Generation of in-silico MS/MS mass spectra using combinatorial algorithms and reaction prediction expert systems

239th ACS National Meeting 2010
San Francisco, CA

CINF: Division of Chemical Information
Metabolomics: A Field at the Boundaries between Chemistry and Biology

Tobias Kind, Kwang-Hyeon Liu, Do Yup Lee, Oliver Fiehn
FiehnLab – Metabolomics
UC Davis Genome Center, Davis, USA
Outline

1) History and motivation (NIH glue grant of 70 Mio. Dollars)

2) Molecule creation using combinatorial algorithms

3) Modeling of in-silico MS/MS spectra

4) Outlook and Conclusions
Tandem mass spectrometry

sn1 = alkyl or acyl rest

sn2 = alkyl or acyl rest

Precursor ion
m/z=760.64

Product ions of
m/z=760.64

Mass spectrum

MS/MS spectrum
In-silico mass spectra:
• **m/z fragments** and **abundance** calculation required
• **statistical** (computer derived) and **heuristic rules** (experience of a human expert)
Idea: Consistent lipid fragmentation (CID 35 V)

Phosphatidylcholine - PC (16:0/16:1) or short PC 32:1
[M+H]+ MS/MS precursor m/z = 732.55

\[ \text{sn1 acyl/alkyl chain} \]

\[ \text{sn2 acyl/alkyl chain} \]

\[ \text{headgroup} \]

Abundance

m/z

[\text{[M+H]-C3H9N (-59)}] 673.3751991201706

[\text{[M+H]-H2O (-18)}] 713.4246597050723

[\text{[M+H]-sn2-H2O}] 496.3987288063696

[\text{[M+H]-sn1}] 549.40657225188

[\text{[M+H]-C5H14NO4P (-183)}] 613.2630013581689

[\text{[M+H]-sn1-H2O}] 393.40846996748957

[\text{[M+H]-sn2}] 496.3987288063696
Existing in-silico approaches for tandem mass spectrometry modeling

1) Peptides (Proteomics) – o.k.

2) Oligosaccharides (Glycomics) – o.k.

3) Not for small molecules – or not validated on larger sample sets (*)

In-silico spectra only "easy" to generate when consisting and repeating building blocks exist. For example amino acids in peptides or sugar building blocks in oligosaccharides.

(*) Matching Structures to Mass Spectra Using Fragmentation Patterns: Are the Results As Good As They Look?
A total of 29,027 MS/MS publications exist (22,991 excluding peptides)
What went historically wrong?

Challange: Name that graph! (*)

(*) Internet meme from Chemical blogspace http://cb.openmolecules.net/
Promise: You're not gonna get rickrolled.
What went historically wrong?

The largest commercial MS/MS database (NIST08) contains 14,802 MS/MS spectra of 2857 unique compounds (85 lipids).

The largest public source (Massbank) contains 8,337 MS/MS spectra of 2572 unique compounds.

8 Million commercial unique chemicals available (eMolecules)
50 million molecules in CSLS DB
What went historically wrong?

A) Scientists (we) do not publish machine readable MS/MS spectra
B) Scientists (we) publish MS/MS as bitmap picture in PDF
C) Scientists (we) do not share spectra (Open Access, commercially)
D) There are no easy to use technologies in place to enable data sharing

Do we need to push OCR technology?
Enable electronic data (MS spectra) sharing!

Digital structures and spectra

Analog paper publication

Digital database from OCR data

Data reduction and loss
remove noise and uninteresting data

Extreme data loss
OCR and text mining conversion errors

Hamburger to Cow algorithm or "Wishful Thinking"
Requires Jurassic Park Technology

Kind T, Scholz M, Fiehn O
Eureka! Create in-silico MS/MS spectra

Experimental MS/MS precursor m/z = 732.55

in-silico MS/MS match precursor m/z = 732.55

Head to Tail MF=437 RMF=666

PC 32:1; [M+H]+, GPCho(16:0/ 16:1(7Z))
Combinatorial library algorithms for structure generation

1) **LipidMaps** Tools (Perl)  
   based on open source MayaChemTools by Manish Sud

2) **SMILIB** (JAVA)  
   open source Modlab Uni Frankfurt Schüller/Hähnke/Schneider

3) **Reactor** (JAVA)  
   virtual reaction processing tool by ChemAxon

A) **Instant-JChem database** (ChemAxon)  
   for structural handling

B) **MassFrontier** (HighChem/Thermo)  
   for mass spectrometry based reactions and fragmentations
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

choline
inositol
glycerol

+ 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
Instant JChem structure handling

Lipid database of 44,000 glycerophospholipids, 444,080 diacylglycerols, and mostly triacylglycerols from LipidMaps

Export of structures from Instant-JChem into EXCEL

Structures created with LipidMaps tools
MS/MS search with NIST MS search program using precursor search and dot-product match

Search speed ~ 100 MS/MS spectra per second (without GUI)
### Library size and coverage of lipid classes

<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>plasményl-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>plasményl-PE</td>
<td>222</td>
<td>666</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Phosphatidyldserines</td>
<td>PS</td>
<td>5123</td>
<td>15369</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>Sphingomyelines</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>Phosphatidylinositol</td>
<td>PI</td>
<td>5476</td>
<td>5476</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Phosphatidyglycerols</td>
<td>PG</td>
<td>5476</td>
<td>5476</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>Cardiolipins</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>Diacylglycerols</td>
<td>DAG</td>
<td>1764</td>
<td>1764</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>Triacylglycerols</td>
<td>TAG</td>
<td>264</td>
<td>5280</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>Monogalactosyldiacylglycerols</td>
<td>MGDG</td>
<td>5476</td>
<td>21904</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>Digalactosyldiacylglycerols</td>
<td>DGDG</td>
<td>5476</td>
<td>10952</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>Sulfogluconosyldiacylglycerols</td>
<td>SQDG</td>
<td>5476</td>
<td>5476</td>
<td>1</td>
</tr>
<tr>
<td>19</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>95326</strong></td>
<td><strong>184164</strong></td>
<td><strong>40</strong></td>
</tr>
</tbody>
</table>

**Covered adduct libraries**

[M+H]+  [M+Na]+  [M+NH4]+  [M-H]-
[M-2H](2-)+  [M+NH4-CO]+  [M+Na2-H]+  [M]+  [M-H+Na]+
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:
723.49409 999.00  [M+Na]-C3H9N (-59)
599.50155 600.00  [M+Na]-C5H14NO4P (-183)
544.33807 20.00  [M+Na]-sn1
526.32751 20.00  [M+Na]-sn1-H2O
518.32243 20.00  [M+Na]-sn2
500.31187 20.00  [M+Na]-sn2-H2O
467.25401 40.00  [M+Na]-59-sn1
441.23837 40.00  [M+Na]-59-sn2

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

LC/MS Analysis of Bronchoalveolar Lavage Fluid Phospholipids as Biomarkers for Chronic Lung Inflammation;
Agilent application note; 5989-1491EN; Barroso, Bischoff
Example: Electrospray-ion trap mass spectrometer

MW: 795 ID#: 75218 DB: lipidblast-neg
Comment: Parent=795.50478 Mz_exact=795.50478 ; MGDG 38:9; [M-H];
2 largest peaks:

2 m/z Values and Intensities:
301.21662 999.00 | 275.20098 999.00 |

275.5
301.5
447.4
537.5
795.5

MGDG(20:5/18:4)
(4 candidates in database)
(512 double bond isomers)
Example: Hybrid Ion-Trap (IT) and Time-of-Flight (TOF)

Source: A Chloroplastic UDP-Glucose Pyrophosphorylase from Arabidopsis Is the Committed Enzyme for the First Step of Sulfolipid Biosynthesis
Y Okazaki, M Shimojima, Y Sawada et al. The Plant Cell 21:892-909 (2009);
Example: ion trap mass spectrometer

**Name:** LipidA PP [14/14/14/14/3O-(12)/3O-(14)]; [M-H]-

**MW:** 1796

**ID#:** 64304

**DB:** lipidblast-neg

**Comment:** Parent=1796.21157 Mz_exact=1796.21157 ; LipidA PP [14/14/14/14/3O-(12)/3O-(14)]; [M-H]-; C94H178N2O25P2; LipidA-PP-[R2(14:0)(3-OH)/R3(14:0)(3-OH)/R2'(14:0)/R3('14:0)/R2'-3-O-(12:0)/R3'-3O-(14:0)]

**9 largest peaks:**

<table>
<thead>
<tr>
<th>m/z</th>
<th>Intensity</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>1552.00785</td>
<td>999.00</td>
<td>[M-H]-</td>
</tr>
<tr>
<td>1796.21157</td>
<td>500.00</td>
<td>[M-H]-PO3H</td>
</tr>
<tr>
<td>1470.02587</td>
<td>300.00</td>
<td>[M-H]-PO4H3</td>
</tr>
<tr>
<td>1568.00277</td>
<td>250.00</td>
<td>[M-H]-PO4H3-R2'-O-FA</td>
</tr>
<tr>
<td>1552.00785</td>
<td>999.00</td>
<td>[M-H]-R2 acyl FA</td>
</tr>
<tr>
<td>1498.05715</td>
<td>300.00</td>
<td>[M-H]-PO4H3-R2'-O-FA</td>
</tr>
<tr>
<td>1470.02587</td>
<td>300.00</td>
<td>[M-H]-PO4H3-R3'-O-FA</td>
</tr>
<tr>
<td>1454.03095</td>
<td>250.00</td>
<td>[M-H]-R2-PO4H3</td>
</tr>
</tbody>
</table>

**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

Example: hybrid quadrupole ion mobility spectrometry time-of-flight

Name: PC 32:0; [M+Na]+; GPCho(16:0/16:0)
MW: 756
ID#: 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; 720002444en
Emmanuelle Claude, Marten Snel, Therese McKenna, James Langridge;
Library curation costs money

This library will be:
CC-BY This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation.

Source: http://creativecommons.org/licenses/by/3.0/
Applications and future developments

A1) Energy research – lipid profiling

A2) Health research – lipidomics

A3) Fundamental research – understanding spatial and temporal distribution of lipids in plants and animals

S1) Side effect: Lipidomics for the masses (use low-cost ion traps)

F1) Oxylipids and different oxygenated species for medical and age research require sensitive triple-quadrupole MS (QTRAP) or hybrids

F2) Rare lipid species from health related species (tuberculosis, pestilence)

F3) Regiospecific databases (from MS$^3$ and MS$^4$ data)

F4) Translation to other molecule classes (requires diverse validation sets)
Thank you!

Fiehn Lab

**Dr. Oliver Fiehn (Principal Investigator)**
Mine Palazoglu (Library, GC-MS, GCT)

**Dr. Tobias Kind (Cheminformatics)**
Dinesh Kumar Barupal (Bioinformatics)
Gert Wohlgemuth (BinBase)
Kirsten Skogerson (NMR, GCxGC)

**Dr. Kwang-Hyeon Liu (LC, Pharma)**
Sangeeta Kumari (GCT, GC-MS)
Sevini Shahbaz (Library)
Kristie Cloos (Lipids, MS, GC-MS)

**Dr. Pierre Ayotte (Docking)**
John Meissen (UPLC, LC)

**Dr. Do Yup Lee (now LBNL Berkley)**

Sponsors Fiehn Lab

NIH R01 ES013932
NIH GM078233 & ARRA RC2
NIH R01 DK078328
NIH 1 R21 AI073323-01A1
UC Discovery itl07-10167
NSF MCB 0520140
EU FP7 Health-2007-2.1.4.1/Dupont
Agilent, LECO, Waters

Thanks to the useful LipidMaps service!
Please apply for beta-testing!
Tandem mass spectrometry (MS/MS)

Iontrap MS/MS spectra creation

Iontrap MS/MS spectra creation

sn1 = alkyl or acyl rest

sn2 = alkyl or acyl rest