Recent Developments in Chemoinformatics Education

Val Gillet
University of Sheffield

Chemoinformatics as a Discipline

- Chemical Information Systems and Services have been established for many years
  - Chemical Abstracts started in 1907
  - First computerised systems established in 1960s
- Recent emergence of Chemoinformatics as a discipline
  - Chemoinformatics; Cheminformatics; Chemi-informatics; Chemical Informatics; Molecular Informatics
  - "Chemoinformatics - a new name for an old problem?" Hann M., Green R. Current Opinion in Chemical Biology, 3, 379-383, 1999.

Chemoinformatics: Definitions

- "mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization"

  Brown, F. K. Annual Reports in Medicinal Chemistry, 33, 375-384, 1998

- "Chem(o)informatics is a generic term that encompasses the design, creation, organization, storage, management, retrieval, analysis, dissemination, visualization and use of chemical information"

  Greg Paris (August 1999 ACS Meeting) quoted by Wendy Warr at www.warr.com

Chemoinformatics: The Rationale

- The development of high-throughput screening and combinatorial chemistry techniques has led to a huge increase in the volumes of data about structures and their bioactivities

- The explosion of data has increased the need for integration of chemical information (archival functions) with molecular modelling techniques

- Now an increasingly wide range of "informatics"
  - Bioinformatics, Medical Informatics, Health Informatics, Educational informatics etc

Skills Required (www.warr.com)

- Tasks involved
  - database design, SAR, programming, multivariate analysis, pattern recognition, library design, calculation of properties, data mining, chemical registration, and library enumeration.
- Skills required
  - understanding of (medicinal) chemistry; ISIS, Web, Visual Basic and ORACLE experience; ability to analyse and correlate data from massive data banks; programming (UNIX scripting C++); interpersonal skills; enthusiasm; ability to look at the bigger picture; and project management skills

Supply vs Demand

- World-wide industry shortage
  - Chemoinformatics experts: supply and demand (www.warr.com August 1999 ACS Meeting)
  - Ideal CV
    - First degree in Chemistry (or related);
    - PhD eg Cambridge, Erlangen, Leeds, Oxford, Sheffield, York, UCSF, Texas
    - *2 years Post-doc experience*
  - supply ~12 per year
  - demand ~40 per year

- Head hunters
  - 1 approach every 2/3 weeks

Meeting The Demand

- UK
  - December 1999: EPSRC Call for proposals
    - Masters Level Training Packages (MTPs)
    - Chemoinformatics: priority area
  - Sheffield MSc Chemoinformatics - October 2000
  - UMIST MSc Cheminformatics - October 2001
- US
  - Indiana University MSc Chemical Informatics - Sept 2001

Chemoinformatics At Sheffield: I

- Recruitment of Michael Lynch from Chemical Abstracts Service in 1965, shortly after the creation of the Department, to teach "computing in libraries"
- Initiated research in database processing, both textual (which is common in an LIS environment) and chemical (which certainly is not common)
- Three of the four current members of the Chemoinformatics Research Group (Gillet, Holliday and Willett) worked for Lynch as PhD students and post-docs

Chemoinformatics At Sheffield: II

- In addition to its research activities, the Department has long provided a one-term course in chemoinformatics to students on (what is now) the MSc Information Management taught programme
- Proved the starting point for many people who later did doctoral studies, but the numbers of scientists of all sorts (not just chemists) applying to the MSc programmes have been reducing steadily over the years….
- …while the need for specialist training has increased

Sheffield's MSc Chemoinformatics

- Developed in collaboration with a range of pharmaceutical, agrochemical and software companies
  - AstraZeneca, CCDC, ChemWeb,Eli Lilly, GlaxoSmithKline, Merck, Novartis, Oxford Molecular, Pfizer, Roche, Syngenta,Tripos….
  - Support via industrial placements, lectures, software

- EPSRC funding includes
  - 10 studentships over 5 years
  - tuition and 2/3 maintenance (1/3 coming from industry)

- 2000-01: 10 graduates in chemistry; biological sciences; food science
- 2001-02: 10 graduates (1 part-time): chemistry; biological sciences

Teaching Staff

- Peter Willett, Val Gillet (Programme Coordinator), John Holliday, Nick Rhodes
- Other staff in Information Systems (Beaulieu, Nunes, Eaglestone, Sanderson)
- Contributions from departments of
  - Automatic Control and Systems Engineering; Chemistry; Computer Science; Molecular Biology and Biotechnology
- External lecturers

Aims

- Develop an awareness of IM and IT techniques used in the design and implementation of chemoinformatics systems
- Enable students to demonstrate skills learned by carrying out a small-scale industrially relevant chemoinformatics research project
- Basic structure
  - Two semesters of taught modules (like other masters programmes in the Department)
  - One semester dissertation working at the site of one of the companies supporting the programme

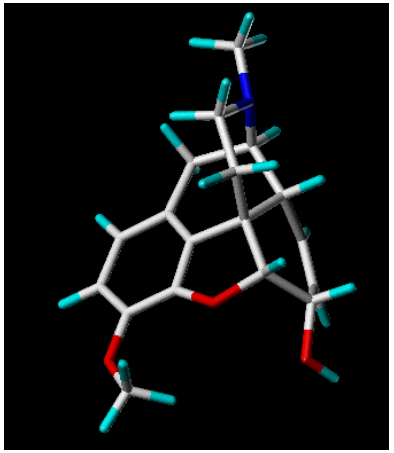
Taught Modules

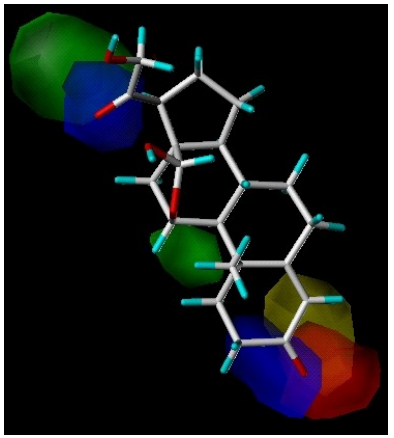
- Core modules in chemoinformatics, numeric and textual information systems and computer programming
  - Chemoinformatics I & Chemoinformatics II ;
  - Information Systems Modelling; Database Design
  - Information Storage and Retrieval;
  - Foundations of Object-Oriented Programming (CS)
  - Practical Computing
  - Research Methods and Dissertation Preparation

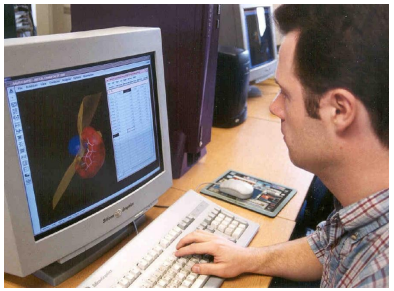
- Two from:
  - Molecular Modelling (Chemistry); Information Storage and Retrieval Research; Healthcare Information; E-Business & E-commerce

Chemoinformatics I

- Introduction to computational techniques
  - Representation and searching of chemical structures
    - 2D & 3D; exact; substructure; similarity searching
  - Techniques used to design bioactive compounds
    - Drug discovery process; QSAR; Combi-chem; SBDD
  - Representation and searching of biological databases
- Introduction to variety of chemoinformatics software
- Essay on state-of-the-art of an aspect of Chemoinformatics

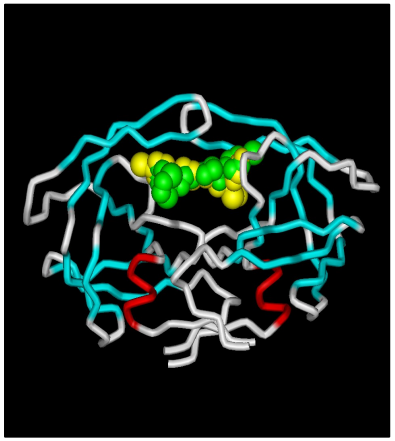


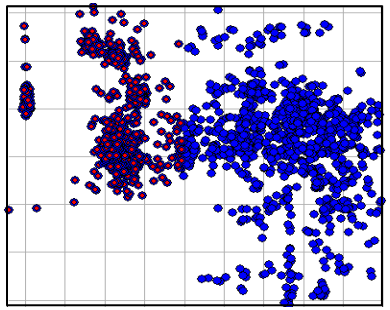


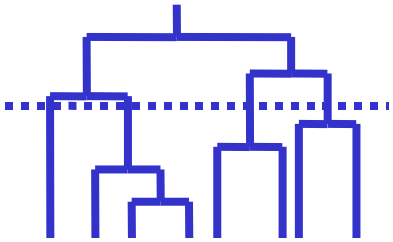


Chemoinformatics II

- Practical implementation of techniques introduced in Chemoinformatics I
  - Richard Lewis (QSAR); John Delaney (Compound Selection); Andrew Leach (SBDD); Frank Allen (CCDC); Bill Town (ChemWeb); Simon Cross (Software Industry)
- Data analysis techniques
  - Clustering; EAs; NNs; Graph Theory
- Programming exercise (see later)







Numeric Data

- Information Systems Modelling
  - Systems analysis and design; data flow modelling; entity modelling; prototyping; introduction to object-oriented methods
- Database Design
  - Relational DB; conceptual DB design; logical DB design; SQL; Distributed DBs; OO and O-relational DBs; Web DBs
  - Practical experience using MS ACCESS/Oracle
- Case-study analysis in groups

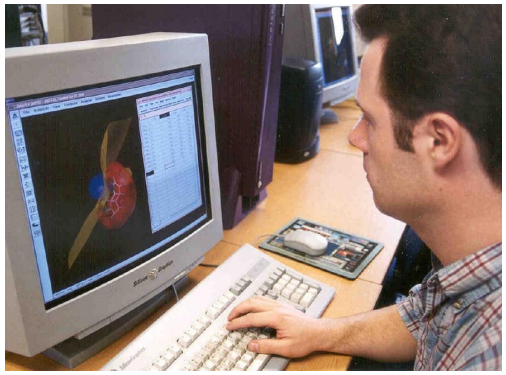
Textual Information

- Information Storage and Retrieval
  - Classical keyword searching; web searching; subject analysis; record description; user interface issues
- Information Storage & Retrieval Research
  - One of the Electives that provides a development of the basics in the first-semester module
  - Current research in IR; statistical approaches; cognitive and behavioural approaches; evaluating IR systems; AI and NLP; speech and image retrieval
- Essay or literature review

Computer Programming

- Foundations of Object Oriented Programming
  - Introduction to JAVA
  - Series of programming tasks and exam

- Chemoinformatics II
  - Students choose one from six projects, each illustrating some aspect of the course, e.g., a chemical sketcher, selecting a structurally diverse combinatorial library, selecting a set of 2D fragments for chemical substructure searching
  - Use of both industry-standard software (SYBYL, UNITY, BCI, CONCORD, CLOGP) and programs they write themselves



Practical Computing

- Introduction to practical IM techniques
- Semester 1
  - Email; word processing; web searching; presentation software; spreadsheets; databases; web page authoring
- Semester 2
  - HTML forms; HTML frames; JavaScript basics; SPSS; EndNote….

Research Methods…

- Introduction to research techniques
  - qualitative data collection and management
  - quantitative data collection and management
  - statistics

…and Dissertation Preparation

- Chapter One of the final dissertation
  - research proposal
  - full literature review
  - familiarity with software
- Critique of a previous dissertation

Rationale For The Placements

- A dissertation is an important part of any PGT course
  - distinguishes an MSc from a Diploma or Certificate
- Part of all of our existing MA/MSc programmes
- Substantial industrial involvement now required for EPSRC-funded PGT programmes
  - we've had many successful CASE PhD collaborations in the past, and decided to use these as a model for the dissertation component of the MSc

Timetable

- Mid-November
  - Initial contact with companies to discuss possible projects
- Late-January
  - Final list of projects sent to all students (in 2000/01 had 18 projects from 12 companies, this year have 19 projects from 9 companies)
  - Students asked to select three projects after any necessary discussion with Sheffield supervisors
- Mid-February: final allocations
  - One student per company
  - Student to get one of three choices
  - Take account of Sheffield staff expertise
- Early-June
  - Students leave Sheffield
  - Regular email/phone conversations with Sheffield supervisor
  - At least one on-site visit with submission by 1st September

Example Projects

- Four main types of project
  - Development and/or testing of an existing or novel piece of software for some specific application
  - Development of a Web front-end to an existing system or service
  - Comparison of different programs for some specific application
  - Analysis of chemical and/or biological dataset(s)
- Wide range of application domains, including
  - Substructure searching
  - De novo design
  - Pharmacophore identification
  - Property prediction

Problems In 2000/01

- Some teething problems
  - e.g. programming modules; setting of deadlines
- Industrial Placements
  - Organisational difficulties
    - changes at collaborating companies (site closure, rebuilding)
  - Academic
    - pressure of work in semester two meant that dissertation preparation suffered
    - students didn't make as much use of the Sheffield supervisor during the writing-up as expected
- Hope it is better this year

Successes in 2000/01

- Despite the problems, the placements went well
  - All the students finished on time
  - The final dissertations were generally of a high standard, and a couple of them were very good indeed (one currently being converted to a journal article)
  - Students felt that they received a lot of support, both academic and organisational, from the companies
- Employment Data for 2001 Graduates
  - 5 Chemoinformatics jobs
    - 2 pharmaceutical; 1 software company; 2 academic research
  - 1 general IT
  - 1 Health Service
  - 1 seeking PhD position
  - 1 unknown

Summary

- The emergence of Chemoinformatics as a discipline has led to the development of masters programmes
  - Sheffield; UMIST; Indiana
- The programmes aim to provide students with skills that are currently in high demand
- Collaboration between programmes
  - Currently via exchange of lectures
  - Possibilities for the future include the development of Training Packages