Recent Developments In Chemoinformatics Education

Helen Schofield, Gary Wiggins and Peter Willett

Abstract

Chemoinformatics techniques are increasingly being used to analyse the huge volumes of chemical and biological data resulting from combinatorial synthesis and high-throughput screening programmes. Scientists with both the chemical and the computing skills required to carry out such analyses are currently in very short supply, this resulting in the establishment of MSc programmes for the training of chemoinformatics specialists.

Contact Details

Helen Schofield. Chemistry Department, UMIST, Sackville Street, Manchester M60 1QD, UK. Tel: +44 (0)161 2004468. Fax: +44 (0)161 2004559. Email: helen.schofield@umist.ac.uk

Gary Wiggins. Chemistry Library, Indiana University, 800 E. Kirkwood Avenue, Chemistry Bldg C003, Bloomington, IN 47405-7102, USA. Tel: +01 812 8559452. Fax: +01 812 8556611. E-mail: wiggins@indiana.edu

Peter Willett. Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK. Tel: +44 (0)114 2222633. Fax: +44 (0)114 2780300. Email: p.willett@sheffield.ac.uk

Development of chemical information systems and services

Ever since the beginning of "modern" chemistry in the 19th century, chemists have been prolific in their production of information. Publishers of chemical information have been faced with greater challenges than those in most other disciplines, due to the special language of chemistry which focuses on chemical structures and reactions. As a consequence, sophisticated printed secondary chemical information systems were developed to organise the primary journals and patents and aid information retrieval by the practising chemist, who in turn was prepared to invest time and effort in learning to use these systems effectively.

Inspired by the guaranteed customers of the chemical and pharmaceutical industries, chemical information publishers were quick to embrace computer technology in the 1960s, initially with a view to streamlining the production of their printed sources. However, it was soon realised that in addition the new technology could provide search functionality, not just to printed text but also to the structure diagrams and reaction schemes that comprise so much of the chemical literature. This realisation led to the development of increasingly sophisticated chemical information retrieval systems, such as the Chemical Abstracts Registry File[1] (searchable since the 1970s) and the Cambridge Structural Database[2]. The first desktop systems were pioneered by MDL Information Systems Inc.[3] and based on their MACCS software, which has been used by organisations for managing in-house chemical information since 1979. The full impact of desktop systems was felt when the Beilstein CrossFire[4] (with coverage to 1771) and Chemical Abstracts Service's SciFinder[5] (coverage now back to 1947 and due to be further increased back to 1907 in 2002) services became available in the mid-1990s. Desktop systems enable researchers to have access to information directly with no metered charging for online time, searches or displays.

The latest developments include availability of full-text electronic journals, web interfaces for chemical information applications, enhanced three-dimensional structure and reaction searching, and electronic archiving of old (pre-1960s) information. Thus the transition of medium from print to electronic for storage and retrieval of chemical information is nearing completion.

Chemoinformatics: why now?

Since chemical information systems have been available, first in printed and then in computer form, for many years, one may well ask why the current interest in chemoinformatics. Indeed, one might reasonably ask exactly what is meant by chemoinformatics, as terminology is still variable, with a search of the literature or a quick surf of the Web revealing references not just to chemoinformatics but also to cheminformatics and chemiinformatics, as well as to related phrases such as chemical informatics, chemical information science and molecular informatics.

Perhaps the first formal definition of chemoinformatics was provided by Frank Brown[6] who stated that "The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is 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 area of drug lead identification and organization" It will be clear that this definition ties chemoinformatics very firmly to the pharmaceutical industry and to the process of drug discovery. A more wide-ranging definition is provided by Greg Paris, as quoted by Wendy[7]: "Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information". We believe that this latter, more encompassing definition, better describes the contribution that chemoinformatics makes to modern chemistry (although it is certainly the case that pharmaceutical research - and related specialty areas such as agrochemicals discovery - will continue to provide the driving force for much of the methodological development of the subject). This contribution is evidenced by the increasing frequency of coverage of computer-based topics not just in specialist journals but also more generally in the chemical literature, *e.g.*, *Chemical & Engineering News*.

So why chemoinformatics now? There are several reasons for this. Firstly, and most importantly, the technological developments in combinatorial synthesis and high-throughput screening have brought about an increase by several orders of magnitude in the volumes of data that need to be processed in drug discovery programmes. Secondly, this explosion of both structural and bioactivity data has further hastened the need to integrate two areas of chemical computation that had previously developed, in large part, separately. Chemical information techniques have been developed for the storage and retrieval of information from databases of chemical articles and chemical structures, both corporate and public. The computer processing required for such systems is relatively simple in nature, although extremely impressive in terms of the data volumes involved (hundreds of thousands or millions of molecules). Molecular modelling techniques, conversely, have traditionally been used for the detailed analysis of datasets that contain a few tens, or at most a few

hundreds, of molecules, with the aim of using knowledge about their conformations and energies, *inter alia*, to predict their biological activities. Extending these methods for analysing structure-activity relationships to data volumes typical of those routinely handled in chemical information systems is a data mining challenge that is now being faced by most drug discovery organisations. Thirdly, there is no doubt that informatics is an idea whose time has come, or is coming, in an increasingly wide range of disciplines. Bioinformatics is, of course, now a widely recognised discipline, the establishment of which has been driven in large part by the data explosion resulting from the Human Genome and related sequencing projects. Medical informatics and health informatics are also well established and references are starting to appear to, *e.g.*, educational informatics and neural science informatics.

Educational requirements

Despite the fact that nearly every field of modern chemistry relies on the ability to use information technology in one form or another, each field tends to focus only on those aspects of chemoinformatics that are of most importance to them. We have already mentioned the significant contributions of chemical information specialists and of molecular modellers, but there are also other aspects of chemistry that require sophisticated use of chemoinformatics, *e.g.*, the storage and manipulation of spectra and other numeric data, or the increasing use of visualisation techniques for data mining in high-dimensional chemical spaces. While highly welcome, this widespread underpinning of chemical research means that chemoinformatics can fail to be recognised as a sub-discipline in its own right. As a result, there has been a lack of the rigorous academic courses that characterise other chemistry sub-disciplines and that could provide a steady output of graduates with the chemoinformatics skills required by industry.

This shortfall was recognised in the UK by the Engineering and Physical Sciences Research Council[8] (EPSRC), who sent out a request for proposals at the end of 1999 for funding for development of MSc courses under their "Masters Training Package" (MTP) programme. EPSRC claim that: "The MTP System will enable universities to meet better the changing needs of students, employees and employers. Innovation in the provision of training at this level is considered the key issue. MTPs can include full or part-time training and/or continuing professional development (CPD) including.....courses leading to appropriate postgraduate qualifications (MSc and/or MRes) at the Masters level.....Training Packages are awarded in areas for which EPSRC support is considered especially critical. Such as: innovation in provision, including anticipation of new areas of demand and new forms of provision; interdisciplinary and multidisciplinary study, which might need particular nurturing; gaps in subject coverage and the emergence of new subject areas."

Chemoinformatics was one of the key areas identified where proposals were sought by EPSRC. Both UMIST and the University of Sheffield independently recognised that they had the expertise available to mount MTP courses in chemoinformatics, responded to the call for proposals and were subsequently granted substantial funding for development of the new courses.

In the United States, no national initiative such as EPSRC's was available to spur the development of chemoinformatics. Nonetheless, an effort has been underway for several years at Indiana University to create a masters-level program in the subject. Indiana University has always been an early adopter of computer-based chemical information systems; for example, one of the first current awareness systems based on the Chemical Abstracts tapes was developed here in the late 1960s. With an awakening interest in informatics in general, Indiana University recently created its first new school in three decades, the School of Informatics.[9] The new school, which has admitted the inaugural group of graduate students for the fall 2001 academic year, offers five master's programmes: Media Arts and Science; Human Computer Interaction; Health Informatics; Bioinformatics; and Chemical Informatics.

Content of the programmes

The content of the Indiana, UMIST and Sheffield programmes is detailed elsewhere[10], and the descriptions that follow hence summarise just the main features. The overall programme structures are very similar (with the exception that that at Indiana lasts for two years, as against just a single year for the two UK programmes). Thus, all are at the MSc level, with students being assumed to have a first degree in chemistry or a chemistry-related subject, so that they can understand the chemical concepts that underpin much of the material presented to them. All three have a first part that involves a set of both required and elective taught modules, these including both chemistry-focused and informatics-focused modules and with each having associated tutorials, workshops and course work; and a second part that involves a research project or internship that leads to the presentation of a dissertation. There is, moreover, a fair measure of agreement in terms of actual content, although the relative amounts of chemical and informatics material does vary across the programmes; for example, the Sheffield programme is based in that University's Department of Information Studies and thus provides a strong

informatics focus with courses that, *e.g.*, discuss not just chemical but also textual and numeric database systems.

Indiana University The Indiana programme starts in September 2001. There are two required introductory courses that are common to all of the School of Informatics graduate MSc programmes (with the exception of the Media Arts and Science program): Introduction to Informatics; and Information Management. There are then two further modules that are required for all students on the MSc in Chemical Informatics program: Chemical Information Technology; and Computational Chemistry and Molecular Modeling. The taught-part of the programme is completed by a range of elective courses, drawn from those offered within or outside the school, to round out the students' education, such as: Algorithms Design and Analysis; Bioinformatics: Theory and Application; Bioinformatics in Molecular Biology and Genetics: Practical Applications; Chemical Instrumentation; Introduction to Human Computer Interaction; and User Interface Design for Information Systems.

Since the chemical informatics programme (and the bioinformatics programme) are being offered at two campuses (Bloomington and Indianapolis) of Indiana University, considerable effort is being made to keep the curricula uniform at both locations. Videoconferencing and other distributed education techniques will be used to share guest speakers and expertise in a broad range of topics.

UMIST The UMIST programme will initially run as a one-year full-time course, commencing in October 2001. The taught part requires the students to take the following modules: Chemical Information Sources; Chemical Informatics Applications; Computer-Aided Molecular Design 1; Computer-Aided Molecular Design 2; Database Design and Programming; Fundamentals of Bioinformatics; Research Methodology and Feasibility Study (which is part of the dissertation component); and Spectroscopy and Drug Discovery; Students complete the taught part of the course by taking two of the following elective modules: Algorithm Design for Chemical Problems; Combinatorial Chemistry; Knowledge Management; and Management of Intellectual Property. Most course units will originate from the Chemistry Department, but some will be delivered by the Biomolecular Sciences Department and by the School of Management. The course will also be enhanced by guest speakers from organisations involved with chemoinformatics, either information provision software or those with practical experience of implementation of chemoinformatics solutions.

In view of the objective of EPSRC to make masters level training courses attractive to people already in employment, the aim at UMIST is to convert the programme from a traditional MSc, obtained by full-time attendance at the university, to a series of distance learning modules. Under this arrangement, students may be expected to spend a few days at UMIST for each module, but with the bulk of the work being done from home or place of employment. As at Indiana, videoconferencing and other distributed education techniques will be employed to facilitate this. Units may be built up over a period of several years and when sufficient have been passed the student will be eligible to proceed to a research project leading to a dissertation in the same way as a full-time attendee. Although dissertation opportunities will be provided at UMIST, it is anticipated that the majority of both full-time and distance-learning students will undertake their dissertations at their place of employment or other industrial organisation.

University of Sheffield The Sheffield programme welcomed its first cohort of students in September 2000. The required modules here are: Chemoinformatics I; Chemoinformatics II; Computer Programming I; Computer Programming II; Information Storage and Retrieval; Information Systems Modelling; and Molecular Modelling. For their single elective module students choose from: Database Design; Human-Computer Interaction; and Multimedia Information Systems. Most of these modules are given by the Department of Information Studies but there are contributions from other departments, most noticeably from the Departments of Computer Science and of Chemistry for the two programming modules and the Molecular Modelling module, respectively.

The programme has been designed with the support of a consortium of organisations that includes examples of most of the major employers of people with chemoinformatics expertise: pharmaceutical and agrochemical companies; chemical software companies; and chemical database producers. Members of these organisations (currently including AstraZeneca, Barnard Chemical Information, Cambridge Crystallographic Data Centre, ChemWeb Inc., Eli Lilly, GlaxoWellcome, Merck Sharpe and Dohme, Novartis, Pfizer, Syngenta and Tripos Inc.) provide lectures for the two chemoinformatics modules and, most significantly, projects and funded student dissertation placements during the summer of the one-year programme: the students are now (July 2001) working on-site on their dissertation projects.

Conclusions

The emergence of chemoinformatics as a distinct sub-discipline of chemistry has spurred the development of high-level educational programmes to provide students with skills that are currently in high demand in industry.

It is hoped that the contacts already established between the programmes at Indiana, UMIST and Sheffield can grow into further co-operative activities. The taping or broadcasting of lectures in special areas of expertise, the development of canned instructional modules, or perhaps even exchanges of students are all possibilities. One particular area of interest is exploring the possibilities of distributed education, given that the Indiana programme is separated from the others by an ocean and many time zones.

References

1. This database is produced by Chemical Abstracts Services, at http://www.cas.org. See also: Fisanick, W. *et al*. (1998) Chemical Abstracts Service Information System. *Encycloped. Comput. Chem. 1,* 277-315.
2. This database is produced by the Cambridge Crystallographic Data Centre, at http://www.ccdc.cam.ac.uk/
3. MDL Information Systems Inc. is at http://www.mdli.com
4. This database is produced by the Beilstein Institute, at http://www.beilstein.com/beilst_2.shtml. See also: Heller, S.R. (1998) *The Beilstein System: Strategies for Effective Searching*. American Chemical Society: Washington DC; Meehan, P. and Schofield, H. (2001) CrossFire: a structural revolution for chemists. *Online Inf. Rev.* in press.
5. Ridley, D.D. (2000) Strategies for chemical reaction searching in SciFinder. *J. Chem. Inf. Comput. Sci. 40*, 1077-1084; Schwall, K. and Zielenbach, K. (2000) SciFinder: A new generation of research tool. *Chem. Innovat. 30(10),* 45-50. Also at http://pubs.acs.org/subscribe/journals/ci/30/i10/html/10ziel.html
6. Brown, F.K. (1998) Chemoinformatics: what it is and how does it impact drug discovery? *Ann. Report. Med. Chem. 33*, 375-384.
7. Warr, W.E. (1999) Paper presented at the 218TH ACS National Meeting, New Orleans, Aug. 22-26.
8. The Engineering and Physical Sciences Research Council is at http:/www.epsrc.ac.uk
9. Indiana University School of Informatics is at http://informatics.indiana.edu and http://www.informatics.iupui.edu/
10. Details of the Indiana University MSc in Chemical Informatics are at www.indiana.edu/~cheminfo/informatics/ci_iu.html, of the UMIST MSc in Cheminformatics at http://www.umist.ac.uk/chemistry/MScChemInf.htm, and of the University of Sheffield MSc in Chemoinformatics at http://www.shef.ac.uk/~is/courses/pgrad/mscci/mscci.html