Education for chemoinformatics

Peter Willett, University of Sheffield

Chemistry has for long been recognised as one of the most information-rich disciplines: the first specialist chemical journal appeared as long ago as 1778 and the principal abstracting service, Chemical Abstracts Service, will be celebrating its centenary in 2007. Computer tools for the processing of chemical information, in particular information pertaining to the structures of molecules, have been available for many years [1, 2], but it is only in the last few years that chemoinformatics, "the application of informatics methods to the solution of chemical problems" to quote the definition of Gasteiger [3], has established itself as a key component of many aspects of modern chemical research. This has happened principally as a result of developments in the pharmaceutical and related industries, where the introduction of combinatorial chemistry and high-throughput screening has resulted in a massive growth in the volumes of chemical and biological information available in research programmes [4, 5].

Interest in chemoinformatics is now becoming widespread, but this greatly increased exposure has highlighted the fact that there are very few people with high-level chemoinformatics skills. The principal source of such individuals in the past has been doctoral students and post-doctoral staff who have spent time in one of the few academic groups world-wide who carry out research in this area, with job opportunities also becoming available to individuals who have worked in areas of chemistry that involve significant computation – such as X-ray crystallography or computational chemistry – or in related areas such as bioinformatics or computational biology. However, there are still too few trained staff available to meet the emerging need, and this has spurred the development of university courses that can provide students with the necessary skills, at both undergraduate and postgraduate levels. What should these two types of course contain and how should the material be presented? There is, of course, no single answer to these two linked questions since they will depend on a range of local factors, such as the amount of time available in the curriculum, other modules already available in a degree programme, the particular interests and expertise of the staff and the extent of the external private-sector and public-sector involvement in the programme. As an example of possible curricula we include in the Appendix a brief summary of courses presented at the University of Sheffield; other such courses are discussed by Wiggins and Wild [6].

Material about the printed chemical literature has formed a component of undergraduate programmes in chemistry for many years, and such modules have recently been extended to cover the wide range of electronic resources that are now available, in particular the increasing amounts of structural information in resources such as SciFinder. With further enhancement such modules could provide a good introduction to chemoinformatics, and it is possible to visualise two types of module. The first would be a very simple one delivered as a core module in the first-year of a degree programme and introducing students to the basic components of chemoinformatics – chemical databases and search systems, molecular modelling and QSAR, and the IT and research environments in which systems operate. At this level, the focus should be on awareness, rather than detailed understanding, of key techniques such as structure drawing and display software, substructure searching using SciFinder Scholar or Beilstein, and chemoinformatics resources on the Internet. A second, probably final-year elective, module could then be used to provide more detailed insights into the techniques described previously and to introduce further, more specific types of system, such as flexible ligand docking and QSAR.

While undergraduate modules will enable students to become familiar with the practical aspects of chemoinformatics, they are unlikely to have gained much understanding of the

theoretical basis of the subject and may well remain unaware of many other important applications, e.g., computer-aided structure elucidation, virtual screening, chemical expert systems, 2D and 3D descriptors, pharmacophore mapping, and molecular diversity analysis, *inter alia*. Topics such as these, together with the underlying computational infrastructure and the IT skills needed to develop and maintain chemoinformatics systems, can be covered adequately only in a specialist programme. Such a multi-disciplinary programme is best offered at the masters level where sufficient time can be made available to permit the teaching of the subject in some detail to students. Such a programme will also provide sufficient space for the extended research project that differentiates an MSc from Diploma- or Certificate-level postgraduate programmes. It is our experience in Sheffield that students benefit immensely from the research project being based in industry, where they can see, and contribute to, the application of chemoinformatics methods in an operational setting. Indeed, it is our experience that all stakeholders benefit from a work placement: students gain valuable experience for their future careers; the work organisation gains an enthusiastic worker able to carry out small-scale projects that might otherwise never get done by the full-time staff; and the academic gains expert knowledge that can then be fed back into their teaching. And there is always the possibility that the project may result in, or at least contribute to, a publication in the academic or professional literature.

A dedicated MSc programme should encompass all of the material mentioned previously for undergraduate modules, albeit at a more detailed level. In addition, students need to become familiar not just with the practical aspects of the various tools that exist but also with the data structures and algorithms that underlie such systems, and to develop skills in computer programming and database design, using, e.g., C, Perl and Oracle. These core components, together with the research project, are likely to comprise more than half of a one-year MSc programme. If resources permit then one could clearly expand the amount of time available for such topics; alternatively, and in our view preferably, the remaining space in the teaching programme could be given over to modules that are already available in existing MSc programmes. Examples of chemistry or informatics modules that might prove beneficial to a chemoinformatics MSc include bioinformatics, combinatorial chemistry, information systems analysis and design, molecular modelling, and Web search engines.

At the undergraduate level, the assumption has been made that the chemoinformatics material is being delivered as part of a chemistry (or strongly chemistry-related) programme. At the MSc level, however, one may wish not only to provide chemists with informatics skills but also to consider how a course might be offered to students with computer science or other IT-related first degrees, a practice that is quite common for MSc programmes in bioinformatics. Different departments will take different views on this: thus our MSc programme in Sheffield is only open to graduates with a strong chemistry background; the programme at Indiana, conversely, is hospitable to both types of graduate.

A further factor that needs to be considered carefully, at both undergraduate and postgraduate levels, is the availability of software to support a teaching programme. At the risk of considerable over-simplification, much if not most bioinformatics software has been developed in academe and is available at zero or minimal cost. It is thus not difficult to assemble a body of data and software for teaching purposes typical of that used in industrial research and development environments. Conversely, much if not most of the software used in operational chemoinformatics systems has been developed in the private sector, and can thus be expensive to acquire for teaching purposes. This continues to be the case even with the significantly reduced charges that database hosts and software suppliers offer for educational purposes, and with the public resources (such as open-source software and RoadMap data) that are starting to become available.

In conclusion, there is a current unmet need for graduates with skills in chemoinformatics, and the shortfall is likely to grow with the continuing data explosion that is characterising

many areas of chemistry, both academic and industrial. Programmes of the sort outlined above will provide effective ways of meeting this need.

1. M. Hann and R. Green, "Chemoinformatics – a New Name for an Old Problem?", *Curr. Opin. Chem. Biol*., vol. 3, pp. 379-383 (1999).
2. W.L. Chen, "Chemoinformatics: Past, Present and Future", *J. Chem. Inf. Model*, in the press (2006).
3. J. Gasteiger, "The Central Role of Chemoinformatics", *Chemomet. Intell. Lab. System*., vol. 82, pp. 200-209 (2006).
4. A.R. Leach and V.J. Gillet, "*An Introduction to Chemoinformatics*", Kluwer (2003).
5. J. Gasteiger and T. Engel, eds., "*Chemoinformatics*", Wiley-VCH (2003).
6. D.J. Wild and G.D. Wiggins, "Challenges for Chemoinformatics Education in Drug Discovery", *Drug Discov. Today*, vol. 11, pp. 436-439 (2006).

Appendix

As an exemplar of the sorts of course that can be offered, the appendix describes two such that are presented at the University of Sheffield.

The first is a 10-credit module (which equates to one-twelfth of a year of study) that is offered to first-year students in the Department of Chemistry. The module – entitled Introduction to Chemoinformatics – introduces students to the computer handling of chemical structure information and the role of these techniques in the drug discovery process. The brief nature of the module – a one-hour lecture and a two-hour practical session for each of eleven weeks – means that the module cannot be comprehensive; instead, we focus on just a few key areas. First, the drug discovery process and how computational methods can assist in this process, and then basic techniques of structure representation and searching, with lectures by Sheffield staff on these two topics being complemented by presentations from a major pharmaceutical company and from Chemical Abstracts Service, respectively. Second, virtual screening, covering both ligand-based and structure-based methods, and with an extended practical exercise that introduces students to an industry-standard docking program (GOLD) when applied to a set of known protein-ligand complexes taken from the CCDC/Astex validation set. Other practical sessions cover structure input and display, and property prediction, using SMILES, ISISDraw and MOLSOFT software.

The second course to be described here is the one-year MSc in Chemoinformatics (see http://www.shef.ac.uk/is/prospectivepg/courses/chem/index.html) that has been run in the Department of Information Studies since 2000, at which time it was the first such programme of its type anywhere in the world. The programme is designed for students with a first degree in chemistry (or a closely related topic such as biochemistry) and seeks to provide them with in-depth knowledge both of chemoinformatics and of informatics more generally. There are thus compulsory 15-credit modules covering topics such as Web search engines, database design methodologies, and computer programming (a Java module given by the Department of Computer Science, and an introduction to Perl), as well as additional elective 15-credit modules in topics such as human-computer interaction, health informatics and electronic publishing. We are able to provide such a strong slate of informatics modules since the MSc programme draws on modules that are also open to other departmental programmes in Information Management, Information Systems and Librarianship.

The chemoinformatics components of the taught part of the programme comprise 110 of the 180 credits that comprise the MSc, and are in three parts. The first-semester Principles of Chemoinformatics covers the basic components of the subject, with lectures and practical sessions that cover: representing and searching 2D and 3D chemical structures; similarity searching; chemical patent searching; structure-activity relationships; combinatorial library

design and molecular diversity; and structure-based drug design. The lectures are complemented by an extensive set of practical sessions that provide the students with practical skills using a wide range of industry-standard software. The practicals include both those in the undergraduate module mentioned previously and also work on 2D substructure searching, 3D substructure searching and pharmacophore mapping using Merlin, Unity and GASP, respectively. Then, in the second semester, the students receive lectures in two areas: first, lectures covering algorithmic approaches that underlie chemoinformatics systems, covering topics such as graph theory, cluster analysis and machine learning; second, lectures from visiting speakers from the pharmaceutical, agrochemical, chemical database and chemical software industries, illustrating the application of the basic techniques that had been introduced in the first semester. The lectures are complemented by a small-scale software project that draws on the first-semester Java programming course and that might, e.g., involve a comparison of diversity selection algorithms or creating a web-based applet in which a user can draw a compound . Also during the second semester, students carry out the literature review and other preliminary work on the dissertation project that runs from the end of semester-2 (mid-June) through to the end of the programme at the start of September. The project is carried out in collaboration with, and at the premises of, one of the companies that support the programme. In its six years, a total of 16 companies have provided projects, initially just in the UK but subsequently more generally in Europe. We believe that this is the most important part of the MSc since it enables students to practice and develop their skills in an operational environment that makes very clear the application of the theoretical material that they had met during the taught part of the programme.