ChemComm
www.rsc.org/chemcomm

Developing tools and standards in molecular informatics


Robert Glen

Let's put a 'toe in the water' of molecular informatics. There are 50 million or so accessible chemical substances, around 6 million available reagents, 7 million published chemical reactions, as well as nearly 16,000 protein X-ray crystal structures and 250,000 readily available small molecule X-ray structures. This is the tip of a large (and growing) information iceberg. One of the biggest challenges (and opportunities) in the chemical sciences today is how best to manage the mountains of data and information associated with compounds and their structures. Unilever and the University of Cambridge have set out to address this problem, in a unique partnership. The Unilever Centre for Molecular Informatics at the University of Cambridge is dedicated to the exciting new discipline of molecular informatics, under the leadership of Robert Glen, formerly Vice President of Collaborative Research at Tripos Inc. (St Louis, Missouri), a leading company in life sciences software. Professor Glen previously set up the Computer-aided Molecular Design group at the Wellcome Foundation; he is the co-inventor of the migraine drug, Zomig (AstraZeneca), and of two other compounds that have entered into Phase 2 clinical trials.

MOLECULAR INFORMATICS covers all aspects of scientific information involving molecular structures, whether this comes from experimental measurement, hypothesis, or observation. The field has grown from a synthesis of information and computational sciences[1,2] and spans the disciplines. It is not unusual to combine theory and experimental data from many different sciences and sources.

"Major resources have been directed to collect enormous amounts of data, and data collection has become more automated," explains Professor Glen. "This means that the volume and diversity of the data has increased, with the associated problems of quality control. There are now opportunities both to analyse the data in a much more systematic way (curation) but also to use it in a more inventive way." Among the beneficiaries of developments in molecular informatics will be the food, pharmaceutical and biotech industries. Bioinformatics[3], which is concerned mainly with genomic and protein sequence information, is fast becoming an essential component of drug discovery, but there is an interface between molecular informatics and bioinformatics, of which a prime example would be a potential drug binding to its protein target. One of the many projects underway at the Unilever Centre, in collaboration with Roche, GlaxoSmith Kline, and the Cambridge Crystallographic Data Centre, involves further development of the GOLD[4] (Genetic Algorithm for Ligand Docking) program, which was first put forward by Gareth Jones, Peter Willett and Robert Glen. GOLD is already widely used within the pharmaceutical industry to explore the binding of small molecules to target proteins. The aim now is to overcome previous limitations of the programme by building in flexibility for target protein structures, introducing more relevant scoring functions and speeding the method – which should allow accurate and rapid *in-silico* screening of millions of potential drug molecules.

Developing this and related molecular informatics projects is clearly going to be of great interest to an industry desperate to fill its drug pipeline.

"I believe that one of the big misconceptions of the pharmaceutical industry over the last 5 to 10 years is that if more data is collected then more medicines will be generated," says Professor Glen. "The advent of combinatorial chemistry and high throughput screening meant that the amount of data that could be collected has increased enormously. But what we have actually seen is a big fall-off in the quality of the data. High throughput screening methods were inaccurate and required a lot of follow up to establish leads. We also have an interesting situation in that high throughput chemistry provides many more molecules than are available from traditional synthetic methods (generating more 'hits') but this process is very inefficient, with typically 40 per cent of solvent-based reactions in high throughput mode not producing molecules of high enough purity to test reliably – and with real problems associated with the reproducibility of high throughput

DOI: 10.1039/b207793k

screening data. So what you're doing is collecting a tremendous amount of data, some of which contains valuable information, but much of which is destined for the rubbish bin. Also, due to the nature of the screening libraries used, the hit molecules discovered tended to be much larger (and more similar to each other) than in the past, resulting in little opportunity for lead optimisation".

One of the major challenges is to be able to produce data that is of high quality, and also to be able to filter the data in such a way so that the quality can be raised. "This is one of the major functions that I hope molecular informatics will contribute to in the near future – being able to harvest this data and to refine it using clearly designated and accepted methods; increasing its relevance to the problems being addressed," says Professor Glen.

Bioinformatics *vs* molecular informatics

But molecular informatics has a way to go before it catches up with bioinformatics in its widespread use and acceptability. "We've had an explosion of data analysis and interest in genomic projects because the basic information is available, much of it in accessible public databases" says Professor Glen. "But in molecular sciences, although much information is being gathered, very little of it will actually see the light of day."

This is because much of the data is collected in private by companies or is published by journals in an inaccessible form and is not available for use outside. The publishing process actually results in the widespread destruction of data. A simple example might be the submission of a molecule to a journal in connection table format as part of a paper. This contains much information on the compound and is computer searchable. The conversion of the structure to a gif or pdf format file loses this key facility. Commercial secrecy raises another problem. One of the best ways of validating data is to have scientists peer review it. Mistakes are often found when data is reviewed by scientists who are analysing and using it, but were not involved in its original production. If the data is confined within one organisation, then it will not be peer-reviewed and most mistakes will never be found.

"The pharmaceutical industry has been very short-sighted in setting up institutes that allow it to share experience and data in collaboration with each other, government and other organisations," says Professor Glen. "It would benefit everybody if such shared experiences could be pooled. For instance, information on simple properties of molecules like solubility, acid/base constants or partition coefficients could have enormous benefit in many research and development activities, but such things are seen as trade secrets, so they're not published."

Nor can the academic community escape responsibility for the large gap between bioinformatics and molecular informatics. In a typical chemistry department, many small groups work independently and do not produce common databases that are interchangeable, although they will access commercial databases, when they buy reagents, for example, or search the literature for reactions. "It is for the universities now to invest in their informatics infrastructures and catch up with what has been happening in industry for the past 20 years. The recent Grid initiatives in the UK are a step in the right direction" says Professor Glen.

Although molecular modelling began in universities, it quickly migrated to the pharmaceutical industry and small start-up companies. Since then, universities have not really encouraged the development of systems that allow the inter-operability of data nor the development and widespread sharing of data – except in the bioinformatics community. It is always difficult to find reliable data (particularly on the internet – as an example, try identifying the structure of the natural enantiomer of epinephrine) – one approach could be to identify trusted university departments where the data is created. A chemical database approach developed here is an index of university chemistry departments and chemistry journals (http://www.ch.cam.ac.uk/c2k/). This site is run by Jonathan Goodman, a lecturer at the Unilever Centre. Assembling such a database is not a major problem, keeping it up to date and reliable is the big challenge. This is achieved through automated procedures, which regularly revalidate the data collection. Another approach we are following is the establishment of tools and standards for the input, storage and retrieval of chemical data. A schematic of our 'world wide molecular matrix' (WWMM) concept which is under development, is shown below. This is a peer-to-peer system with annotated molecules having validated data from experimental and computational sources.

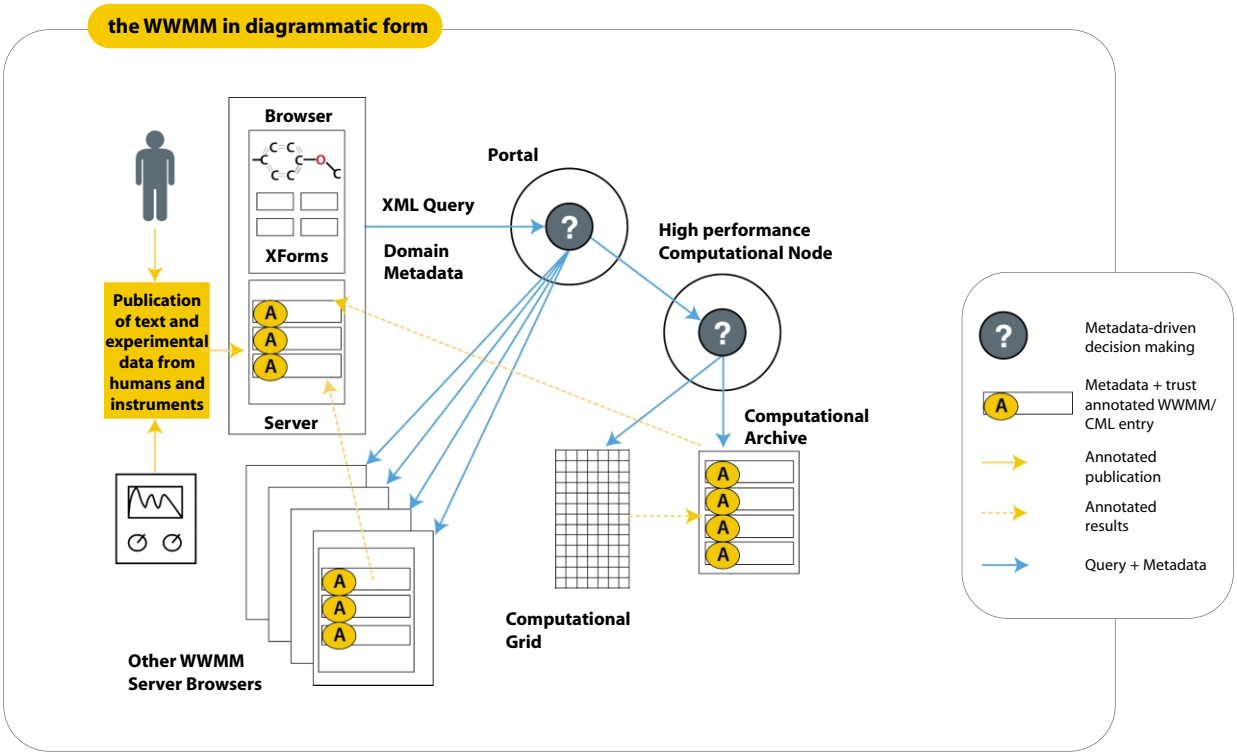
Tools and standards

The Unilever Centre is not creating its own database, but rather working to create tools and standards so that other molecular databases can be better accessed, managed, and understood. There is much emphasis on a new language called XML, and its derivative for the chemical community, CML[5] (Chemical Markup Language) which has been developed by Peter Murray-Rust, formerly of GlaxoWellcome and currently a Lecturer in the Centre, and Henry Rzepa of Imperial College. This allows the generation of molecular ontologies (which are the 'semantics and grammar' of molecules) and unambiguous dictionaries which better define the relationship between different types of molecular data. This also allows a clearer definition of the actual nature of the data (how the data was collected, its units and its relationship with other data) and the uses it can be put to. This represents a vast improvement on the current situation, where there over 50 different formats for storing molecules, all containing different kinds of data and all incompatible with one another.

What CML does is to unambiguously define a molecule so that everybody who gets a copy in XML (or even a robotic system accessing the data) knows exactly what they're getting – at least as far as people can agree on how to describe a molecule on a computer. Of course, one of the fundamental problems we face is that chemists usually deal with an abstraction of a molecule that is very far from reality. A 'real' molecule is best described by a complex wavefunction, usually too complex to compute for even the simplest of molecules, and it exists in an environment which may alter its properties (*e.g.* in water). The chemist will typically view a structure of a molecule as being made out of atoms connected by bonds. It is a static structure and it doesn't really include all the properties of the molecule (conformation, electronic structure, polarisation *etc.*). In CML, it is possible to come to an agreement on how to define an abstraction of a molecule for a particular use (the underlying 'meta data'). The key

Examples of Bio/chemo - informatics

Biological Data	Chemical Data	Patents
• Over 23 billion nucleotides bases	• 34,000,000 chemical substances	• European patents – 150,000,000 pages
• Over 800 organisms	• 3,700,000 chemical reactions	• 150,000 applicants/year
• 18838 protein x-ray crystal structures	• 613, 000 available reagents	• 400,000 chemical patent hyperstructures
• 12 million citations medline	• Bielstein has 600,000 reaction abstracts	• Over 100 countries in patent cooperation treaty (PCT)
• 20 mainstream databases from EBI	• 250,000 organic x-ray structures	
• Ensembl: 22,980 gene predictions	• Combichem libraries of **billions** of compounds	



here is the use that the information will be put to. The simple example Professor Glen gives is in the area of molecular similarity. A scientist specialising in the chemistry of dyes may regard two red substances as being similar as they are both red, while a biochemist may say they are different because they behave different in biochemical assays.

The annotated description of molecules is a significant step forward in describing a molecule in a way that makes the description more reliable. This, of course, will change over time – after all, the way we look at molecules now is very different from how they were viewed 100 years ago.

Such tools should help the Centre work towards developing new standards for chemical information which will help chemists to gain better trust of their data sources. Professor Glen wants to address how to describe molecules in a more unambiguous way so that the origin of the molecule and its associated data is trusted and chemists can be sure the data has not been interfered with or modified – in short, that it is still applicable to its intended use.

To this end, the Centre is working with learned societies and international bodies such as IUPAC, RSC, NIST, FDA and WHO with the objective of creating a set of common standards for describing molecules.

Similar developments are already occurring in many fields including for example, crystallography – where researchers are increasingly using the crystallographic information file (mmCIF) format as a common standard.

Meanwhile, the Cambridge Crystallographic Data Centre (CCDC) next door to the Unilever Centre, is a good role model, having made crystallographic data available to the community for the last 20 years in a validated form. However, one of the problems it faces is that although they have carefully abstracted the data and presented it in a very high quality format there are just 250,000 structures in the CSD, compared to around a few million small molecule structures that could be available, if only they were published. "Ways have to be found to allow people to publish the data and make it available generally, as it is in the bioinformatics community," says Professor Glen.

In the future, he is looking for the production of two levels of chemical data and information. First, there would be raw data – straight out of the lab – which can be made available to the general community. Second, there would be an abstracted form of data, where a trusted organisation, such as the RSC, harvests the data, verifies it and makes it available in a marked up format such as XML. Further steps would include further abstracting and curating the data, and increasing its quality and reliability, which should be left to organisations like CCDC or NIST. The difficulty in the past has been in making the raw data available – especially from many highly productive companies that are producing excellent data all the time and then squirreling it away.

Professor Glen returns to the current problem in molecular sciences – in sharing data, sharing empirical experience and in establishing standards. "We have a role in all three of these areas in encouraging scientists to at least publish the raw data more fully and giving them the tools to do that." The Unilever Centre is currently working with the RSC to create tools that will make it easier for scientists to put their data into a journal format that will result in enhancement of content. The tools will also check data upon entry so there will be fewer of the more common mistakes, such as spectroscopic data which are not self-consistent or mis-drawing of structures.

Professor Glen concludes: "We have to be very ambitious and become the institute that sets the standards in molecular informatics. We also want to develop methodologies that move the boundaries of how molecular informatics can be used and applied to ever wider areas of scientific interest."

Robert Glen was talking to Susan Aldridge

REFERENCES

1. *Molecular Modelling: Principles and Applications* (2nd Edition) by Andrew R. Leach Prentice Hall; ISBN: 0582382106.
2. *Three-Dimensional Chemical Structure Handling (Computers and Chemical Structure Information Series, No. 1)* by Peter Willett. John Wiley & Sons; ISBN: 047193108X; (October 1991).
3. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, Second Edition by Andreas D. Baxevanis (Editor, Wiley-Interscience; ISBN: 0471383910; 2nd edition (April 6, 2001).
4. G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *ACS Symp. Ser.*, 1997, **719**, 271.
5. G. V. Gkoutos, P. Murray-Rust, H. S. Rzepa, C. Viravaidya and M. Wright, *Internet J. Chem.* (www.ijc.com), 2001, **4**, 12.