ELSEVIER

Challenges for chemoinformatics education in drug discovery

David J. Wild and Gary D. Wiggins

School of Informatics, Indiana University, 901 E 10th Street, Bloomington, IN 47408, USA

Chemoinformatics is rapidly becoming a core part of drug design informatics, yet the educational opportunities in the field are currently limited. This article reviews the academic and commercial educational programs that are available in chemoinformatics, considers the current challenges and takes a look at emerging trends, such as distance education and intensive short courses.

Like many of today's emerging life science fields, chemoinformatics has become a 'hot topic' while it is still in the process of finding its identity. Indeed it is not yet clear how to spell the name of the field: some prefer cheminformatics – no 'o' – and others, including ourselves, use entirely different terms, such as chemical informatics. What is clear is that the techniques that this field concerns itself with – the processing of chemical and related information on computers – are becoming central to the processes of modern drug discovery [1].

The use of computers for the processing of chemical information had its origins in various scientific and computer-related disciplines, as well as some pioneering academic work, such as that at the University of Sheffield, UK, in the late 1960s on the representation and use of chemical structure information. For many years, the field of chemical-information handling had a well-defined and important part to play in library and information systems, where the techniques were used for the storage and searching of chemical and patent databases. In addition, computational chemistry methods were applied to aid understanding of theoretical chemistry and pharmaceutical drug discovery where a variety of techniques, such as similarity searching and docking, helped scientists find potentially active molecules and model how those molecules bind to protein targets in the body. This latter application, known as computer-aided drug design or molecular modeling, rapidly became an essential part of modern drug discovery, and thus the pharmaceutical industry became a strong supporter of the field. In particular, the race to develop antiretroviral drugs for AIDS resulted in a strong push to use computational techniques to speed up the discovery of new drugs, and the existence of several marketed drugs, such as Viracept® (Pfizer), can be traced directly to the application of computer-aided drug design [2]. It is important to recognize that despite its strong support from the pharmaceutical industry, the field, unlike other life-science-related computer disciplines, also has strong support and application outside the life sciences arena.

During the mid to late 1990s, the technological innovations that began to impact the processes of early stage drug discovery resulted in a vast increase in the amount of chemical information available. HTS enabled hundreds of thousands of compounds to be screened for biological activity in a short period of time, and combinatorial chemistry enabled the automatic synthesis of thousands of new structural analogs at once. Likewise, improvements in X-ray crystallography and NMR spectroscopy meant that the number of proteins with resolved 3D structures increased dramatically, and new experimental pharmacology and toxicity methods led to a dramatic increase in useful information about the effects of drugs in the body. More recently, the advent of genomics and particularly microarray assays has meant that cellular samples (including those treated with chemical compounds) can be tested for up- or down-regulation of thousands of genes at once. Therefore, Modern-day early-stage drug discovery involves the generation of vast amounts of pertinent data from a diverse set of sources. However, effectively managing, organizing and analyzing that data can be extremely difficult, particularly because the volumes are much more than can be analyzed manually. There is therefore much demand for techniques that allow the data to be turned into useful information and knowledge [3].

Demand is also increasing for people who have training and experience in these techniques. However, there is still only a very small number of academic institutions that offer teaching and

Corresponding author: Wild, D.J. (djwild@indiana.edu)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.03.010

qualifications in chemoinformatics or related disciplines, and these institutions generally only offer substantial graduate courses that require relocation to the site of the institution and a time commitment that is incompatible with maintaining parallel full-time employment.

Academic programs in chemoinformatics

A small number of universities have established chemoinformatics programs [4–6]. The most widely recognized and well-established research and teaching base in the field is the Department of Information Studies at the University of Sheffield, which offers Master of science (MSc, or MS) degree and PhD qualifications in chemoinformatics. Subsequent programs have been developed at the University of Manchester Institute of Science and Technology (UMIST), now merged with the University of Manchester, UK, and the School of Informatics at Indiana University (IU), IN, USA. The programs offered by the institutions, and their URLs, are summarized in Table 1.

University of Sheffield

At Sheffield, the Master of Science (MSc, or MS) program includes courses in database design, information searching and retrieval, principles of chemoinformatics, foundations of object-oriented programming, research methods and dissertation preparation, and a dissertation in chemoinformatics. Sheffield also offers periodic intensive short courses in chemoinformatics, which emphasize practical chemoinformatics skills including 2D databases and database searching, diversity and compound selection, QSAR, computational methods for 3D, 3D databases, combinatorial libraries, and analysis of HTS data.

University of Manchester

At the University of Manchester, the MSc course has been developed (originally created at UMIST before their merger) in the School of Chemistry, in conjunction with sponsors from the chemical informatics community. Core modules include research methodology and feasibility study, chemical information sources, molecular simulation and design, spectroscopy and crystallography in cheminformatics, database design and programming, cheminformatics applications, and fundamentals of bioinformatics. There is also a Master of Enterprise (MEnt) degree at the Manchester Science Enterprise Centre, which offers the opportunity to combine chemoinformatics and business courses.

Indiana University

The Indiana University School of Informatics has programs in many areas of science informatics, including bioinformatics and chemical informatics, at the Bloomington (IUB) and Indianapolis (IUPUI) campuses. In addition, IUPUI has MS graduate programs in laboratory informatics and health informatics and a developing medical informatics program, plus a Bachelor of Science (BS) undergraduate program in health information administration. The School has a variety of opportunities for training in chemical informatics for undergraduate and graduate students, including a BS in informatics with a cognate in chemistry, an MS in chemical informatics, and a PhD in informatics with a track in chemical informatics. Core chemoinformatics courses for these programs include: chemical information technology, computational chemistry and molecular modeling, programming for science informatics and an independent study module, with other required courses in introductory informatics and information management. MS students must complete a substantial capstone project. Since the first year they were taught, in 2001, the core chemical-informatics courses have used remote conferencing technology to allow participation from several sites, and they are now offered to any US-resident graduate as part of a distance-education chemical informatics certificate program. The classes are carried out live on the internet, currently using the Raindance meeting edition web conferencing system (www.raindance.com).

Other institutions

There is evidence of increasing interest in the creation of chemoinformatics programs at other institutions. The Beilstein Institute has funded an endowed chair for chemoinformatics at Johann Wolfgang Goethe University of Frankfurt, Germany (http://gec-co.org.chemie.uni-frankfurt.de), and Unilever has funded the Centre for Molecular Informatics at the University of Cambridge, UK (http://www-ucc.ch.cam.ac.uk). The Department of Computer Science at the University of Massachusetts, Lowell, USA, has created both an undergraduate and MS program in bio- and chem-informatics (http://cs.uml.edu).

There is a one-year postgraduate distance-education diploma program at the Institute of Cheminformatics Studies in Noida, India (www.cheminformaticscentre.org/files/index.asp). A pharmaco-chemistry MS degree can be obtained from the University of Strasbourg, France (http://bioinfo-pharma.u-strasbg.fr/education.php), and at the Friedrich Alexander University of Erlangen–Nuremberg, Germany, chemoinformatics is part of the course on molecular science (www2.chemie.uni-erlangen.de/education/cheminf).

Michigan Technological University, USA now offers a BS in Cheminformatics (www.chemistry.mtu.edu/pages/undergrad/index.php). Several undergraduate institutions list chemoinformatics as possible areas of specialization, for example, Rensselaer Polytechnic Institute, NY, USA, has a science IT concentration

TABLE 1

Institutions offering academic programs in chemoinformatics[a]

Institution and web address	Masters degree	PhD	Certificate	Short courses	Distance learning
University of Sheffield Department of Information Studies (www.shef.ac.uk/is)	✓	✓		✓	
Indiana University School of Informatics (www.informatics.indiana.edu)	✓	✓	✓		✓
University of Manchester School of Chemistry (www.chemistry.manchester.ac.uk)	✓			✓	✓

[a] Institutions offering only PhD programs are excluded.

option in cheminformatics (http://admissions.rpi.edu/update.-do?artcenterkey=19andsetappvar=page(1)), and at the University of Pittsburgh, Bradford, PA, USA, a molecular informatics BS was proposed in the 2004–2005 academic year. A BS in molecular informatics, which is essentially the art of organizing and accessing vast amounts of molecular data, would provide graduates with the opportunity to pursue advanced study in such areas as proteomics, bioinformatics, chemoinformatics, drug design, computational chemistry and molecular biology. Graduates would also be prepared for technical positions in the pharmaceutical industry.

Other institutions that do not have formal programs still contributed to learning in the field of chemoinformatics. For example, although it was never continued beyond pilot stage, an interesting program was developed at the University of Nottingham, UK, in conjunction with Pfizer called the Virtual School of Molecular Sciences, which established an internet-based course on structure-based drug design.

The academic activities in this field have also spawned a small number of textbooks. Of particular note are Leach and Gillet's very accessible and readable *An Introduction to Chemoinformatics* [7], Gasteigher and Engel's comprehensive *Chemoinformatics: A Textbook* [8] and the multi-volume *Handbook of Chemoinformatics* [9].

Commercial offerings

Chemoinformatics education has not seen much uptake in the commercial sector, although there are a few offerings which stand out. Molecular Conceptor (www.molecular-conceptor.com) have a computer-based course for teaching the fundamentals of medicinal chemistry, drug design, molecular modeling and chemoinformatics for universities and Industry. According to their website, the course is currently used by more than 200 academic institutions worldwide. Included in the course are modules on molecular modeling, protein structure and modeling, rational drug design, structure- and pharmacophore-based drug design, QSAR, synthesis and library design, peptidomimetics, ADME properties and predictions, and database searching. Mesa Analytics and Computing (www.mesaac.com) recently won a Phase II small-business innovation research (SBIR) grant to develop web-based teaching tools for chemoinformatics, and it is likely that these tools will be used to form a 'virtual classroom' for teaching chemoinformatics. Network Science Corporation offer a brief, but free, set of introductory course materials (www.netsci.org/courseware) focusing on drug design.

Historically, some of the chemoinformatics conferences, particularly the more intimate corporate ones like the Daylight MUG (www.daylight.com) and OpenEye CUP (www.eyesopen.com), have also served as informal places of learning about chemoinformatics in general. Several companies arrange more formal short courses and conferences that focus on particular areas of life sciences, including chemoinformatics, and which serve as intensive learning opportunities in the field. Most notable have been Cambridge Healthtech (www.healthtech.com) and more recently eChemInfo (http://echeminfo.colayer.net).

A look to the future

The future of the field of chemoinformatics is still in flux and is, of course, dependent on the directions that life science research takes. The pharmaceutical industry is still a major driver in the field, but the growing open-source movement [10,11] and government sponsorship of chemoinformatics programs, such as PubChem [12], is helping to broaden the accessibility and use of chemoinformatics tools and techniques. The increase in external licensing deals for drugs by pharmaceutical companies (as opposed to research being primarily internal to the company) is likely to result in more support for smaller, corporate or university drug-development research units, and thus a demand for affordable chemoinformatics tools for these groups, and a corresponding need for education in chemoinformatics techniques. Moreover, the vibrant young pharmaceutical industries developing in India and China are clearly going to be technology-focused, with computational specialties like chemoinformatics and bioinformatics at the center of the drug development programs. Research on the interfaces of drug design chemistry and emerging fields like proteomics, genomics and metabolomics will likely spur further application of chemoinformatics techniques and some degree of blurring of the previous distinctions between computational domain specialties.

According to John Reynders, Informatics Officer, Discovery and Development Informatics at Eli Lilly and Co. in Indianapolis, IN, USA, the most desirable people are now those with a wide range of skill in informatics, what he calls a 'technology stack' of software development skills plus 'meta-skills', including the demonstrated ability to work well in teams, to rapidly innovate in new areas, to easily cross traditional boundaries, to learn domain specifics quickly, and to apply techniques from one domain to another. Reynders observes that in the rapidly-changing world of modern life science, these abilities are often more important than extensive training in one highly specialized area.

This implies that as well as training people as specialists in chemoinformatics, we should also be training them as agile 'informaticians', with the ability to cross over traditional boundary lines when necessary. Encouragingly, Reynders says that diversity in backgrounds and training is one of the great strengths of the informatics staff at Lilly: 'we have biologists cross-trained in software development, software engineers with chemoinformatics skills, formally trained bioinformaticians – this broad and eclectic mix gives us great innovativeness and flexibility in what we do'.

This attitude has also been reflected elsewhere. In 2005, the IU School of Informatics had the benefit of sound advice on educational programs from two outstanding groups of industry experts: the members of our Science Informatics Advisory Board and the external advisors for a recently received US National Institutes of Health grant project for an exploratory center for cheminformatics research. Again, it was highlighted that prospective industry employees in scientific computing needed to have skills that go beyond the traditionally instilled academic ones, including change management, negotiation skills, modeling and statistics, project management and generalized data-mining.

There seemed to be varying interest from industry in external chemoinformatics education programs. Some reported little interest in prepackaged educational modules, but greater interest in short, intensive courses. We were advised to develop an industry-exchange program to keep a finger on the pulse of industry. One suggestion was to have IU staff work in industry for a couple of weeks (or even have faculty spend a sabbatical there); alternatively, we could have private sector workers come to IU to present

problems facing the industry. Of course, any program aimed at industry professionals must take into account the competing time-demands on students, and it is still unclear how this works best in practice (e.g. should students be required to commit time to the course during normal working hours?).

We consider distance education to be an essential part of any successful chemoinformatics program in the coming years. The great advantage of using distance education is that students do not have to relocate to take a course, and thus the potential exists for a small number of institutions to resource chemoinformatics learning around the world. Experiments in distance education at Indiana University are, at the time of writing, looking very promising. Online and telephone conferencing technologies now permit easy participation in live lectures by students regardless of location, and even allow more flexibility than traditional classroom courses (enabling guests to lecture without traveling, teleconferenced panel discussions and so on). Asynchronous technologies such as the Sakai-based university course repositories (www.sakaiproject.org) permit the sharing of documents and interaction between students and lecturers outside of live classes.

Challenges

Thus the biggest current challenges for those of us involved in chemoinformatics education are to appropriately define the scope of chemoinformatics (particularly with relation to emerging disciplines such as chemogenomics, systems biology and proteomics), to contextualize it in wider training in transferable informatics skills, and to steer the development of courses and programs now so that they will meet the educational needs in the next 2–5 years. Currently, demand for graduate chemoinformatics education at the MS level is fairly small, with the programs mentioned in this article typically attracting only a handful of students each year. However, we do see a much greater interest in short courses and certificate programs that can enable industry professionals to get 'up to speed' with the state of the art, and we expect demand for these kinds of programs to grow in the coming years. Whether increasing visibility of chemoinformatics will encourage more students to choose a more involved Masters or PhD program remains to be seen. What is clear, is that chemoinformatics is now established as an essential ingredient of modern drug discovery and is likely to remain so for the foreseeable future.

References

1 Russo, E. (2002) Chemistry Plans a Structural Overhaul. *Nature* 419, 4–9
2 Henry, C.M. (2001) Structure-Based Drug Design. *Chem. Eng. News* 79, 69–74
3 Mullin, R. (2004) Dealing with Data Overload. *Chem. Eng. News* 82, 19–24
4 Borkent, H. (2004) Cheminformatics and Chemistry Teaching. In *Cheminformatics Developments: history, reviews and current research* (Noordik, J.H., ed.), pp. 203–217, IOS Press
5 Cooke, H. and Willett, P. (2004) Masters level training in chem(o)informatics in the U.K.. *228th National Meeting of the American Chemical Society*, Pittsburgh, PA, USA. CAS AN 2004:656494
6 Schofield, H. *et al.* (2001) Recent developments in Chemoinformatics Education. *Drug Discov. Today* 6, 931–934
7 Leach, A.R. and Gillet, V.J. (2003) *An Introduction to Chemoinformatics*. Kluwer Academic Publishers
8 Gasteiger, J. and Engel, T. (2003) *Chemoinformatics: A Textbook*. Wiley-VCH Verlag
9 Gasteiger, J. (2003) *Handbook of Chemoinformatics*. Wiley-VCH Verlag
10 DeLano, W. (2005) The case for open-source software in drug discovery. *Drug Discov. Today* 10, 213–217
11 Stahl, M.T. (2005) Open source software: not quite endsville. *Drug Discov. Today* 10, 219–222
12 Austin, C.P. (2004) NIH Molecular Libraries Initiative. *Science* 306, 1138–1139