Chemical Informatics

http://www-jmg.ch.cam.ac.uk/cil/partii/

Artemisinin

synthetic antimalarial drug candidate

Jonathan Goodman
jmg11@cam.ac.uk

Robert Glen
rcg28@cam.ac.uk
Chemical Informatics

Lectures 1 - 2: Jonathan Goodman
A: Chemical Information
The use of some of the main databases available will be outlined.

Lectures 3 - 4: Robert Glen
B: Computer representation of molecules and molecular data analysis
Molecules are represented in the computer in different ways for different contexts. Some of these methods will be described. Molecular data can be generated and analysed from these stored molecules. Methods of property generation will be described along with some example analyses.

A: Chemical Information
1. Google and Google Scholar
2. Web of Knowledge
3. Scopus
4. Chemical Abstracts (SciFinder Scholar)
5. Reaxys
6. Cambridge Structural Database
7. ChemSpider

Practical Assessment
You must complete the two practicals for the course, one on each section of the course.
The practicals must be submitted by Friday December 6th, 2019

Marking: You will get seven out of ten for a workmanlike report; eight, or possibly more, for one which shows particular insight; six or five or less if you make serious mistakes; zero if you do not submit a report.

Resources: linked from the Chemistry Department Library webpages:
http://www-library.ch.cam.ac.uk/

Admitto passwords available from:
https://apps.ch.cam.ac.uk/password-collection/password-collect
1. Google and Google Scholar

Chemical informatics is the computer-assisted storage, retrieval and analysis of chemical information, from data to chemical knowledge. More than a hundred million small molecules have been identified over the last century. Finding information from these data requires computational tools.

Google and Google Scholar are not able to search for chemical structures, and it is not able to limit its searches to the scientific literature. Databases contain errors. Some of these are because papers make mistakes in their references, and the database has copied the original authors' mistakes. Some of these are new errors introduced in the abstracting process.

2. Web of Knowledge

This resource grew out of the Science Citation Index – a book that listed the papers that papers cited. Starting with an interesting paper, it was possible to find out which newer papers had cited it, and so work forwards through the literature. Without this tool, it was only possible to find out which papers were cited by a new paper, and so work backwards through the literature. It also made it possible to find out how well cited papers were, and so, how influential the papers and their authors were. A significant paper will probably be cited many times.

On-line access to the Web of Knowledge is available using Raven userIDs and passwords.

http://wok.mimas.ac.uk/

3. Scopus

The University of Cambridge has a subscription to Scopus. This covers similar ground to the Web of Knowledge, and has some additional features, but surveys fewer years.

http://www.scopus.com/

4. SciFinder Scholar

The American Chemical Society's Chemical Abstracts Service (CAS) has been abstracting chemistry and chemistry-related scientific papers for over a century. A large team of experts abstracts papers.

Molecules are indexed using the CAS registry number – an arbitrary number assigned as the papers are abstracted. These registry numbers are widely used in databases, and provide a convenient and concise index of molecules. However, it is not possible to create a CAS registry number for a new molecule without consulting CAS. For CAS, the fundamental unit of scientific knowledge is the published paper (or patent). Each paper has its own abstract, which highlights the
important contributions of the paper. SciFinder Scholar is a convenient program for searching the huge CAS database. For on-line access to Chemical Abstracts, consult:

http://www-library.ch.cam.ac.uk/scifinder/

You will need to register to use the software; this can be done on-line and is quite quick.

5. Reaxys
A second database covers more than a century of chemical information: Beilstein. This is based on the idea that a molecule, and not a paper, is the fundamental unit of chemistry. The database can be compared to CAS, which regard the scientific paper as the fundamental unit of knowledge. However, searching CAS is likely to produce different results to searching Beilstein, and, ideally, both should be used. Reaxys provides a method to search the Beilstein database, and integrates it with an inorganic molecule database, Gmelin, and a database of patents. A particular strength of Reaxys is that it helps to extract synthetic procedures from papers and patents.

http://www.reaxys.com/

6. Cambridge Structural Database (CSD)
This database is an index of the world's X-ray crystal structures of organic molecules. All of the data included in the database is of a high standard. However, even crystal structures can contain misleading data. All of the data in the CSD has been carefully checked.

https://www.ccdc.cam.ac.uk/structures

7. ChemSpider
ChemSpider is a free-to-access collection of compound data from across the web, provided for the benefit of the community by the Royal Society of Chemistry (RSC). It aggregates chemical structures and their associated information into a single searchable repository and makes it available to everybody.

http://www.chemspider.com/

8. PubChem
“Quickly find chemical information from authoritative sources”

Resources

Chemistry department library web pages:
http://www-library.ch.cam.ac.uk/

Course page: specific information on this course, and any new information that becomes available.
http://www-jmg.ch.cam.ac.uk/cil/partii/

Some of the department's web pages require a *Raven* account for access. You should already have a *Raven* account and password.

MCS
The chemistry MCS, which you can use in the library, has the Cambridge Crystallographic Database installed. You should already have a MCS account.

ChemDraw:
The department has purchased an annual university-wide license for ChemDraw. For details, look for the link on the library web page.
Practical A: Chemical Informatics. Answer all four parts (a), (b), (c), (d)
(The practical for the second half of the course must also be completed and is described separately)

The results of practical A should be submitted, by e-mail, to Jonathan Goodman (jmg11@cam.ac.uk) by the end of Friday December 6th. A printed version is not required. These questions require only short answers. Do not submit every piece of information that you find, but summarise your results to answer the questions succinctly. Do not send PDFs, .doc, .odt, etc. Send you answer in the e-mail without an attachment.

(a) What was Professor W Clark Still's most cited paper? (Give the title and the full reference)
How many papers have tried to cite it? How many citations can Scopus and SciFinder give for this paper? E-mail your results, with the date you did the search and a one or two sentences of explanation, to jmg11@cam.ac.uk as soon as you can, with the subject line beginning "WCStill:"

Hint: the answer is probably more than 8 700
A graph of the results will be build up on the website as they are received
The graphs generated by the last years' classes are available

(b) The 2010 Nobel Prize for chemistry was awarded for palladium-catalysed coupling reactions (http://nobelprize.org/nobel_prizes/chemistry/laureates/2010/). Is it possible to do a transition-metal-free Suzuki coupling reaction? Find a paper that says this has been done and one that contradicts this view. In your answer, give both references, and comment briefly on whether such a reaction is possible. Send your answer by e-mail, with the subject line "Palladium free:"

(c) Jan Hendrik Schönh has reported a series of breakthroughs in semiconductor physics. Look him up in (i) Google Scholar, (ii) Web of Knowledge. Do they give a consistent picture of his contribution? Now try Google. E-mail your answer with the subject line "Schon:“. Only one or two sentences are needed.
(d) Contribute to global chemical knowledge

Details of part (d) will be available on the course website.