc. Competition is a good thing, but the wheel has already been invented. These computational technologies could, in most eyes in the pharmaceutical industry, now be seen as commodities. Truly novel developments have been mostly in 1) infrastructure, e.g., providing models on company intranets; 2) pipelining tools for workflow or in hardware (shift away from SGI to clusters running Linux, GRID and on the cloud, etc.); 3) applying models for other areas, e.g., ADME/Tox (14–17); and 4) a philosophical mindset change in the recent development of an open chemistry development kit (CDK) (18), crowdsourcing (19), combinatorial model building (20), parallel assessment of many model algorithms/descriptor combinations (21,22), secure sharing of chemical information (21) and collaborations between groups. While some of the latter developments are in the early stages, we think they are positioned well to stand apart from what has existed before.

As we have perhaps the most experience with QSAR methods and descriptors, we will use this as our example throughout. From our experience using very large datasets (tens to hundreds of thousands), we see that the best models on average have a test set correlation of approximately 80% (a best case) (22). This is comparable to the predominantly *in vitro* data used to build the models, as these will have a cutoff for predictivity. As has been noted in a prior commentary, we have reached a peak (or at least a plateau) in prediction accuracy for blood brain barrier (BBB) modeling, and this is likely the case elsewhere (23). We therefore do not believe that the development of new descriptors (whether 2D or 3D), or even quantitative structure-activity relationship (QSAR) algorithms, is really needed for cheminformatics, as these will only provide incremental advances, although we place some caveats to this statement. There may be some difficult targets that are intractable to model with the currently available algorithms and descriptors (although to our knowledge no one has exhaustively tried to build ligand-based models for all known biological drug targets). Some have tried to focus on, for example, GPCRs but did not comment on the model quality for each protein (24,25), so there is still much to be done, and perhaps our collective energies can be focused on the following important issues.

THE UNIVERSE

While some methods like CoMFA may by their very nature limit you to molecular structures that are highly similar to the training set (local model), others may enable you to make predictions for compounds far away from the training set without providing the user with any reference to whether the model is interpolating or extrapolating (concepts we can all understand). While a model may be useful for making predictions for molecules close to the training set, this may not be the case for compounds that are far away in chemical similarity space. The uninitiated may have no idea of this limitation and have no concept of what they have stumbled into. Therefore, the area of concern to most using a predominantly ligand-based computational model, e.g., QSAR or pharmacophore, is answering the question "Is the prediction reliable?" Providing the user with some confidence in the prediction has only been partially addressed by the efforts at providing an applicability domain using Tanimoto similarity, PCA, clustering, Mahalanobis distance, etc. (28–36).

Generally, any QSAR model will either be local (narrow structural diversity to one chemotype) or global (diverse array of molecules), but even the latter case will just have coverage of a small fraction of chemical space. What may be needed is some anchoring or benchmarking of a model based on the known chemical space (measured using some physicochemical or substructure descriptors) of a set of drugs, chemicals or reference databases, e.g., taking a ChemGPS-type approach (26). This would provide the user with only the coverage of a reference space for their model and goes only part way to providing some confidence. Model quality based on testing data could be another component; prediction probability based on the QSAR algorithm used provides a further dimension, while a consensus across all these approaches could be a stronger measure than any in isolation. We feel there is still an immense unmet need and opportunity here to develop standards for the applicability domain of a model and when predictions should be avoided.

THE BLACK HOLE

Another major issue that cheminformatics has failed to address is that current computational chemistry software companies have generally catered to the computational

modeling community and have not done well in translating their tools to bench biologists and chemists (in comparison to some of the bioinformatics tools like BLAST searching that are widely used by non-bioinformaticians). Subsequently, in most cases, you have to be an expert to use most computational chemistry tools, with prior knowledge of what a method can and cannot do. These tools do not teach you as you go. Future cheminformatics tools have to consider their audience before assuming that any scientist will use them; the barrier to entry has to be as low as using Google, Facebook and Twitter, etc. Admittedly, this may be a lot to ask because these tools are very general in nature and have to appeal to non-scientists, whereas cheminformatics is much more specialized. However, if cheminformatics is to spread its user base, it is essential the tools become used by non-experts with minimal training. Methods should ask the user what they want to do, then provide a path to achieving their aims. The software complexities should be translated in a way that is understandable by anyone regardless of whether or not they have any prior computational experience.

Allied with this is the interpretability of a model prediction. Part of the problem with some QSAR approaches is that a model output is not inherently understandable. While the models may be black boxes, the outputs could be thought of as black holes (taking time and resources away from other projects), as they are not widely embraced. There have been efforts made at ADME data visualization (27–32), while expanding these approaches to show outputs from multiple computational models in a color-coded or symbolic manner may be preferable. Again, development of truly novel, simple and interpretable data visualization methods capable of handling the massive growth in experimental and computational data is long overdue.

THE RED DWARF

Having identified some long unresolved issues which we still face, what can each stakeholder do (industry, academia, software developer, etc.)? As the industry is changing so rapidly, this places the software providers in an undeniably difficult position, so what can they do? They might want to ask their customers before developing some new software or software suite. Building it will not guarantee that people will buy it. Please do invest in R&D. For a healthy cheminformatics field, we all need to support educational and training efforts to bring in more talented minds (e.g., University of Sheffield, UK http://www.shef.ac.uk/is/prospectivepg/courses/chem/index.html and Indiana University, USA http://cheminfo.informatics.indiana.edu, etc.). Generally, innovation happens at the intersection between different research fields; therefore, it is important to ensure that we bring in fresh blood from perhaps non-intuitive fields. This may lead to new uses in cheminformatics for older technologies from elsewhere. For example, we have seen the auto industry and pharma sharing best practices for manufacturing, e.g., just in time provision of product, etc. What would it take to bring in some of the data modelers, data miners, computer animators, interface developers (from other industries), defense industries or simulator designers to provide help with pharmaceutical-derived cheminformatics data visualization? Bringing together those skilled in chemistry and informatics or these other disciplines may be analogous to the field of systems biology (33,34), which tries to integrate those skilled in mathematics, engineering, biology, etc. It is important that there is a common language that connects researchers from different backgrounds to aid in integration (35). This may be academia's role, and the NIH, NSF and grant reviewers could ensure that they are funding at least some truly ground-breaking initiatives, as well as covering well-trodden ground. There may also be a role here for other mechanisms to get involved, such as the pre-competitive initiatives (36). More diversification of a portfolio of efforts may be a good thing to drive innovation in cheminformatics. While in the past the industry has been an innovator in computational chemistry software (e.g., Merck molecular forcefield (37)), can we expect this to be the case in future? In a financially constrained environment it is likely that such developments and risk will be borne outside the industry or at least shared in a collaborative manner (e.g., Pistoia Alliance). Already the open source cheminformatics data pipelining initiatives like KNIME (38) (http://www.knime.org) are nipping at the feet of the commercial software tools. Cheminformatics companies need to "innovate or die" and "innovate quickly" before they become red dwarfs (see the recent merger between Accelrys, Inc. and Symyx Technologies, Inc. (http://accelrys.com/about/news-pr/0410-announcement.html). We think the cheminformatics industry could learn much from "the innovators dilemma" (39). There are still significant holes. For example, there has been very limited development of open pharmacophore tools to our knowledge (http://pharmacophore.org), and this could do with some attention. Mobile computing devices present a new frontier (and business opportunity) with constraints in how much can be shown on very small screen real estate, which might drive cheminformatics software developers to consider how they expose their tools to new users (40) in the pharmaceutical industry or academia.

TECHNOLOGIES TO NAVIGATE THE CHEMICAL UNIVERSE

What computational technologies could be combined or applied in a new way to provide the catalyst for the field that would enable us to navigate the chemical universe?

What could be created from a mashup of technologies, hybrids or truly new tools? For example, now we have a hybrid in which *in vitro* data and *in silico* approaches are used side by side for absorption, distribution, metabolism, excretion and toxicity (ADME/Tox) properties. As we had predicted in the past (15), it took between 5–10 years for ADME/Tox data in the industry to be of a scale useful for building reliable predictive computational models. In the industry, we now have tens to hundreds of thousands of datapoints for individual ADME/Tox assays used in modeling, depending on the company size and screening capability. How long will it be before we almost completely replace the *in vitro* models with computational models for absorption, metabolic stability, P450 inhibition and P-gp in the same way that we have for logP and solubility? Perhaps another 5–10 years, maybe sooner based on the developments we are seeing.

Integration of tools that impart some medicinal chemistry expertise (like DrugGuru (41), rule bases (42) or multiobjective library enumeration methods (43)), multiple QSAR models and multidimensional optimization methods (44) may help the user or go beyond making predictions for many properties with computational models and instead propose a synthetically reasonable alternative with improved properties that considers all options of interest. This certainly goes beyond highlighting a problematic part of a molecule, whether reactive (42) or toxic, etc.

We certainly cannot possibly have all the answers yet hope that by raising these challenges, in the future we (and the reader) can apply some new technologies that will greatly expand the influence of cheminformatics in pharmaceutical research and simultaneously provide increased confidence and improved interpretability of the outputs. This may be one small step for mankind but a very big one for cheminformatics.

ACKNOWLEDGMENTS

We are grateful to discussions with colleagues and Dr. Maggie A.Z. Hupcey for astronomical inspiration.

Conflicts of Interest SE consults for Collaborative Drug Discovery, Inc. on a Bill and Melinda Gates Foundation Grant#49852 "Collaborative drug discovery for TB through a novel database of SAR data optimized to promote data archiving and sharing."

REFERENCES

1. Kuhn TS. The structure of scientific revolutions. Chicago: The University of Chicago Press; 1996.
2. Williams AJ, *et al.* Free online resources enabling crowdsourced drug discovery. Drug Discovery World. 2009; Winter.
3. Ekins S, Williams AJ. Precompetitive preclinical ADME/Tox data: set it free on the web to facilitate computational model building to assist drug development. Lab on a Chip. 2009; in press.
4. Bingham A, Ekins S. Competitive collaboration in the pharmaceutical and biotechnology industry. Drug Discov Today. 2009;14:1079–81.
5. Cramer RD *et al.* Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc. 1988;110:5959–67.
6. Schneider G, Bohm H-J. Virtual screening and fast automated docking methods. Drug Discov Today. 2002;7:64–70.
7. Mason JS *et al.* 3D pharmacophores in drug discovery. Curr Pharm Des. 2001;7:567–97.
8. Sprague PW. Automated chemical hypothesis generation and database searching with Catalyst. Perspect Drug Discov Des. 1995;3:1–20.
9. Barnum D *et al.* Identification of common functional configurations among molecules. J Chem Inf Comput Sci. 1996;36:563–71.
10. Martin YC. 3D database searching in drug design. J Med Chem. 1992;35:2145–54.
11. Hahn M. Three-dimensional shape-based searching of conformationally flexible compounds. J Chem Inf Comput Sci. 1995;37:80–6.
12. Todeschini R *et al.* New molecular descriptors for 2D and 3D structures. Theory. J Chemom. 1994;8:263–72.
13. Willett P *et al.* Chemical similarity searching. J Chem Inf Comput Sci. 1998;38:983–96.
14. van de Waterbeemd H *et al.* Estimation of Caco-2 cell permeability using calculated molecular descriptors. QSAR. 1996;15:480–90.
15. Ekins S *et al.* Progress in predicting human ADME parameters in silico. J Pharmacol Toxicol Methods. 2000;44(1):251–72.
16. van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov. 2003;2:192–204.
17. Waller CL *et al.* Modeling the cytochrome P450-mediated metabolism of chlorinated volatile organic compounds. Drug Metab Dispos. 1996;24:203–10.
18. Steinbeck C *et al.* Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. Curr Pharm Des. 2006;12(17):2111–20.
19. Oprea TI *et al.* A crowdsourcing evaluation of the NIH chemical probes. Nat Chem Biol. 2009;5(7):441–7.
20. Zhu H *et al.* Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. J Chem Inf Model. 2008;48(4):766–84.
21. Kaiser D *et al.* Similarity-based descriptors (SIBAR)—a tool for safe exchange of chemical information? J Comput Aided Mol Des. 2005;19(9–10):687–92.
22. Gupta RR, *et al.* Using open source computational tools for predicting human metabolic stability and additional ADME/TOX properties. 2010, submitted.
23. Ekins S, Tropsha A. A turning point for blood–brain barrier modeling. Pharm Res. 2009;26(5):1283–4.
24. Keiser MJ *et al.* Predicting new molecular targets for known drugs. Nature. 2009;462(7270):175–81.
25. Keiser MJ *et al.* Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007;25(2):197–206.
26. Rosen J *et al.* Novel chemical space exploration via natural products. J Med Chem. 2009;52:1953–62.
27. Ekins S, Shimada J, Chang C. Application of data mining approaches to drug delivery. Adv Drug Deliv Rev. 2006;58:1409–1430
28. Maniyar DM, Nabney IT, Williams BS, Sewing A. Data Visualization during the Early Stages of Drug Discovery. J Chem Inf Model. 2006;46:1806–1818

29. Yamashita F, Itoh T, Hara H, Hashida M. Visualization of large-scale aqueous solubility data using a novel hierarchical data visualization technique. J Chem Inf Model. 2006;46:1054–1059
30. Yamashita F, Hara H, Itoh T, Hashida M. Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: application to structure-activity relationship analysis of cytochrome P450 metabolism. J Chem Inf Model. 2008;48:364–369
31. Stoner CL, Gifford E, Stankovic C, Lepsy CS, Brodfuehrer J, Prasad JV, Surendran N. Implementation of an ADME enabling selection and visualization tool for drug discovery. J Pharm Sci. 2004;93:1131–1141
32. Stoner CL, Cleton A, Johnson K, Oh DM, Hallak H, Brodfuehrer J, Surendran N, Han HK. Integrated oral bioavailability projection using in vitro screening data as a selection tool in drug discovery. Int J Pharm. 2004;269:241–249
33. Kitano H. Computational systems biology. Nature. 2002;420:206–10.
34. Ekins S *et al.* Systems biology: applications in drug discovery. In: Gad S, editor. Drug discovery handbook. New York: Wiley; 2005. p. 123–83.
35. Ermondi G, Caron G. Recognition forces in ligand-protein complexes: blending information from different sources. Biochem Pharmacol. 2006;72(12):1633–45.
36. Barnes MR *et al.* Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. Nat Rev Drug Discov. 2009;8(9):701–8.
37. Holloway MK *et al. A priori* prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. J Med Chem. 1995;38(2):305–17.
38. Patel H *et al.* Knowledge-based approach to de novo design using reaction vectors. J Chem Inf Model. 2009;49(5):1163–84.
39. Christensen CM. The innovator's dilema. Harvard Business School Press; 1997.
40. Williams AJ. Mobile chemistry—chemistry in your hands and in your face. Chemistry World. 2010; May.
41. Stewart KD *et al.* Drug Guru: a computer software program for drug design using medicinal chemistry rules. Bioorg Med Chem. 2006;14(20):7011–22.
42. Metz JT *et al.* Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. J Comput Aided Mol Des. 2007;21(1–3):139–44.
43. Gillet VJ *et al.* Combinatorial library design using a multi-objective genetic algorithm. J Chem Inf Comput Sci. 2002;42 (2):375–85.
44. Ekins S *et al.* Evolving molecules using multi-objective optimization: applying to ADME. Drug Discov Today. 2010;15:451–60.