Welcome!

Mass Spectrometry meets Cheminformatics
Tobias Kind and Julie Leary
UC Davis

Course 9: Prediction and simulation of mass spectra

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
History of artificial intelligence and mass spectrometry

Dendral project at Stanford University (USA)
Started in 1960s
Pioneered approaches in artificial intelligence (AI)

Aim:
Prediction of isomer structures from mass spectra
Idea: Self-learning or intelligent algorithm

Participants:
Lederberg, Sutherland, Buchanan, Feigenbaum, Duffield, Djerassi, Smith, Rindfleisch, many others…

Figure: Heuristic DENDRAL:
A Program for Generating Explanatory Hypotheses in Organic Chemistry
Prediction and simulation of mass spectra

A) Prediction of the isomer structure or substructures from a given mass spectrum
   The structure is directly deduced from the mass spectrum or generated by
   a molecular isomer generator or existing structures can be found in a structure database

B) Simulation of a mass spectrum from a given isomer structure
   The mass spectral peaks and abundances are generated by a machine learning algorithm
   The structures can be obtained from a isomer database (PubChem, LipidMaps)
   or a sequence database (Swiss-Prot, NCBI) in case of proteins
Application of machine learning for detection of substructures from mass spectra

Data Preparation
- Basic Statistics, Remove extreme outliers, transform or normalize datasets, mark sets with zero variances

Feature Selection
- Predict important features with MARS, PLS, NN, SVM, GDA, GA; apply voting or meta-learning

Model Training + Cross Validation
- Use only important features, apply bootstrapping if only few datasets; Use GDA, CART, CHAID, MARS, NN, SVM, Naive Bayes, kNN for prediction

Model Testing
- Calculate Performance with Percent disagreement and Chi-square statistics

Model Deployment
- Deploy model for unknown data; use PMML, VB, C++, JAVA

What is machine learning?
Prediction of substructures from mass spectra

Working examples for EI mass spectra:
Varmuza classifiers in **AMDIS** and **MOLGEN-MS**

Substructure algorithm (Stein S.E.)
Implemented in NIST-MS search program

Mass spectral classifiers for supporting systematic structure elucidation
Chemical Substructure Identification by Mass Spectral Library Searching
Substructures deduced from mass spectra for generation of isomer structures

1) **Molecular formula** must be known - can be detected from molecular ion and isotopic pattern
2) **Good-list** (substructure exists) and **bad-list** (substructure not existent) approach
3) Sub-structures are combined in **deterministic** or **stochastic** (random) manner
4) **Database** or **molecular isomer generator** (combinatorial, graph theory) approach for generating or finding possible structure candidates

**Example:**
Molecular formula C₆ClH₅O;
calculated from molecular ion

**Goodlist:**
-benzene
-hydroxy
-chlorine

**Badlist:**

<table>
<thead>
<tr>
<th>Database (Chemspider): 25 hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>(including all possible existing structures)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MOLGEN Demo:</th>
</tr>
</thead>
<tbody>
<tr>
<td>All constructed isomers: 8372</td>
</tr>
</tbody>
</table>

Total: 3 possible results
Simulation of mass spectra

Why is simulation of mass spectral fragmentation important?

Imagine – you have a **structure database** of all molecules
Imagine – you can **simulate mass spectra** for all these molecules
Imagine – you can **match** your **experimental spectra** against a database of **calculated spectra**

If the calculation is simple the database is not needed;
In-silico MS fragments can be calculated on-the-fly
Simulation of alkane mass spectra (I)

**Approach**
Use of artificial neural networks (ANN) (machine learning)
Electron impact spectra 70 eV
Substructure descriptors were used for calculation
Selection of 44 \( m/z \) positions – training was performed for correct intensity

117 noncyclic alkanes and 145 noncyclic alkenes
training set: 236 molecules
prediction set: 26 compounds

**Problems**
Prediction or validation set very small (should be 30%)
Prediction of molecular ion (usually very low abundant)
Overfitting possible, works only for selected substance classes

Source: Jalali-Heravi M. and Fatemi M. H.; *Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network*
Simulation of alkane mass spectra (II)

2,3,3-trimethylpentane (a and b) and 2,3,4-trimethylpentane (c and d).

Source: Jalali-Heravi M. and Fatemi M. H.; Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network
Analytica Chimica Acta; Elsevier permission use for coursepack/classroom material

Structures: Chemspider
Simulation of lipid tandem mass spectra (I)

Similar structures; plus CH2 in side chains sn1 and sn2; double bonds possible
Similar and almost constant fragmentation rules
Loss of head group (diagnostic ion in MS and MS/MS spectrum)
Loss of rest one (R1) and rest two (R2) can be observed in MS/MS spectrum

Simulation of lipid tandem mass spectra (II)

Simulation of tandem mass spectra or MS/MS fragment data from LipidMaps

Experimental Mass spectrum

In-silico prediction of MS/MS mass spectral fragments

<table>
<thead>
<tr>
<th>Mass</th>
<th>C</th>
<th>DB</th>
<th>Abbrev.</th>
<th>M-sn1+H</th>
<th>M-sn1-H2O+H</th>
<th>M-sn2+H</th>
<th>M-sn2-H2O+H</th>
<th>sn1 acid(-)</th>
<th>sn2 acid(-)</th>
<th>HG</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>797.5180</td>
<td>31</td>
<td>0</td>
<td>14:0/17:0</td>
<td>587.3196</td>
<td>569.309</td>
<td>545.2727</td>
<td>527.2621</td>
<td>227.2011</td>
<td>269.2481</td>
<td>GPIns</td>
<td>C_{40}H_{77}O_{10}P</td>
</tr>
<tr>
<td>797.5180</td>
<td>31</td>
<td>0</td>
<td>17:0/14:0</td>
<td>545.2727</td>
<td>527.2621</td>
<td>587.3196</td>
<td>569.309</td>
<td>269.2481</td>
<td>227.2011</td>
<td>GPIns</td>
<td>C_{40}H_{77}O_{10}P</td>
</tr>
<tr>
<td>796.5128</td>
<td>37</td>
<td>5</td>
<td>17:0/20:5(5Z,8Z,11Z,14Z)</td>
<td>544.2675</td>
<td>526.2569</td>
<td>512.2988</td>
<td>494.2882</td>
<td>269.2481</td>
<td>301.2168</td>
<td>GPSer</td>
<td>C_{43}H_{74}NO_{10}P</td>
</tr>
<tr>
<td>796.5128</td>
<td>37</td>
<td>5</td>
<td>20:5(5Z,8Z,11Z,14Z)17:0</td>
<td>512.2988</td>
<td>494.2882</td>
<td>544.2675</td>
<td>526.2569</td>
<td>301.2168</td>
<td>269.2481</td>
<td>GPSer</td>
<td>C_{43}H_{74}NO_{10}P</td>
</tr>
<tr>
<td>796.5856</td>
<td>37</td>
<td>4</td>
<td>17:0/20:4(5Z,8Z,11Z,14Z)</td>
<td>544.3403</td>
<td>526.3297</td>
<td>510.3559</td>
<td>492.3453</td>
<td>269.2481</td>
<td>303.2324</td>
<td>GPCho</td>
<td>C_{45}H_{82}NO_{10}P</td>
</tr>
<tr>
<td>796.5856</td>
<td>37</td>
<td>4</td>
<td>20:4(5Z,8Z,11Z,14Z)17:0</td>
<td>510.3559</td>
<td>492.3453</td>
<td>544.3403</td>
<td>526.3297</td>
<td>303.2324</td>
<td>269.2481</td>
<td>GPCho</td>
<td>C_{45}H_{82}NO_{10}P</td>
</tr>
</tbody>
</table>

Spectrum Source: Lipidmaps.org
Simulation or prediction of oligosaccharide spectra (carbohydrate sequencing)

Consistent building blocks (sugars)
Consistent fragmentation allows in-silico fragment prediction
Pre-calculated fragments from known structures can be stored in database (use NIST-MS-Search)
Algorithm works also on-the-fly without database
De-novo algorithms work for truly unknown structures

See Oscar and FragLib
See GlySpy
Simulation of peptide fragmentations
(De-novo sequencing of peptides)

**Principle:**
De-novo sequencing of peptides (determine amino acid sequences)
De-novo algorithms can perform permutations and combinatorial calculations from all 20 amino acids (superior if the sequence is not found in a database)
Highly dependent on good mass accuracy (less than 1 ppm) of precursor ion and MS/MS fragments
Generate match score by matching in-silico fragments against experimental MS/MS spectrum

**Problems:**
Leucine and isoleucine have same mass
Post translational modifications (PMTs)
Missing fragment peaks

Picture source: MWTWIN help file2 (Monroe/PNNL)
Picture 2 source: Tandem mass spectrometry data quality assessment by self-convolution
Keng Wah Choo and Wai Mun Tham http://www.biomedcentral.com/1471-2105/8/352
The Last Page - What is important to remember:

Fragmentation and rearrangement rules and ion physics can be programmed into algorithms
→ Abundance calculations are problematic

Prediction of isomer substructures from mass spectra is possible
→ Works for reproducible mass spectra

A simplified simulation of mass spectra and simulation of fragmentation pattern
is only possible for certain molecule classes
→ Works only for peptides, lipids, oligosaccharides, alkanes
→ Does not work for all other molecules
→ Does not work with complex (side chain) modifications

Machine Learning Methods for simulation and prediction of mass spectra
require a large pool of diverse experimental mass spectra and MS^n spectra for training
Tasks (42 min):

Download one of the following tools:
MOLGEN, MOLGEN-MS, AMDIS, OMMSA, OSCAR or any free/commercial/demo program for in-silico peptide fragment determination or de-novo sequencing. Report on use.
Literature (36 min):

Mathematical tools in analytical mass spectrometry [DOI]
Metabolomics, modelling and machine learning in systems biology – towards an understanding of the languages of cells [DOI]
Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry [PDF]
Mass Analysis Peptide Sequence Prediction [LINK]
Links:

Used for research: (right click – open hyperlink)

http://scholar.google.com/scholar?hl=en&q=%22Simulation+of+mass+spectra
http://scholar.google.com/scholar?num=100&hl=en&lr=&safe=off&q=+Simulation+of+%22mass+spectral+fragmentation
http://www.google.com/search?num=100&hl=en&safe=off&q=in-silico+prediction+tandem+mass+spectra&btnG=Search
http://www.aseanbiotechnology.info/Abstract/21020883.pdf
http://www.google.com/search?hl=en&q=GNU+polyxmass%2C&btnG=Google+Search
http://www.google.com/search?hl=en&q=C41H76N2O15&btnG=Google+Search
http://www.google.com/search?num=100&hl=en&safe=off&q=MOLGEN+MS&btnG=Search
http://www.google.com/search?hl=en&q=G.+L.+Sutherland&btnG=Google+Search

GlySpy and the Oligosaccharide Subtree Constraint Algorithm (OSCAR)
See Mass Frontier for further discussion
MOLGEN-MS [LINK]

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