

Masters Level Training in Chem(o)informatics in the UK

Helen Cooke (formerly UMIST)
and
Peter Willett (University of Sheffield)

Presented on 23rd August 2004 at the 228th
ACS National Meeting, Philadelphia, PA,

Outline

- Funding support from EPSRC
- MSc programmes at the University of Sheffield and University of Manchester Institute of Science and Technology (UMIST)
- Structure and development of the Sheffield programme
- Lessons learned from the UMIST programme

Funding Support for Chemoinformatics: I

- The Engineering and Physical Sciences Research Council (EPSRC) is one of the main sources of research funding in the UK
- EPSRC's principal type of student support is funding for PhD research but also supports MSc programmes if, and only if, can demonstrate significant industrial needs
- Approach to head of the Chemistry section of EPSRC by Willett outlining the world-wide shortage of people with skills in chemoinformatics, followed by briefing paper detailing the industrial context (Q3/4 1998)

Funding Support for Chemoinformatics: II

- EPSRC discussions with chemical and then pharmaceutical companies (Q1/2 1999)
- Acceptance of industrial need following round-table meeting at Association of the British Pharmaceutical Industry in June 1999
- EPSRC Call for Proposals for new Masters Training Packages (MTPs) in November 1999 included Chemoinformatics as one of about a dozen priority areas

Funding Support for Chemoinformatics: III

- Sheffield proposal developed in collaboration with a range of pharmaceutical, agrochemical and software companies
- Notification of funding received in mid-July 2000 for start in Sept 2000. Funding over 5 years (ca. £500K) to cover
 - Tuition fees (10 students per year)
 - Substantial part of student maintenance costs, with rest paid for by collaborating companies
 - Small contribution to teaching costs
 - Publicity, equipment and odds-and-ends
- UMIST proposal funded for start in Sept. 2001

Principal Characteristics of the Sheffield MSc

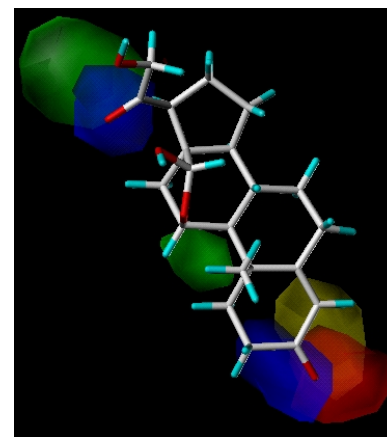
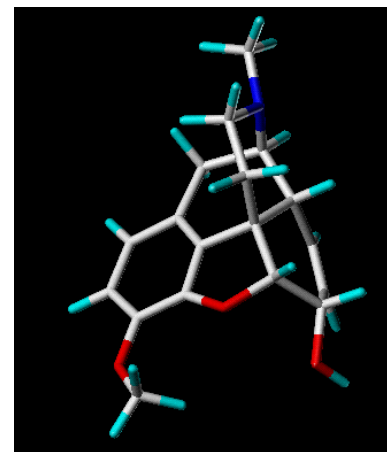
- Informatics focus
 - Aims to teach chemoinformatics and information management skills to chemists (cf many bioinformatics programmes)
 - Teaching input principally from Information Studies, using both existing and new modules
- Collaborative nature
 - Modules from Chemistry and Computer Science, and lectures also from Automatic Control and Systems Engineering and from Molecular Biology and Biotechnology
 - Substantial input from industry: lectures, software and student placements
 - AstraZeneca, Aventis, BCI, CAS/STN, CCDC, ChemWeb, Daylight, Eli Lilly, Evotec OAI, GlaxoSmithKline, Johnson & Johnson, Merck, MIMAS, Novartis, Organon, Pfizer, Roche, Syngenta, Tripos

Structure of the Sheffield MSc

| Semester 1 | Semester 2 | Semester 3 |
|--|--|--|
| Chemoinformatics-I (principles plus introduction to bioinformatics) | Chemoinformatics-II (applications plus data mining methods) | 3 month industry-based dissertation project (UK, mainland Europe) |
| Information Systems Modeling | Database Design | |
| Information Retrieval | Dissertation Preparation | |
| Introduction to Programming | Choice of two from Molecular Modeling Healthcare Information E-publishing Advanced Information Retrieval | |

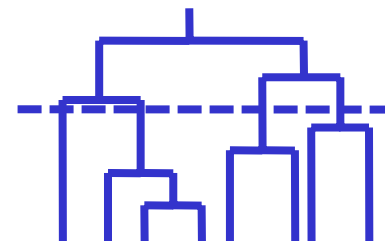
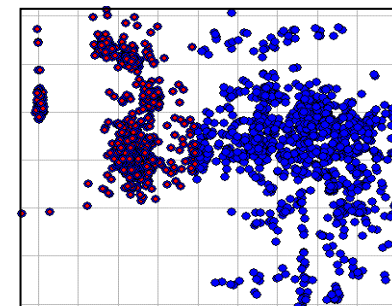
Chemoinformatics I

- Representation and searching of chemical structures
 - Structure, substructure & similarity searching in 2D & 3D
 - Markush and reaction searching
- Techniques used to design bioactive compounds
 - Drug discovery process; QSAR; Combi-chem; SBDD
- Representation and searching of biological databases
- Introduction to variety of commercially available chemoinformatics software



Chemoinformatics II

- Practical implementation of techniques introduced in Chemoinformatics I
 - Richard Lewis (QSAR); John Delaney (Compound Selection); Andrew Leach (SBDD); Frank Allen (CCDC); Simon Cross (Software Industry); Andrew Poirrette (AstraZeneca); David Clark (Argenta)
- Data analysis techniques
 - Clustering; Evolutionary algorithms; Graph theory; Neural networks;
- Small scale chemoinformatics programming project



Dissertation Project: I

- 3 month placement at a collaborating company
 - Pre-placement visit by both student and Sheffield supervisor and one visit by the Sheffield supervisor during the placement
- Preparation begins in Sheffield in previous semester
 - Allocation of projects, literature review, fleshing out project proposal
- Original bid, and first three years of the programme, involved only UK placements
 - The success of these and our industrial links have meant that four of this year's dissertations are being carried out overseas

Dissertation Project: II

- Three main types of project
 - Development and/or testing of an existing or novel piece of software for some specific application
 - Comparison of different programs for some specific application, e.g., docking, pharmacophore identification etc
 - Analysis of chemical and/or biological dataset(s) using, e.g., different descriptors for property prediction or different techniques for analysing HTS data etc

Example – Comparison of Pharmacophore Mapping Methods: I

- Catalyst/Hiphop, GASP, DISCO were used to analyse protein-ligand complexes from the PDB
 - Five proteins: thrombin, CDK2, DHFR, HIV-RT, thermolysin
- Approach
 - Pharmacophore determined manually, based on bound conformations and common features
 - The various ligands were minimised and the programs used to identify the pharmacophore automatically, with this then being compared to the “correct” result
- Collaborators
 - Sheffield (Yogi Patel, Val Gillet)
 - GlaxoSmithKline (Andrew Leach, Giampa Bravi)

Example – Comparison of Pharmacophore Mapping Methods: II

- Comparisons are subjective to some degree due to nature of problem and variation in program features
- Variations also in ease-of-use, software limitations, e.g., number of molecules that can be considered, pharmacophore feature definitions
- GASP and Catalyst generally out-performed (the original version of) DISCO
- Patel *et al.* *Journal of Computer-Aided Molecular Design* 2002, **16**:693-681.

Employment Destinations

| | |
|------------------|---|
| 2000-2001 (9) | 6 Chemoinformatics (3 Pharmaceutical; 1 Software; 2 PhDs) 1 General IT 1 Health Service 1 Unknown |
| 2001-2002 (9) | 8 Chemoinformatics (or related, all PhDs) 1 Unknown |
| 2002-2003 (9) | 5 Chemoinformatics (4 Pharmaceutical; 1 PhD) 2 General IT 2 Unknown |

Chemoinformatics Short Course: I

- Original MTP call specified that should be a substantial distance-learning component
- The preparation of such material is very time-consuming and discussions with pharmaceutical companies revealed little interest in same
- Alternative approach to CPD by an intensive short course
 - “A Practical Introduction to Chemoinformatics” sponsored by MGMS and CSA Trust
 - Run annually over 4 days since 2001

Chemoinformatics Short Course: II

- Delivered to ~16 delegates from UK, mainland Europe and US
 - Mostly from industry; 2 student bursaries a year
- Restricted numbers due to intense nature and practical component
 - Over-subscribed every year it has been run
- Tutors
 - Academic: Val Gillet, John Holliday, Nick Rhodes and Peter Willett
 - Industrial: Andrew Leach (GSK), Richard Lewis (Eli Lilly) and Robin Taylor (CCDC);

Chemoinformatics Short Course: III

- Content
 - Mixture of lectures and hands-on sessions discussing applications to modern drug discovery
 - Group problem-solving exercise
- Topics
 - 2D and 3D database searching
 - Diversity analysis and compound selection
 - QSAR
 - Experimental 3D data sources
 - Pharmacophore generation and protein-ligand docking
 - Analysis of high-throughput screening data
 - Combinatorial library design

The UMIST MSc

- UMIST course funded by EPSRC via same mechanism as the Sheffield course
- Assessment by examination, coursework and dissertation
- Course run by the Chemistry Department: content reflects the research interests of the Department
- Somewhat different “flavour” of cheminformatics

UMIST Applications

| | 2001/2 | 2002/3 | 2003/4 | 2004/5 up to 15 April |
|---------------------------------------|--------|--------|--------|-----------------------|
| Enquiries | 34 | 61 | 62 | 41 |
| Applications | 16 | 30 | 35 | 14 |
| Offers | ? | 25 | 31 | 10 |
| Acceptances (UK – EPSRC funded) | 9 | 7 | 10 | 5 |
| Acceptances (European Union) | 2 | 1 | - | - |
| Acceptances (overseas or self-funded) | 1 | 4.5 | 10 | 2 |

Students recruited from the UK, China, France, Greece, India, Italy, Jamaica, Pakistan

Core Modules

- Chemical information sources
- Cheminformatics applications
- Database design and programming
- Spectroscopy and crystallography in cheminformatics
- Molecular simulation and design (2 modules)
- Fundamentals of bioinformatics (double module)

Elective Modules

- Algorithm design
- Combinatorial chemistry
- Management of intellectual property (2001 and 2002); replaced by Technology enterprise (2003)

Workshops

- Hands-on practice of cheminformatics and chemical information systems
- Visits from outside lecturers, e.g. suppliers of information systems and software
 - STN International
 - Chemical Abstracts Service
 - MDL Information Systems
 - Daylight Chemical Information Systems

Dissertations

- Industrial / external placements
 - Pfizer Global Research
 - GlaxoSmithKline
 - AstraZeneca
 - De Novo Pharmaceuticals
 - PSigate (www.psigate.ac.uk)
 - Royal Society of Chemistry
- Internal
 - Department of Chemistry
 - Department of Biomolecular Sciences

Sample Dissertation Topics

- Automatic extraction of genotoxic properties from the biomedical literature
- Molecular modelling of the structure and properties of amino acid strands
- Access to chemical publications and services
- Compound knowledge capture
- Chemistry gateways and portals
- Use of parser programs to analyse the effectiveness of training in chemical database usage

Distance Learning

- Original intention: to convert whole course to distance learning
- Units developed on Chemical Information Sources and Molecular Modelling
- Units to be tested on 2004/5 full-time students
- May be sold as self-paced learning modules

Sample page from “Chemical Information Sources” unit



MYWEBCT | RESUME COURSE | COURSE MAP | HELP

Chemical Information Sources

Home · Course Content · Searching for ... · The Merck Index · Editorial Proc...

6. The Chemical Abstracts Registry System

6.2 Editorial Procedure and Assignment of Registry Numbers

When the editors (who are all trained chemists or other experts in the subject matter covered by *Chemical Abstracts*) receive a new paper for inclusion in *Chemical Abstracts*, they have to **write the abstract** of the paper, **assign keywords** from *Chemical Abstracts*' controlled vocabulary, and **compare compounds** of significance against the compound already present in the database. If the compound is not on the database, then it is deemed to be new and a **Registry Number** is assigned.

Registry number examples

Below are some examples of Registry Numbers:

Formaldehyde 50-00-0

Quinine 56-54-2, 130-95-0, 572-59-8, 572-60-1, 42151-59-7, 47342-58-5, 72402-50-7, 72402-51-8, 72402-52-9, 72402-53-0, 101143-86-6, 101143-87-7, 101143-88-8, 146925-10-2

Water 7732-18-5

Quinine has a number of Registry Numbers, primarily because of the possibility of a **number of stereoisomers**, see the results of a search of *Chemical Abstracts* through *SciFinder* below.

The screenshot shows the SciFinder Scholar interface with three search results for Quinine stereoisomers. Each result displays a chemical structure and associated information:

- 42151-59-7**: -3 References REGISTRY
- 572-60-1**: Regulated Chemicals Listing
- 572-59-8**: Commercial Sources

The chemical structures are stereoisomers of quinine, featuring a quinoline ring system, a quinuclidine bicyclic core, and a vinyl group. The stereochemistry at the chiral centers is indicated with wedges and dashes.

Employment Record (2002)

- 4: PhD research in UMIST's Chemistry Department
- 4: employed as cheminformaticians or chemical information scientists in the pharmaceutical or drug discovery industry
- 1: employed as an analytical chemist
- 1: Post-graduate Certificate of Education course
- 2: unknown

Strengths

- Department's previous experience of running Masters courses
- Potential supply of students from local undergraduate chemistry courses
- Availability of studentships makes course attractive
- Popularity with overseas students
- Modular nature allows for part-time attendance
- Ideal group size for teaching

Weaknesses

- Students with best undergraduate degrees often opt for PhD research
- Not all overseas students who accept places materialise

Opportunities

- Increasing number of applications from overseas
- Merger of UMIST with the University of Manchester
 - Potential for better range of dissertation projects
 - Potential for new course modules
- Increasing awareness of course in industry may enhance industrial support

Threats

- EPSRC funding ceases after 5 years
 - Course will only survive if all students are self-funded, unless more industrial support is received
- Merger of UMIST with the University of Manchester
 - Some staff may change departments or leave

Conclusions: Sheffield and UMIST

- Student supply and demand have balanced well
- Differences in courses have helped ensure their popularity
- Students have been successful in finding jobs
- Recognition of courses by industry is increasing
- Long-term future of courses uncertain unless additional sponsorship by industry and/or UK Government is forthcoming