OBC
www.rsc.org/obc

New horizons in molecular informatics

A collection of papers in this issue of *Organic & Biomolecular Chemistry* tackle an issue familiar to today's scientists who have the good fortune to have at their fingertips a great proportion of the world's knowledge of chemistry—through a few simple keystrokes of their desktop computer. There is a feeling that we are only scratching the surface of one hundred years of chemistry, and allied sciences, that is increasingly appearing in curated, easily accessible databases through the development of new forms of data abstraction and analysis. Indeed, information is now being generated at an unprecedented rate through the automated production of data (as is seen in the biosciences, in particular the human genome project, or in chemistry *via* robotic synthesis of large libraries of compounds, or the automated production of spectra) and through the sheer numbers of scientists at work today, pooling information and ideas through publication and the internet.

One article in this issue traces chemical information methods from earliest times to the present day and to some extent also shows how chemical knowledge, as it has developed, has shaped the chemical language of that knowledge. A very interesting aspect of this is 'structure as an iconic vernacular'—*i.e.* the way chemists communicate structure *via* chemical diagrams requires an external representation (the structural diagram), but also a degree of chemical perception, required to extract the chemical 'knowledge' from the representation. This of course causes problems in computer representation of structure and data—the computer can store the facts, but does not have the contextual framework to interpret the data. This is the subject of a number of articles that describe the meta-data or ontologies which are required to achieve this synthesis of data and the background knowledge of the data.

In molecular informatics, we are seeing a process that started from the communication of chemistry through alchemical symbols, principally designed to obfuscate chemistry, to repositories of chemical knowledge such as Beilstein, CAS or the CSD, or indeed the online availability of journal papers, mature into a source of information that is enabling a new approach to science. The availability of such volumes of data allows researchers to perform informatics experiments within the data space itself, so suggesting new experiments to verify deductions made from the discovery of new scientific principles within the data space. This approach turns on its head the standard scientific paradigm of hypothesis, experiment and theory. This of course brings about the danger of creating a disconnect between theory and experiment, so we should be aware of the insight from Henri Poincaré:

"Science is built up of facts, as a house is built of stones; but an accumulation of facts is no more a science than a heap of stones is a house".

That said, the opportunities for discovery, particularly in those areas where the traditional sciences overlap, are such that the results can be greater than the sum of the parts. This is particularly true in the biosciences, where the overlap between biology and chemistry is fundamental to understanding the mechanistic basis of biology. Examples in this area abound, and are exemplified by a number of articles in this issue.

These address the deeper understanding of biological processes that can be obtained from simulation. There is probably no more fundamental problem of interest to chemists in the biology sphere than the transformation of molecules by enzymes. This phenomenon, which underpins all of biology, has a large amount of associated data in many diverse and disconnected databases. The familiar Boehringer Mannheim chart of biochemical processes is an early 'systems biology' approach to the integration of these data. This is moving ahead, and is described here, with computer representation of molecules, their reactivity, properties and structure connected to metabolic cycles and ancillary data on their biological significance. The underpinning technology brings together many chemical concepts on structure and reactivity developed over the last twenty years coupled to modern hyperlinked data access on the internet and ontologies that allow conceptual searching of the data. This is also complemented in the genomics sphere with meta-data descriptions of biology such as Gene Ontology, a markup language which is a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. Another example is the BioSimGrid which is an example of a new approach to database technology which addresses the dynamic properties of biomolecules. The field of molecular dynamics has matured in recent years and offers insight into the mechanisms of molecular transformations and molecular recognition. To some extent, these approaches have been driven by the enormous increases in computer power, in which Moore's law has been seen to operate with processing power approximately doubling every two years. Since these simulations are voracious users of computer power, the opportunities to simulate ever larger and more complex systems is possible. This leads to the problem of how to handle large and complex datasets. The approach is to develop an ontology, meta-data and analysis tools which enable scientists to deposit their data and perform analysis of data along with many other similar simulations.

Another approach to address the requirements for ever greater computer power is used in the ligand/protein docking problem in which the internet and home computers have been harvested to interconnect the world's computer resources. Indeed, it appears that this approach, to a highly parallel problem, offers at this point more computer power than can be utilised, an interesting situation and fundamentally different to the present state of molecular dynamics problems. This approach will undoubtedly be exploited increasingly as algorithms and problems amenable to this approach are tackled.

The use of the internet as a global repository of data has been to some extent held up by a lack of standards in the exchange and re-use of data. The use of CML, chemical markup language (and other markup languages) will allow more thorough exploitation of these resources. Indeed, our concept of a 'world wide molecular matrix' of computation and data will only be possible through the use of standards in both data generation, curation, and analysis. This will lead to the creation of a 'semantic web', in which the concentration is on data quality, reusability and knowledge extraction. Crystallography has led the way here with standard machine readable formats such

DOI: 10.1039/b414850a

as the macromolecular Crystallographic Information File Format (mmCIF) and associated computer programs which check the quality and consistency of data from X-ray crystallography experiments. One article in this issue extends this focus on quality to authoring tools for organic chemists in a project in collaboration with the Royal Society of Chemistry to read and check submitted journal articles. The use of natural language processing and chemical knowledge can be used to increase the quality and consistency of results reported in organic synthesis papers. This approach is also amenable to the post-processing of data to extract knowledge which is placed in an ontological framework leading to greater re-use and consistency of use of data.

The focus on the re-use of chemical information and knowledge of course would be easier if the capture of that data and knowledge contained the maximum amount of information initially. Much recent interest in this problem has focussed on the creation of a useful electronic laboratory notebook to replace paper alternatives that have been with us (tried and tested) since earliest times. The use of the paper lab notebook, is simple, concise, easily transportable *etc.* However, in the informatics age, it is unfortunate that much early data has been lost by the selective capture of data, completely open to the personal whim of the scientist performing the experiment. Careful experimental design coupled with helpful computer based applications are now allowing a more complete capture of experimental data to be performed. This has developed in recent years with more machine data capture methods available (including *e.g.* digital photography of a tlc plate as opposed to a drawing or the storage of connection tables for molecules—which then allows exercises such as sub-structure searching). These new approaches of course have to satisfy intellectual property and legal requirements as well as being simple and inexpensive to use before wider adoption is seen. However, I believe that this will eventually be driven by the loss of opportunity for the re-use of data and by the needs of the regulatory authorities. Indeed, safety aspects of experimental design are being incorporated into virtual laboratory notebooks and will be an additional driving force to their adoption.

Looking into the future, chemists will increasingly depend on (as opposed to use) computer technology in experimental design, data capture and analysis. The quality of data, its analysis and its re-usability will become major issues in the exploitation of an increasingly pervasive and comprehensive sea of data. This will open up opportunities for the discovery of new chemistry, not just by chemists, but by chemists alongside their robotic companions.

This issue of *Organic & Biomolecular Chemistry* is published in advance of an RSC one-day symposium on "New Horizons in Molecular Informatics" on December 7th 2004 in Cambridge. Further details may be found at www.mmsconferencing.com/informatics.html. We encourage you to attend to hear more about the issues and developments described in this issue.

Robert C. Glen