J Comput Aided Mol Des (2012) 26:47–49
DOI 10.1007/s10822-011-9501-6

PERSPECTIVE

Computational chemistry and cheminformatics: an essay on the future

Robert Charles Glen

Received: 11 November 2011 / Accepted: 29 November 2011 / Published online: 13 December 2011

Abstract Computers have changed the way we do science. Surrounded by a sea of data and with phenomenal computing capacity, the methodology and approach to scientific problems is evolving into a partnership between experiment, theory and data analysis. Given the pace of change of the last twenty-five years, it seems folly to speculate on the future, but along with unpredictable leaps of progress there will be a continuous evolution of capability, which points to opportunities and improvements that will certainly appear as our discipline matures.

Keywords Drug discovery · Cheminformatics · Computer-aided molecular design · Simulation · Computers in chemistry

About 25 years ago, I was a physical chemist at the Wellcome Foundation (the drug company) and on a day like today, I would probably have been staring at a GT40 (unfortunately not the Ford, but with 8K word core memory and selling for "under $11,000") connected to a PDP-11 (a 16-bit minicomputer sold by Digital Equipment Corporation (DEC) from 1970 into the 1990s), transitioning software to a 32-bit VAX-11/780 connected to an Evans and Sutherland PS300. Not much change out of $200,000 and all in all, the wonder of the age but seriously inferior to your current high-end laptop. Little stick-figures of molecules danced on the screen and we pretended that these were something to do with drug discovery. The company was happy, as every so often, one of the calculations actually worked and an improvement, or even a discovery was made. The computer age was still beginning, as most people had never seen a mouse (except in the pet shop), and the moving image on the screen, under the control of the user, was magical to new users. People would jump in their seat when they moved the mouse and the image on the screen moved too. Especially with the (first ever) LCD shutter glasses from Leeds University (seemed to be made of plastic and chewing gum but gave incredible stereo). We even had a visit from the Minister of Science in the UK and had to explain that an enzyme (on the screen) 'was something in your washing powder that made clothes whiter'. Saying that we had derived an x-ray crystallographic structure of dihydrofolate reductase and it catalysed the reaction of dihydrofolate to tetrahydrofolate didn't seem to wash. Buying this equipment, at this early stage, was a leap of faith on behalf of Wellcome, I believe the first drug company to have molecular graphics and drug discovery software (courtesy of innovators Andy Vinter, David White and Garland Marshall) as part of the discovery process.

Sir James Black, Research Director and the doyen of pharmacologists, had persuaded the company that we should help discover new medicines using computers, and all this very expensive equipment duly arrived. We looked at the exciting images on the screen then he said "Well laddie, calculate the solubility of this analogue of 5-HT…". Oops, not as easy as it sounds and actually quite impossible then. Still, we did manage to discover some important medicines, notably in migraine (Zomig), with Alan Robertson (chemistry), Graeme Martin (pharmacology) and significant input from computer-based methods from Alan Hill (bioavailability) and pharmacaphore analysis (a strong interest of mine at the time). The key ingredient was teamwork, combining diverse areas of science and computer-aided design. There were tasks that could be performed on the computer (in this case pharmacophore analysis/visualisation/pattern

R. C. Glen (✉)
The Unilever Centre for Molecular Sciences Informatics,
The University Chemical Laboratory, University of Cambridge,
Lensfield Road, Cambridge CB2 1EW, UK
e-mail: rcg28@cam.ac.uk

recognition) that were not possible on traditional paper (the thin white material they used to make from trees).

Now, I have the latest 3-D desktop screen, connected via my high-end multi-processor computer to a grid of 5,000+ processors and a 256 core, 32 node GPU cluster with QDR IB and 128 Tesla GPUs. Wow, what a change! We now typically simulate protein complexes in solution, transmembrane receptors and complex biological networks. A recent publication shows millisecond simulations of drug docking using molecular dynamics (D.E. Shaw group). Our data sources are world-wide with semantically enriched knowledge. Unfortunately, I still can't calculate the solubility of that pesky 5-HT analogue (well maybe, if I use the crystal structure and wait a few months to complete the density functional calculations).

Which brings me to the future—and a few caveats. Tolkein said "The wise speak only of what they know"—so by necessity we extrapolate from current technology (no quantum computer yet). However, another sage (William Gibson) said that "The future is here. It's just not widely distributed yet". So given what we have, where will improvements of this technology lead us to?

If I had a seriously faster computer, high speed networks to all of human knowledge and algorithms that mimic reality, what would I do? In 25 years time we can probably be certain that Moore's law will continue, the quality of algorithms will improve and that data sources will increase dramatically. We are already seeing vast increases in data from robotics, electronic publishing sources, the huge increase in scientific output from the BrIC and other developing countries. So, here are three areas that I think will develop substantially to realise the potential of computational chemistry.

The first problem is actually finding the most appropriate data. The internet is unstructured, noisy and untrustworthy but it is also wonderful—if we have the right tools. Imagine if all the publishers could collaborate, the companies could contribute, crowdsourcing worked and the data was semantically richer (and trustworthy)—we would then potentially have much of the data we need available. The big challenge would be finding the information from this ocean of data. This will require advanced linguistic tools (e.g. find the answer to: "is there an antibiotic I can extract from grapejuice?"). Developments in Natural Language Processing for example, means that complex human-created documents can be processed by robots to extract useful information, and in particular concepts, that can be searched for or re-used by other robots. We can also imagine having an assistant (robot) which finds, filters and organises knowledge and is specific to our area of interest. It is also about starting from one point in information space and ending at another. Of course, in chemical information 'space' we don't know where we're going (often) so we go down lots of rabbit holes, never to emerge! I've come to the conclusion that chemical information is all about networks. There is an awful lot of it, it's embedded in a large complex network of journals, articles, blogs, books, peoples heads—and the tools to find the 'information' are inadequate. "Thus, science is not merely a set of ideas but also the flow of these ideas through a multipartite and highly differentiated social system" (Martin Rosvall and Carl Bergstrom). Self organising networks of agents could navigate through the data space (six degrees of Kevin Bacon) to quickly hone in on the desired information.

A second problem is how chemistry is represented in computers. Chemistry started with the alchemists, who were most concerned with secrecy and used a complex language of symbols representing opaque alchemical concepts. Those created by Jean Henri Hassenfratz and Pierre Auguste Adet to complement the Methode de Nomenclature Chimique (1787) are representative of this approach. In an enlightened break with the secretive 'alchemist' approach, Berzelius (1779–1848) suggested compounds should be named from the elements which made them up. Archibald Scott-Couper came up with the 'connections' between symbols, which gave rise to structural diagrams (1858). Frederick Beilstein is probably the father of modern 'cheminformatics'. In the Handbuch (1881) the naming of compounds was an integral part of their storage and retrieval from indexes. The obsession for efficient indexing dominated chemical information for the next half century (although, two modestly different compounds may appear similar but their properties may be fundamentally different—so we are often biased by our indexing methods). The introduction of computers led to line notations (to fit the technology of typewriters). So we can see that chemistry has gone through 200 years of reductionism, resulting in a science of materials being reduced (again) to a science of symbols. I think this has primarily been driven by hardware and algorithmic constraints. In the future, I see this trend being reversed as descriptions of materials become increasingly semantically rich, containing greater levels of structural and functional detail and in the end, will be a full and accurate representation of the material and its properties. A simple example is that the line drawing of a chemical structure (e.g. a polymer) will not describe its bulk properties in solution—but it is actually the properties of the material that we are most interested in.

A third area is of course simulation. I say of course, as a major goal of computational chemistry has typically been to replicate reality. Currently, petabyte computing capacity is available with Infiniband connect (300 Gbit/s) and millisecond simulations of real biological systems are possible with hardware specifically designed for simulation. The consolidation of these methods into materials design, drug discovery and other areas will undoubtedly happen (much the same as in engineering design of aircraft and automobiles). Going ahead, the scale and timescales

will increase along with hardware developments. The accuracy of simulations however is difficult to gauge, as (in chemical structure representation) we will always have to limit the complexity of the simulation to fit the hardware capabilities. Breakthroughs in algorithm development will be the key to both the accuracy of the calculations and the size and complexity of the processes that can be modelled. This is sometimes forgotten, when porting exiting methods to faster computers with the expectation that the accuracy of the results will also extrapolate. Sometimes doing longer more complex simulation shows up the deficiencies in the algorithms (or more importantly the sampling) to a greater extent. It has unfortunately been the case in our area that single experiments have been the norm and reproducibility of results is not always as straightforward as it should be (especially with proprietary software and hardware). With unlimited computing capacity, vast memory capability and virtualised environments, it could evolve to the situation where re-calculation and evaluation is more straightforward. I heard recently that IBM are developing 'neural' chips and with the emergence of the cloud, the super high-speed internet and greatly improved levels of theory, the questions themselves will change—and perhaps the machines themselves will start to ask their own questions.