Welcome!

Mass Spectrometry meets Cheminformatics
Tobias Kind and Julie Leary
UC Davis

Course 1: General Introduction

Class website: CHE 241 - Spring 2008 - CRN 16583
Slides: http://fiehnlab.ucdavis.edu/staff/kind/Teaching/
PPT is hyperlinked – please change to Slide Show Mode
What is ChemInformatics?

Chemometrics est. 1975
Cheminformatics est. 1998
Who uses Cheminformatics?

All parts of chemistry heavily depend on cheminformatics. Life sciences, biochemistry, drug industries use cheminformatics.

20 years ago: 80% in lab – 20% in front of computer
Now: 20% in lab - 70% in front of computer (*)

Examples:

- **Organic chemistry** – automated reaction planning, Beilstein search
- **Physical chemistry** – modeling of structure properties (boiling points)
- **Inorganic chemistry** – ligand bond interactions
- **Analytical chemistry** – structure elucidation of small compounds
- **Biochemistry** – protein/small molecule interaction networks

(*) 10% fixing and installing new programs
Motivation for Mass Spectrometry meets ChemInformatics

To be a master of spectra you need to be a master of structures in the first place.

→ Complex MS data interpretations only possible with software
→ MS data obtained by hyphenated techniques (GC-MS, LC-MS)
→ Mass spectral database search and structure search routinely are used
→ Mass spectrometers deliver multidimensional data
Computer Illiteracy – a threat to your research

Your computer is your friend
You don’t have a computer? You don’t have a friend (just kidding)

• Assume you have a computer:
  Please step forward name: CPU, speed, memory, hard disk, OS

• You are a chemist, biochemist, biologist:
  Please step forward name: Computer language or DB you know

OS = operating system; DB = database, CPU = central processing unit
# Fighting Computer Illiteracy - name your PC

<table>
<thead>
<tr>
<th>Component</th>
<th>Brands/Models</th>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td>INTEL, AMD, IBM, HP</td>
<td>Pentium, Opteron, Core Duo 2-3 Ghz</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>GEIL, KINGSTON</td>
<td>DDR, DDR2</td>
</tr>
<tr>
<td><strong>Hard disk</strong></td>
<td>SEAGATE, WD</td>
<td>Raptor, Barracuda, Cheetah 100-1000 GByte</td>
</tr>
<tr>
<td><strong>OS</strong></td>
<td>MICROSOFT, LINUX</td>
<td>Windows, Linux, OSX, Virtual OS</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td></td>
<td>C, Basic, Perl, JAVA</td>
</tr>
</tbody>
</table>

- Bit < Byte < kByte < MByte < GByte
- Single Core < Dual Core < QuadCore < MultiCore
- MFLOP/s < GFLOP/s < TFLOP/s < PFLOP/s
- 1 Thread < Dual Thread < MultiThreaded
Computer Illiteracy – learn a programming language

Why should you?

20% lab time – 80% computer time
Mass spectrometers deliver data – not results

Why shouldn't you? (fake reasons)

You are too old to learn…
You are not good with computers…
Your have more important research to do…
You are so rich you have programmers who work for you…
Computer Illiteracy – learn a programming language

• Learn any language which has a large code and user base (JAVA, Perl, Visual Basic)
• Use IDEs with automatic code completion like MS Visual Express or Eclipse
• Don’t re-invent code - use (and document) code search engines like
  koders.com;
google.com/codesearch
krugle.com

Do *not* learn these working but esoteric languages
There are 1123 programming languages http://99-bottles-of-beer.net/
Program development – Eclipse for JAVA example

```java
package arraybench;

import java.util.*;

public class ArrayListTest {

    public static StringBuffer stringToList(List list) {
        StringBuffer result = new StringBuffer(1024);
        list.add("ABCD");
        list.add("BFGH");
        list.add("IJKL");
        list.addAll(list);
        result.append(list);
        result.append(result.contains("IJKL");
        result.append(result.containsAll(new ArrayList()
            {add("ABCD");add("BFGH");}));
    }

    public static void main(String[] args) {
        List list = new ArrayList();
        list.add("ABCD");
        list.add("BFGH");
        list.add("IJKL");
        list.addAll(list);
        StringBuffer result = stringToList(list);
        System.out.println(result);
    }
}
```

Text output:

```
Performance with queue length = 100000
- class java.util.ArrayList managed 234 iterations in 2000ms
- class java.util.LinkedList managed 32008 iterations in 2000ms
- class arraybenchCircularArrayList managed 234 iterations in 2000ms
- class org.objectweb.proactive.core.util.CircularArrayList managed 234 iterations in 2000ms
```

JAVA or C code
Computer Illiteracy – your emergency helpers

Regular expressions; SQL database requests; EXCEL VBA scripts or Perl scripts are special tools for data handling (Swiss army knifes).

**Regular expressions** (RegEx) are used for finding and replacing text:

- `[0-9]` – represents all numbers
- `[a-z]` – represents all small letters
- `
` – represents new line (CR/LF)
- `	` – represents TAB

Examples:

- `\n\n` – find double empty lines
- `\t` replace with spaces “ “
- find two numbers in brackets `([0-9][0-9])`

Learn about RegEx

**SQL** is used for programming databases

### Large Database Table

<table>
<thead>
<tr>
<th>yr</th>
<th>subject</th>
<th>winner</th>
</tr>
</thead>
<tbody>
<tr>
<td>1901</td>
<td>Chemistry</td>
<td>Jacobus H. van ’t Hoff</td>
</tr>
<tr>
<td>1902</td>
<td>Chemistry</td>
<td>Emil Fischer</td>
</tr>
<tr>
<td>1903</td>
<td>Chemistry</td>
<td>Svante Arrhenius</td>
</tr>
<tr>
<td>1904</td>
<td>Chemistry</td>
<td>Sir William Ramsay</td>
</tr>
<tr>
<td>1905</td>
<td>Chemistry</td>
<td>Adolf von Baeyer</td>
</tr>
<tr>
<td>1906</td>
<td>Chemistry</td>
<td>Henri Moissan</td>
</tr>
<tr>
<td>1907</td>
<td>Chemistry</td>
<td>Eduard Buchner</td>
</tr>
<tr>
<td>1908</td>
<td>Chemistry</td>
<td>Ernest Rutherford</td>
</tr>
<tr>
<td>1909</td>
<td>Chemistry</td>
<td>Wilhelm Ostwald</td>
</tr>
<tr>
<td>1910</td>
<td>Chemistry</td>
<td>Otto Wallach</td>
</tr>
<tr>
<td>1913</td>
<td>…</td>
<td></td>
</tr>
</tbody>
</table>

### SQL query

```
SELECT yr, subject, winner
FROM nobel
WHERE yr = 1909 and subject = 'chemistry'
```

### Result

<table>
<thead>
<tr>
<th>yr</th>
<th>subject</th>
<th>winner</th>
</tr>
</thead>
<tbody>
<tr>
<td>1909</td>
<td>Chemistry</td>
<td>Wilhelm Ostwald</td>
</tr>
</tbody>
</table>

Visit the SQL Zoo
Regular Expressions – example MS data

Task: create a list of 4 columns with names, formulas, CAS numbers and peaks
Problem: 24,000 lines of mass spectral data (*.msp)
Program: Textpad (WIN), Smultron (Mac)
Regular Expressions – example MS data

**Solution:** replace Enter (\n) with TAB (\t) and use Replace ALL
Regular Expressions – example MS data

Solution: copy only lines of interest (Mark ALL – Copy Bookmarked Lines)
Regular Expressions – Result for MS data

Solution: Replace redundant code with nothing, copy tab separated file to EXCEL

Result: 1:30 min for RegEx job
(1 hour manually?)

Average spectrum size: 70 peaks
Minimum size: 5 peaks
Maximum size: 439 peaks
Most spectra have 35 and 45 peaks
Be prepared – visualize your structures

Try Marvin Space via Webstart
**Be prepared - StereoIsomers**

How many stereoisomers can you expect from glucose (KEGG)?

Example: separation of species with ion mobility MS (FAIMS)

Example calculated with [MarvinView](http://www.marvinjs.org) (via [JAVA Webstart](http://www.javawebstart.com))
Be prepared – Resonance (electron shifts)

What are possible resonant structures?
Important for mass spectral interpretation (electron impact, electrospray)

Example calculated with MarvinView

Start via WebStart
Be prepared – Tautomers

How many tautomers can you expect?
Important for mass spectral interpretations.

Example calculated with MarvinView
Start via WebStart
Mass spectral database search – know what exists

How many mass spectra with formula C11H8O3 in NIST DB?

Result: 19 for C_{11}H_{8}O_{3} in NIST05 DB
Download NIST-MS-Search
Mass spectral interpretation

Assign structural elements to mass spectral peaks

Download Mass Spectrum Interpreter Version 2
Structure search – know what could be possible

How many compounds (isomer structures) are found in public databases?

Result:
272 for $\text{C}_{11}\text{H}_{8}\text{O}_{3}$

http://www.chemspider.com/
Molecular Weight Calculator

Calculate isotopic masses
Find formulas from masses
Calculate isotopic patterns

Download **MWTWIN**
Stay tuned – new mass spectrometry publications via Yahoo Pipes

Mass Spectrometry Journal Articles
Pipe Web Address: http://pipes.yahoo.com/pipes/pipes/info?_id=kK85gta72xGq5xWCIw4Mw

Glass coating for PDMS microfluidic channels by sol-gel methods
Sol-gel method is utilized to create a glass-like coating on PDMS microfluidic channels, which greatly increases chemical resistance of the microfluidic devices.

Examination of laser microbeam cell lysis in a PDMS microfluidic channel using time-resolved imaging
The delivery of a highly-focused [small lambda] = 582 nm pulsed laser microbeam with a 540 ps duration is shown to produce rapid lysis of single suspension cells in a PDMS microfluidic channel.

Tandem Immunoprecipitation of Phosphotyrosine-Mass Spectrometry (TIPY-MS) Indicates C19orf19 Becomes Tyrosine-Phosphorylated and Associated with Activated Epidermal Growth Factor Receptor
Jiefei Yang, Paul Taylor, Eleonora Joyceva, Jonathan R. St-Jean, Lily L. Jin, Ana Nikolic, Xiaopeng Gu, Zhi Hua Li, Suzanne Trudel, and Michael F. Moran Web Release Date: Thu, 14 Feb 2008 00:00:00 EST (Article) DOI: 10.1021/pr700363d

Mixed-Effects Statistical Model for Comparative LC-MS Proteomics Studies
D. S. Dai, K. K. Anderson, E. A. Panisko, S. C. Purvine, R. Fang, M. E. Monroe, and S. E. Baker Web Release Date: Wed, 6 Feb 2008 00:00:00 EST (Article) DOI: 10.1021/pr7004411
The Last Page - What is important to remember:

Learn about CPU type, memory, hard disks, bits and bytes;
→ shock you colleagues with random questions about their computer

Think about automation, thinks you would like to do (even if you can’t)
→ shock you colleagues with a small computer script

Use regular expressions for stupid or boring jobs
→ you delete/replace data more than 3x - remember RegEx, RegEx, Regex

Use scripting languages for small problems (EXCEL VBA, PERL)
→ steal some small examples and color your EXCEL data in rainbow color

Generate yourself a collection of programs and databases for MS
→ try such programs in a Virtual Machine without messing up your system
Tasks:

The PowerPoint slides are all hyperlinked.
1) Download and install the mentioned tools (JAVA required)
2) Visit the databases and online websites
3) Repeat shown examples
4) Check notes in PPT for additional information
Literature:
Check notes and links in PPT
Links:

Used for research: (right click – open hyperlink)

- http://www.computerhistory.org/microprocessors/
- http://www.google.com/search?hl=en&q=holy+crap+site%3A.edu&btnG=Search
- http://www.google.com/search?hl=en&q=editor+for+mac+regular+expressions&btnG=Search
- SQL learning http://sqlzoo.net/
- http://www.microsoft.com/windows/products/winfamily/virtualpc/default.mspx (Virtual PC or VMWare - run multiple WINDOWS or LINUX under WIN or vice versa)

Of general importance for this course:
http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Structure_Elucidation/