Welcome!

Mass Spectrometry meets Cheminformatics
WCMC Metabolomics Course 2013
Tobias Kind

Course 7: Prediction and simulation of mass spectra

http://fiehnlab.ucdavis.edu/staff/kind
**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…

**Status:**
Failed; but inspirational until today (50 years later)
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 an isomer database (PubChem, LipidMaps) or a sequence database (Swiss-Prot, NCBI) in case of proteins.
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 \( \text{C}_6\text{ClH}_5\text{O} \); calculated from molecular ion

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

**Badlist:**

Database ([Chemspider](https://www.chemspider.com)): 25 hits (including all possible existing structures)

MOLGEN Demo:
All constructed isomers: 8372

**Total:** 3 possible results
Aristo – ontology classification of EI mass spectra

http://www.ionspectra.org/aristo/
Application: Decision tree supported substructure prediction of metabolites from GC-MS profiles

Hummel J, Strehmel N, Selbig J, Walther D, Kopka J.

http://gmd.mpimp-golm.mpg.de/

Decision tree supported substructure prediction of metabolites from GC-MS profiles.
Hummel J, Strehmel N, Selbig J, Walther D, Kopka J.
METFRAG: In silico fragmentation for mass spectral identification

Submit: mass spectral information.
Result: Ranked hit list of molecules
http://msbi.ipb-halle.de/MetFrag/
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

(*) Theoretical spectrum is depiction only, not truly simulated.
General methods for simulation of mass spectra

Ab-initio or de-novo first principle methods

Superior, because they solve the problem at the root;
Slower to implement

Heuristics or rule based algorithms

Solve practical problems;
Faster to implement


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 $44 \frac{m}{z}$ 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 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

Source: Congruent Strategies for Carbohydrate Sequencing.
3. OSCAR: An Algorithm for Assigning Oligosaccharide Topology from MSn Data
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
In-silico fragmentation with MassFrontier using fragmentation library of 20,000 mechanisms from literature

Result fragmentations are represented as bar code spectra (same abundance)
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
Combinatorial scaffold library design

sn1 = alkyl or acyl rest

sn2 = alkyl or acyl rest

head group

Functional group (variable)  Linker  Scaffold (conserved)

C16  

C14  

C12  

C10  

+ LipidMaps nomenclature name generation
+ accurate isotopic fragment calculation
+ mass spectral peak annotation
+ heuristic peak abundance modeling (CID voltage dependent)
+ conversion into mass spectral library format

Source: LipidBlast
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>GPIIns</td>
<td>C_{40}H_{77}O_{15}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>GPIIns</td>
<td>C_{40}H_{77}O_{13}P</td>
</tr>
<tr>
<td>796.5128</td>
<td>37</td>
<td>5</td>
<td>17:0/20:5(5Z,8Z,11Z,14Z,17Z)</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>GPISer</td>
<td>C_{43}H_{74}NO_{19}P</td>
</tr>
<tr>
<td>796.5128</td>
<td>37</td>
<td>5</td>
<td>20:5(5Z,8Z,11Z,14Z,17Z)/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>GPISer</td>
<td>C_{43}H_{74}NO_{19}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_{48}H_{82}NO_{8}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_{48}H_{82}NO_{8}P</td>
</tr>
</tbody>
</table>

Spectrum Source: Lipidmaps.org
LipidBlast MS/MS mass spectral modeling

In-silico mass spectra:
• **m/z fragments** and **abundance** calculation required
• **statistical** (computer derived) and **heuristic rules** (experience of a human expert)
## LipidBlast in-silico modeling of lipid tandem mass spectra

<table>
<thead>
<tr>
<th>Number</th>
<th>LipidClass</th>
<th>Short</th>
<th>Number compounds</th>
<th>Number MS/MS spectra with different adducts</th>
<th>Number MS/MS LIBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Phosphatidylcholines</td>
<td>PC</td>
<td>5476</td>
<td>10952</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Lysophosphatidylcholines</td>
<td>lysoPC</td>
<td>80</td>
<td>160</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Plasmenylphosphatidylcholines</td>
<td>plasmenyl-PC</td>
<td>222</td>
<td>444</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Phosphatidylethanolamines</td>
<td>PE</td>
<td>5476</td>
<td>16428</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Lysophosphatidylethanolamines</td>
<td>lysoPE</td>
<td>80</td>
<td>240</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>Plasmenylphosphatidylethanolamines</td>
<td>plasmenyl-PE</td>
<td>222</td>
<td>666</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Phosphatidylyserines</td>
<td>PS</td>
<td>5123</td>
<td>15369</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>Sphingomyelins</td>
<td>SM</td>
<td>168</td>
<td>336</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>Phosphatidic acids</td>
<td>PA</td>
<td>5476</td>
<td>16428</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>Phosphatidylinositosls</td>
<td>PI</td>
<td>5476</td>
<td>5476</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Phosphatidylglycerols</td>
<td>PG</td>
<td>5476</td>
<td>5476</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>Cardiolips</td>
<td>CL</td>
<td>25426</td>
<td>50852</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>Ceramide-1-phosphates</td>
<td>CerP</td>
<td>168</td>
<td>336</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>Sulfatides</td>
<td>ST</td>
<td>168</td>
<td>168</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>Gangliosides</td>
<td>[glycan]-Cer</td>
<td>880</td>
<td>880</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>Monoacylglycerols</td>
<td>MG</td>
<td>74</td>
<td>148</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>Diacylglycerols</td>
<td>DG</td>
<td>1764</td>
<td>3528</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>Triacylglycerols</td>
<td>TG</td>
<td>2640</td>
<td>7920</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>Monogalactosyldiacylglycerols</td>
<td>MGDG</td>
<td>5476</td>
<td>21904</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>Digalactosyldiacylglycerols</td>
<td>DGDG</td>
<td>5476</td>
<td>10952</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>Sulfoquinovosyldiacylglycerols</td>
<td>SQDG</td>
<td>5476</td>
<td>5476</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>Diacylated phosphatidylinositol</td>
<td>Ac2PIM1</td>
<td>144</td>
<td>144</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>Diacylated phosphatidylinositol dimannoside</td>
<td>Ac2PIM2</td>
<td>144</td>
<td>144</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>Triacylated phosphatidylinositol dimannoside</td>
<td>Ac3PIM2</td>
<td>1728</td>
<td>1728</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>Tetraacylated phosphatidylinositol dimannoside</td>
<td>Ac4PIM2</td>
<td>20736</td>
<td>20736</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>Diphosphorylated hexaacyl Lipid A</td>
<td>LipidA-PP</td>
<td>15625</td>
<td>15625</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>All libraries</strong></td>
<td></td>
<td><strong>119200</strong></td>
<td><strong>212516</strong></td>
<td><strong>50</strong></td>
</tr>
</tbody>
</table>

Covered adduct ions: 

LipidBlast MS/MS search with NIST MS search program using precursor search and dot-product match can be used by practitioners.

Experimental MS/MS list

Library hit scores

in-silico MS/MS

exp. MS/MS

Search speed ~ 100 MS/MS spectra per second (without GUI)
**LipidBlast example: ion trap mass spectrometer**

**Name:** PC 34:1; [M+Na]+; GPCho(16:0/18:1(11E))

**MW:** 782 **ID#:** 42511 **DB:** lipidblast-pos

**Comment:** Parent=782.56759 Mz_exact=782.56759 ; PC 34:1; [M+Na]+;GPCho(16:0/18:1(11E)); C42H82NO8P

8 m/z Values and Intensities:

<table>
<thead>
<tr>
<th>m/z</th>
<th>intensity</th>
<th>formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>723.49409</td>
<td>999.00</td>
<td>[M+Na]-C3H9N</td>
</tr>
<tr>
<td>599.50155</td>
<td>600.00</td>
<td>[M+Na]-C5H14NO4P (-183)</td>
</tr>
<tr>
<td>544.33807</td>
<td>20.00</td>
<td>[M+Na]-sn1</td>
</tr>
<tr>
<td>526.32751</td>
<td>20.00</td>
<td>[M+Na]-sn1-H2O</td>
</tr>
<tr>
<td>518.32243</td>
<td>20.00</td>
<td>[M+Na]-sn2</td>
</tr>
<tr>
<td>500.31187</td>
<td>20.00</td>
<td>[M+Na]-sn2-H2O</td>
</tr>
<tr>
<td>467.25401</td>
<td>40.00</td>
<td>[M+Na]-59-sn1</td>
</tr>
<tr>
<td>441.23837</td>
<td>40.00</td>
<td>[M+Na]-59-sn2</td>
</tr>
</tbody>
</table>

Fatty acyl side chains (sn1, sn2) best detected in negative ionization mode

**1st Hit group**

PC 34:1  
(42 candidates)
LipidBlast example: Hybrid Ion-Trap (IT) and Time-of-Flight (TOF)

Name: SQDG 34:3; [M-H]-; SQDG(16:0/18:3(6Z,9Z,12Z))
MW: 815 ID#: 106150 DB: lipidblast-neg
Comment: Parent=815.4972 Mz_exact=815.4972 ; SQDG 34:3; [M-H]-; SQDG(16:0/18:3(6Z,9Z,12Z)); C43H76O12S
559.25784 300.00 [M-H]-sn1
537.27348 300.00 [M-H]-sn2
277.21662 100.00 sn2 FA
255.23226 100.00 sn1 FA
225.00690 999.00 fragment C6H9O7S

LipidBlast example: ion trap mass spectrometer

Name: LipidA PP [14/14/14/14/30-(12)/30-(14)]; [M-H];
MW: 1796.6164  DB: lipidblast-neg
Comment: Parent=1796.21157 Mz_exact=1796.21157; LipidA PP [14/14/14/14/30-(12)/30-(14)]; [M-H];
9 largest peaks:
1552.00785  999.00 | 1698.23467  600.00 |
1796.21157  500.00 | 1498.05715  300.00 |
1470.02587  300.00 | 1568.00277  250.00 |
1596.03405  250.00 |

9 m/z Values and Intensities:
1796.21157  500.00  [M-H];
1714.22959  50.00  [M-H]-PO3H;
1698.23467  600.00  [M-H]-PO4H3;
1596.03405  250.00  [M-H]-PO4H3-R2'-O-FA;
1568.00277  250.00  [M-H]-PO4H3-R3'-O-FA;
1552.00785  999.00  [M-H]-R2 acyl FA || [M-H]-R3 acyl FA;
1498.05715  300.00  [M-H]-PO4H3-R2'-O-FA;
1470.02587  300.00  [M-H]-PO4H3-R3'-O-FA;
1454.03095  250.00  [M-H]-R2-PO4H3 || [M-H]-R3-PO4H3

Structural analysis of lipid A from Escherichia coli O157:H7:K- using thin-layer chromatography and ion-trap mass spectrometry;
LipidBlast example: hybrid quadrupole ion mobility spectrometry time-of-flight

Name: PC 32:0; [M+Na]+; GPCho(16:0/16:0)
MW: 756 1D#: 42167 DB: lipidblast-pos
Comment: Parent=756.55190 Mz_exact=756.55190; PC 32:0; [M+Na]+; GPCho(16:0/16:0); C40H80NO8P
5 m/z Values and Intensities:
- 697.47840 999.00  [M+Na]-C3H9N (-59)
- 573.48586 600.00  [M+Na]-C5H14NO4P (-183)
- 518.32238 20.00  [M+Na]-sn1  || [M+Na]-sn2
- 500.31182 20.00  [M+Na]-sn1-H2O  || [M+Na]-sn2-H2O
- 441.23832 40.00  [M+Na]-59-sn1  || [M+Na]-59-sn2

Source: Direct Tissue Imaging and Characterization of Phospholipids Using MALDI SYNAPT HDMS System; Waters 2008; 72002444en
Emmanuelle Claude, Marten Snel, Therese McKenna, James Langridge;
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

⇒ Validation, Validation, Validation.
    Proof must be given that algorithm works for large diverse sets of molecules (n=100..100,000)
Literature (236 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]
GlySpy and the Oligosaccharide Subtree Constraint Algorithm (OSCAR)
Mass Frontier for further discussion MOLGEN-MS [LINK]
http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Structure_Elucidation/
http://fiehnlab.ucdavis.edu/projects/LipidBlast
MetIDB: A Publicly Accessible Database of Predicted and Experimental 1H NMR Spectra of Flavonoids [LINK]
METFRAG: In silico fragmentation for computer assisted identification of metabolite mass spectra [LINK]
Advances in structure elucidation of small molecules using mass spectrometry [Link]
Computational mass spectrometry for small molecules [Link]