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

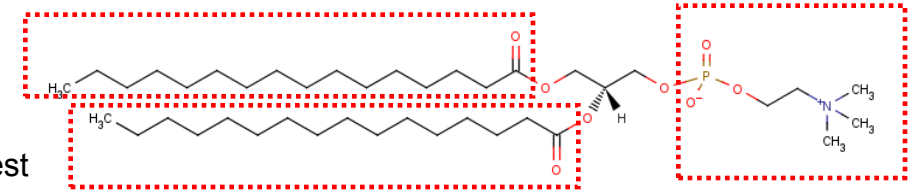
3) Modeling of in-silico MS/MS spectra

4) Outlook and Conclusions

Tandem mass spectrometry

sn1 = alkyl or acyl rest

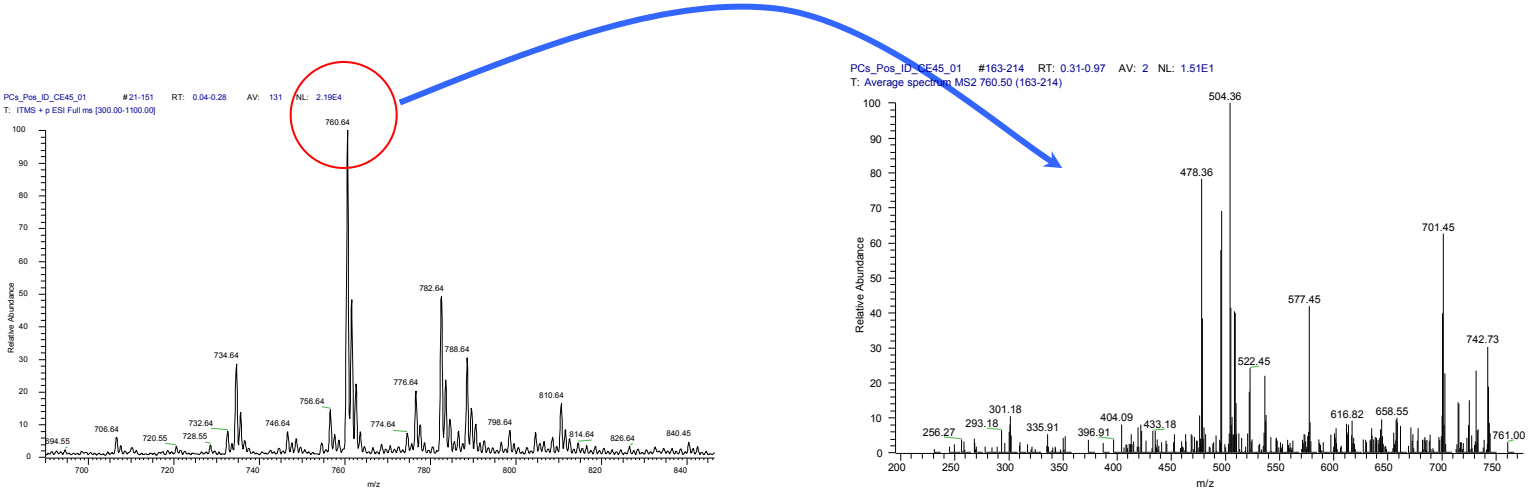
sn2 = alkyl or acyl rest



head group

Precursor ion m/z=760.64

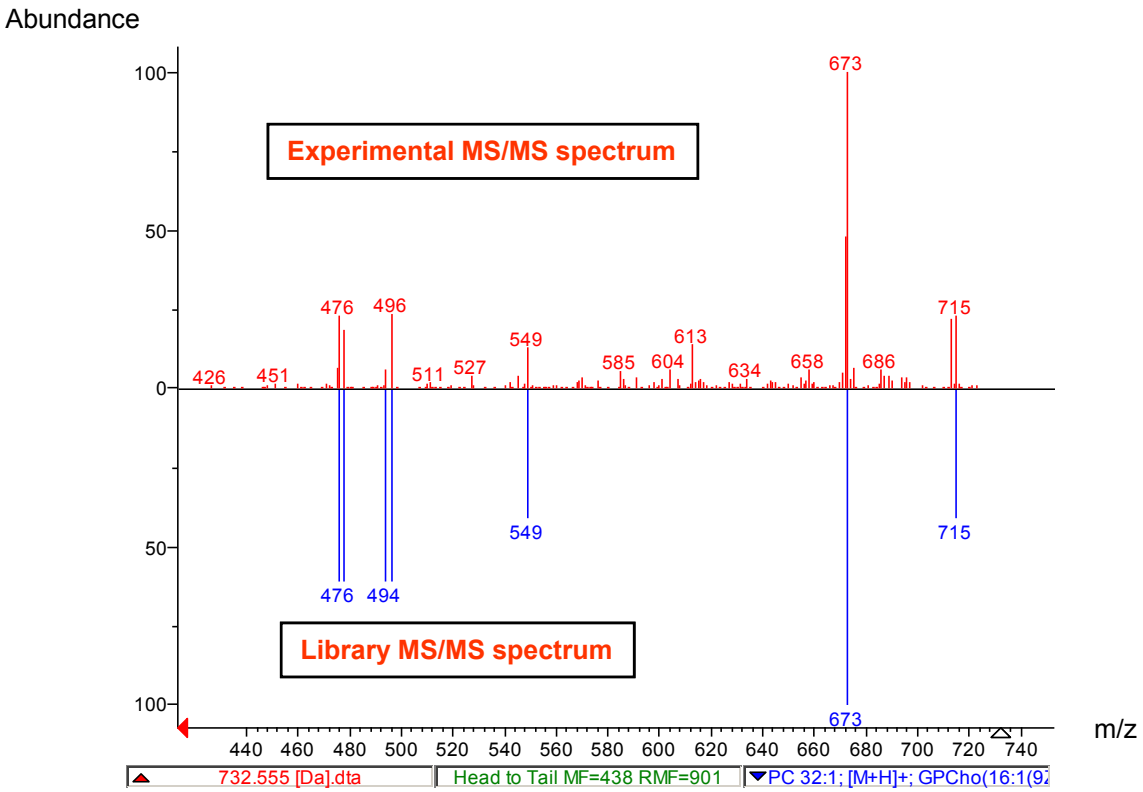
Product ions of m/z=760.64



Mass spectrum

MS/MS spectrum

MS/MS mass spectral library search

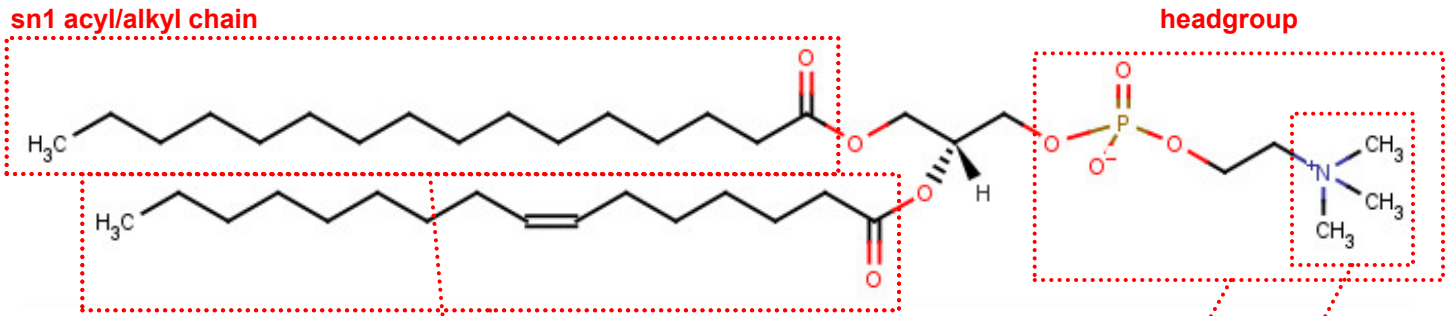


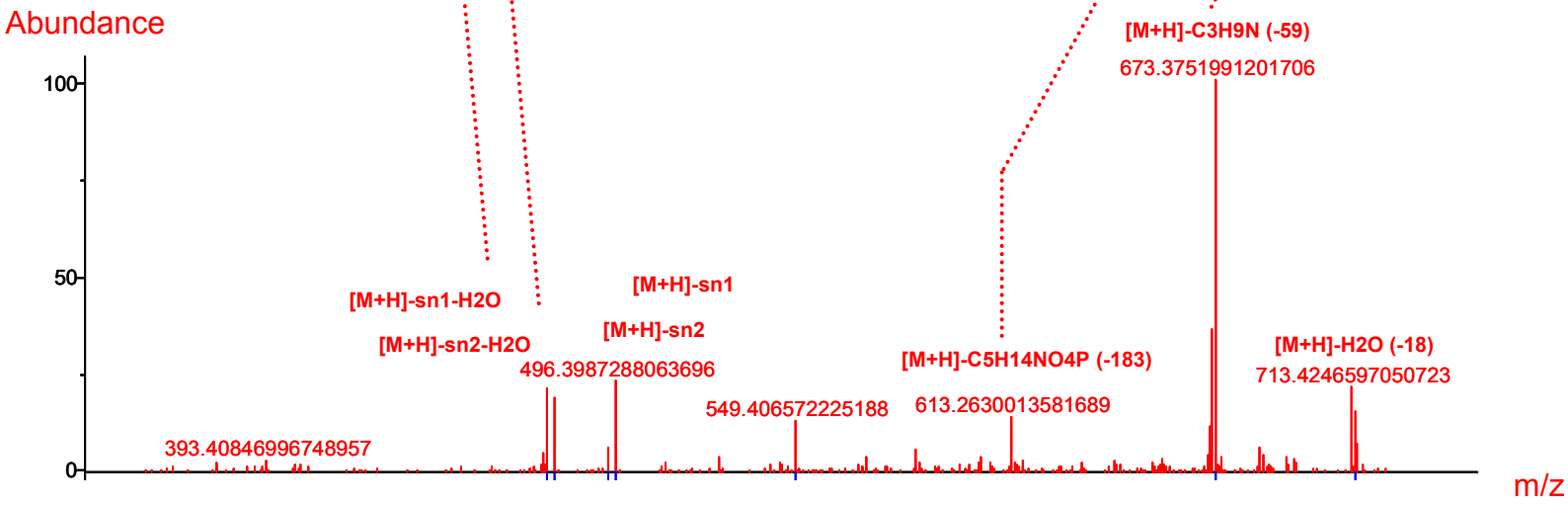
In-silico mass spectra:

- **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





Existing in-silico approaches for tandem mass spectrometry modeling

1) **Peptides (Proteomics) – o.k.**

2) **Oligosaccharides (Glycomics) – o.k.**

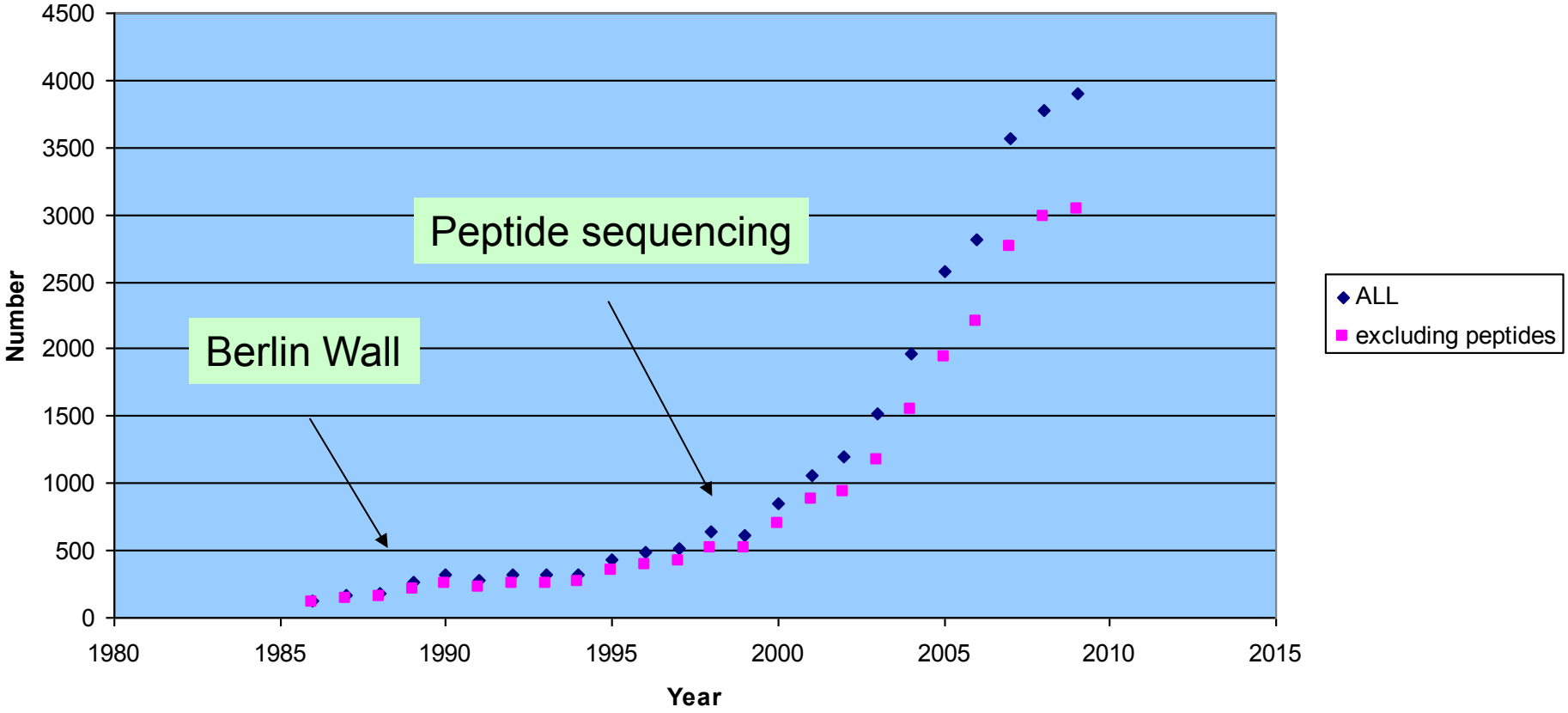
3) **Not for small molecules – or not validated on larger sample sets (*)**

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?**

Schymanski et al. *Anal. Chem.*, 2009, 81 (9), pp 3608–3617

Number of publications about tandem mass spectrometry (MS/MS)

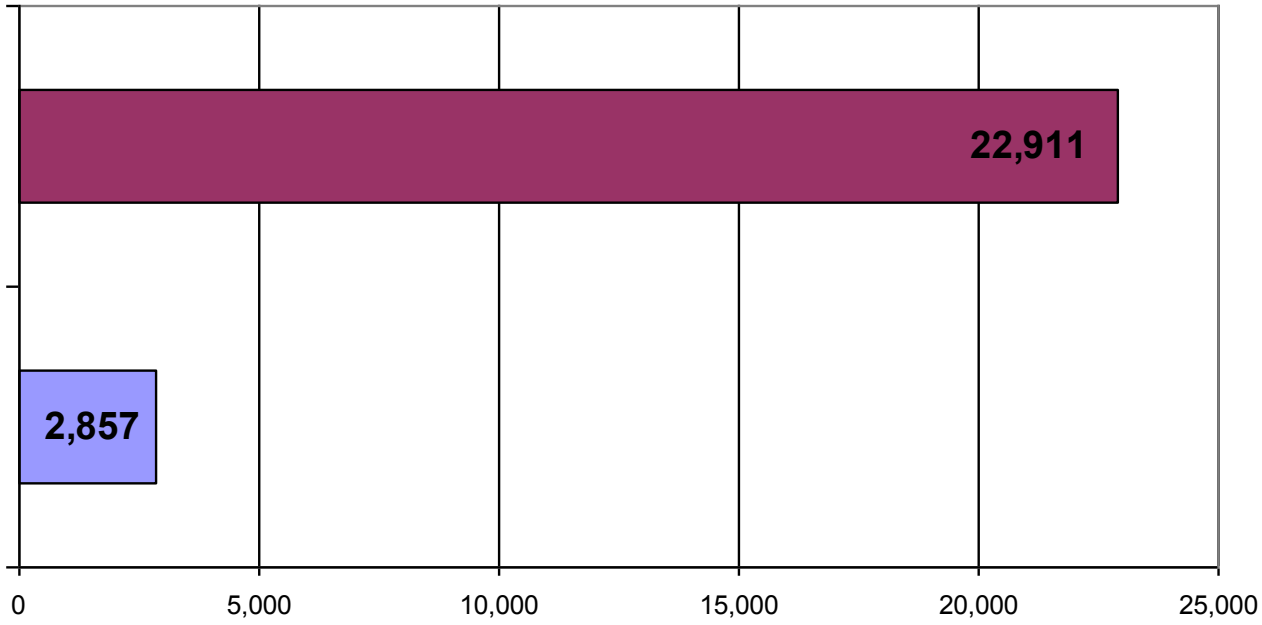


A total of 29,027 MS/MS publications exist (22,991 excluding peptides)

Source: CAS DB

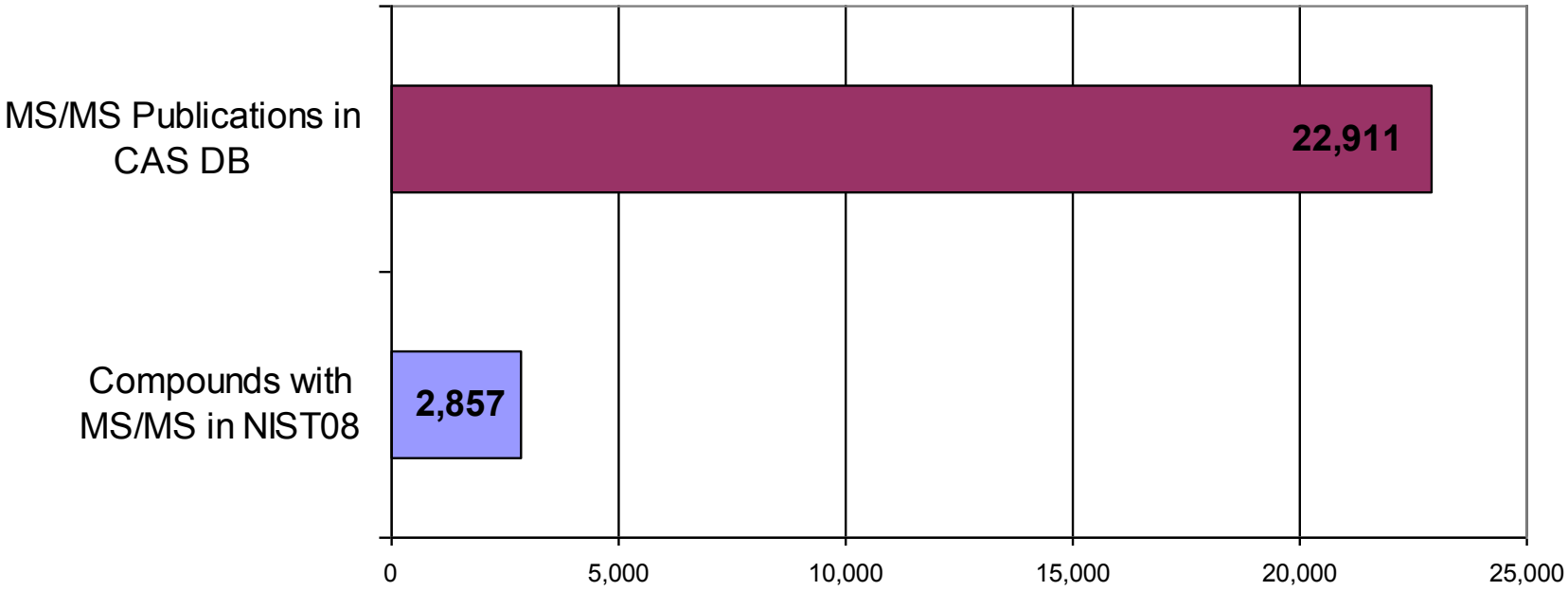
What went historically wrong?

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

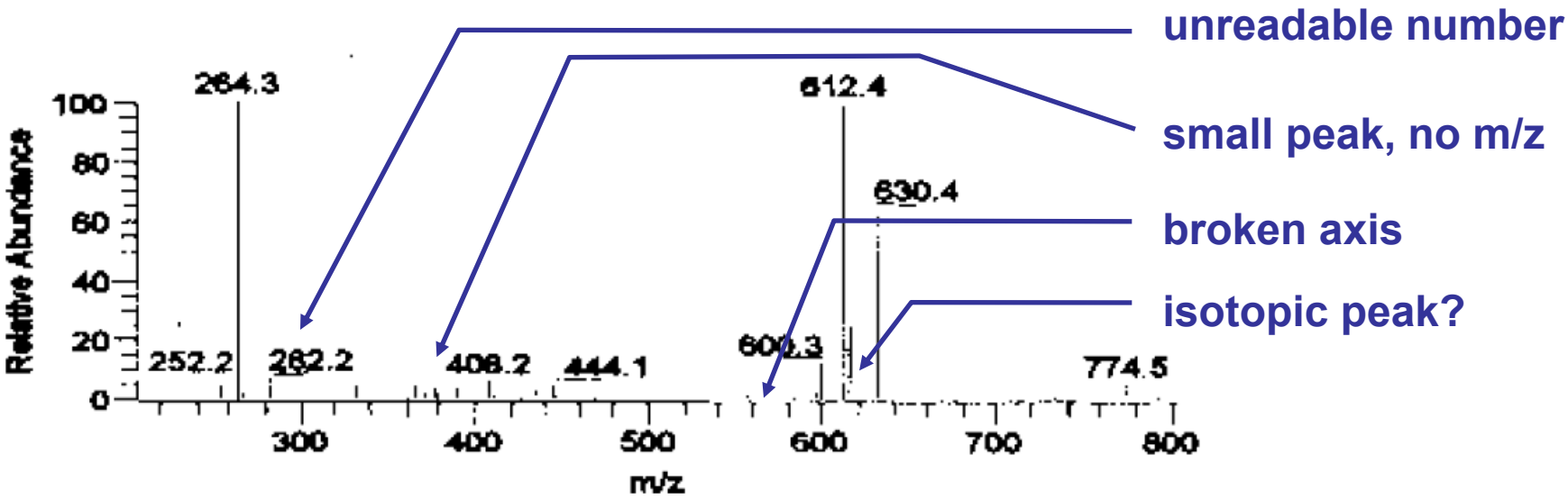
8,337 MS/MS spectra of 2572 unique compounds

8 Million commercial unique chemicals available (eMolecules)
50 million molecules in CSLS DB

What went historically wrong?

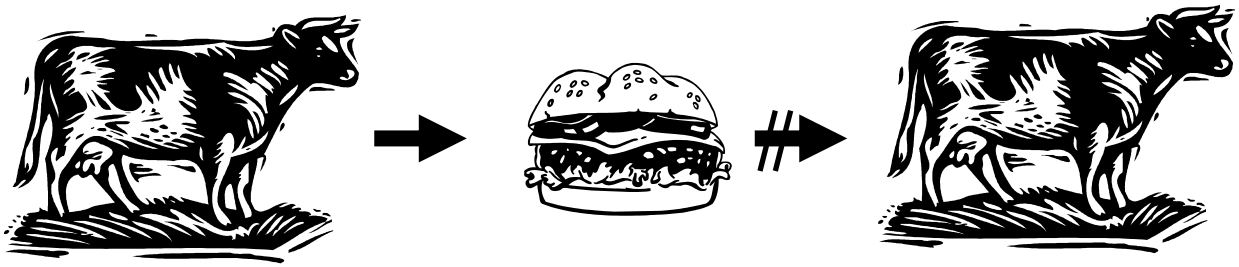
A) Scientists (we) do not publish machine readable MS/MS spectra
B) Scientists (we) publish MS/MS as bitmap picture in PDF
C) Scientists (we) do not share spectra (Open Access, commercially)
D) There are no easy to use technologies in place to enable data sharing

Do we need to push OCR technology?



To cite this Article: Byrdwell, William Craig (2003) 'Dual Parallel Liquid Chromatography/Dual Mass Spectrometry (LC2/MS2) of Bovine Brain Total Lipid Extract', Journal of Liquid Chromatography & Related Technologies, 26:19, 3147 -

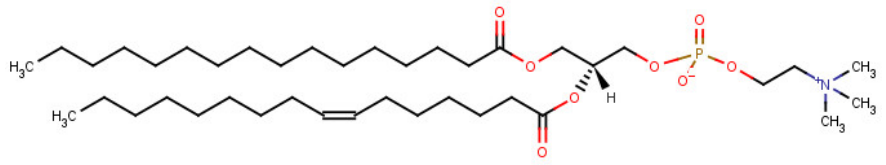
Enable electronic data (MS spectra) sharing!

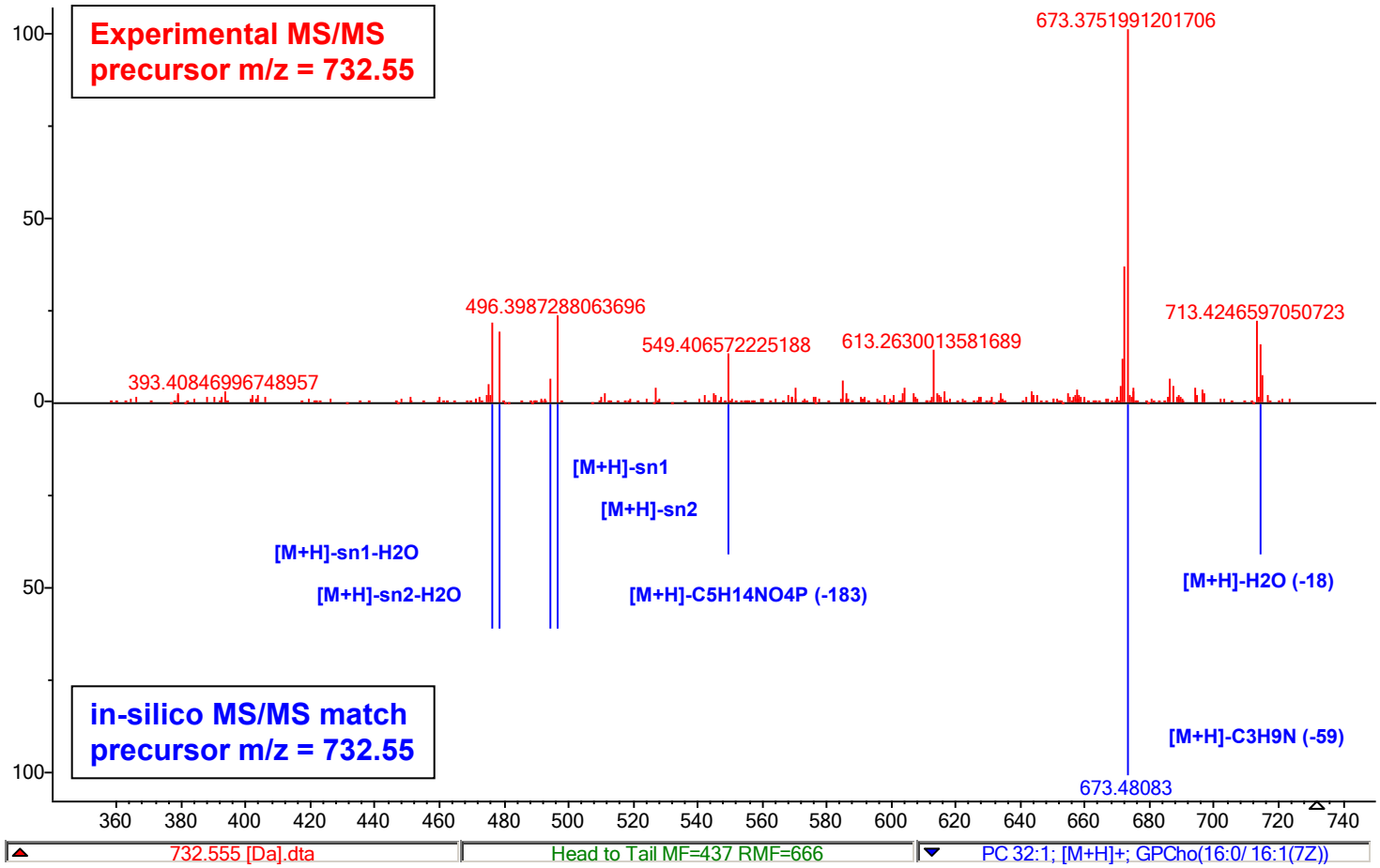


Hamburger to Cow algorithm or "Wishful Thinking" Requires Jurassic Park Technology

Kind T, Scholz M, Fiehn O
How Large Is the Metabolome? A Critical Analysis of Data Exchange Practices in Chemistry.
PLoS ONE 4(5): e5440. (2009); doi:10.1371/journal.pone.0005440

Eureka! Create in-silico MS/MS spectra





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

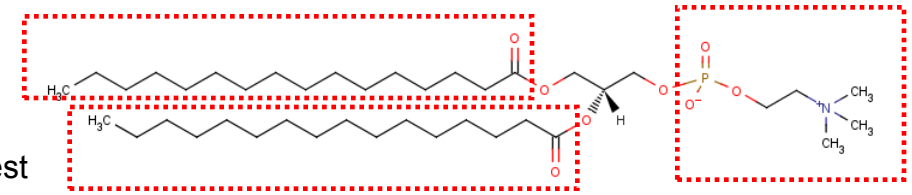
A) Instant-JChem database (ChemAxon)
for structural handling

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

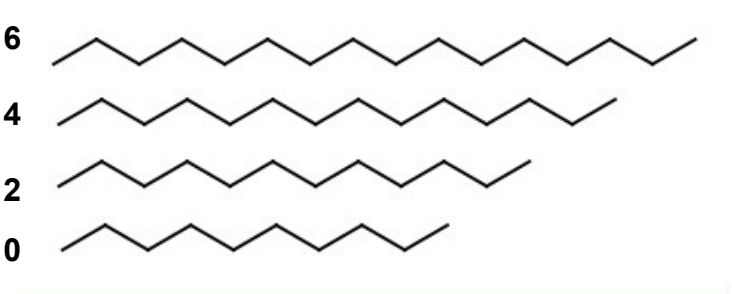
Functional group (variable) **Linker** **Scaffold (conserved)**

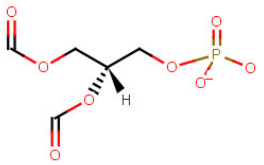
C16

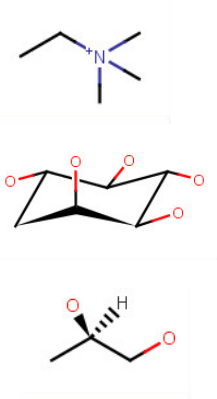
C14

C12

C10







choline

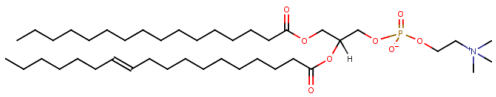
inositol

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

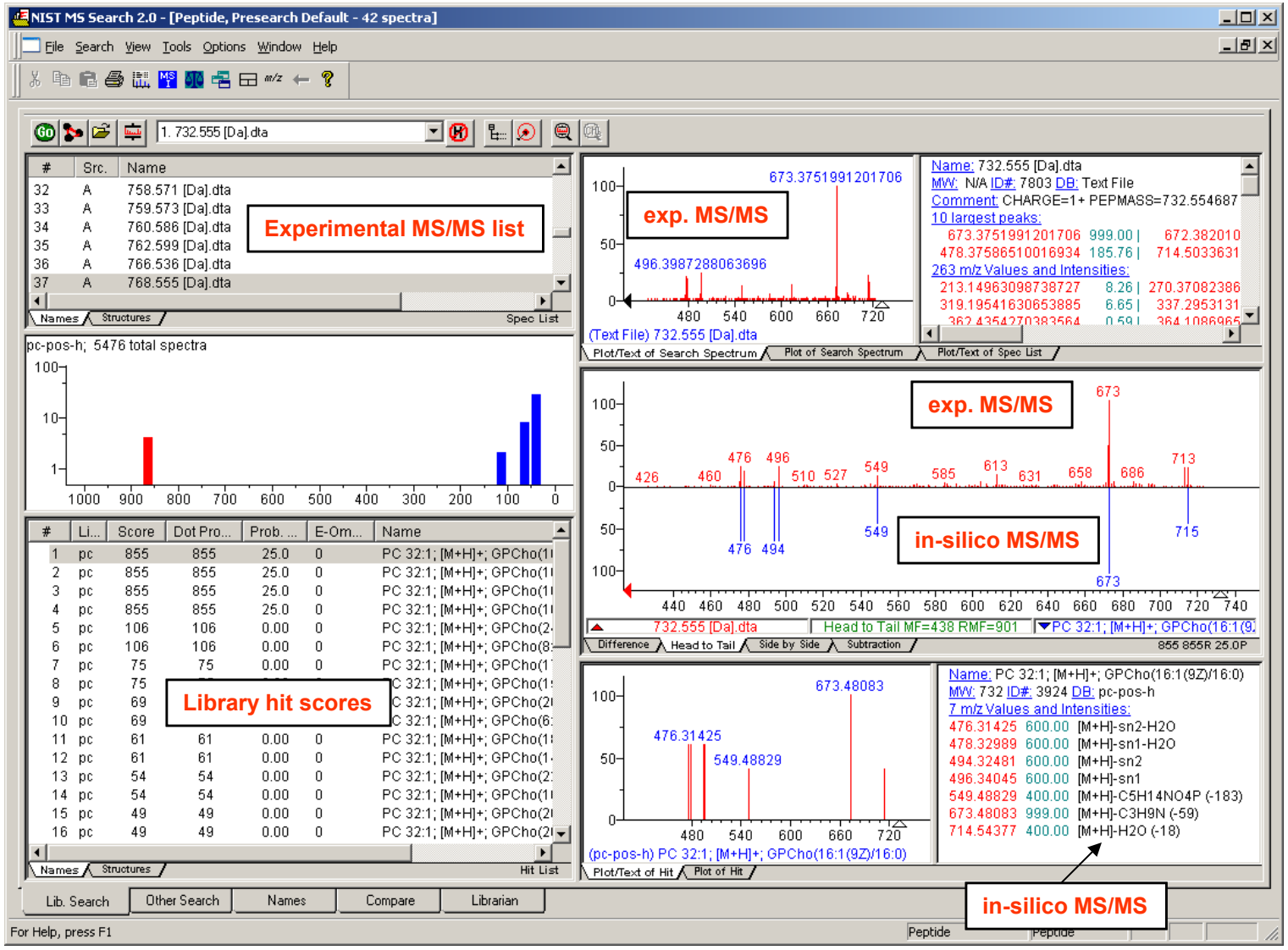
Lipid database of
44,000 glycerophospholipids
444,080 diacylglycerols.
and mostly triacylglycerols
from **LipidMaps**



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)

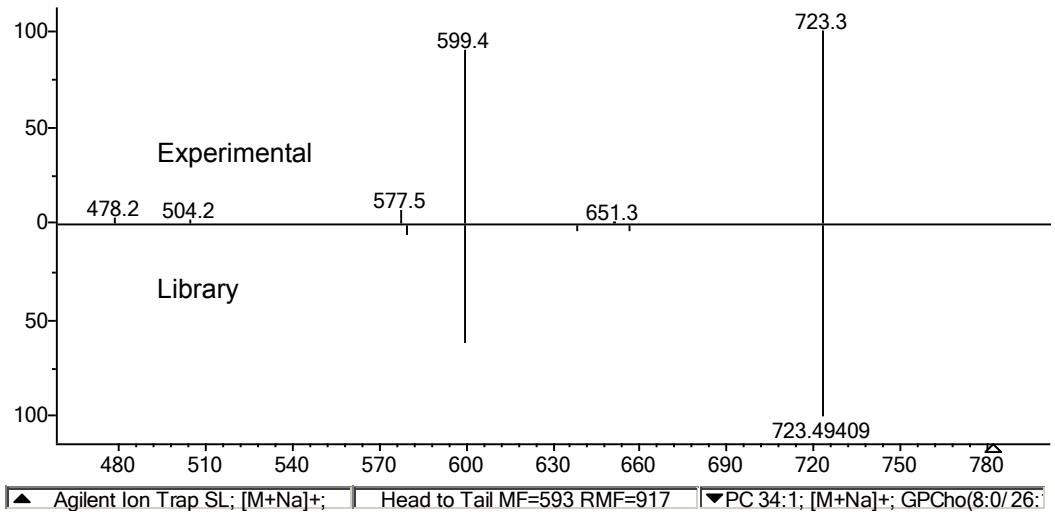
Library size and coverage of lipid classes

Number	LipidClass	Short	Number compounds	Number MS/MS spectra with different adducts	Number MS/MS LIBS
1	Phosphatidylcholines	PC	5476	10952	2
2	Lysophosphatidylcholines	lysoPC	80	160	2
3	Plasmenylphosphatidylcholines	plasmenyl-PC	222	444	2
4	Phosphatidylethanolamines	PE	5476	16428	3
5	Lysophosphatidylethanolamines	lysoPE	80	240	3
6	Plasmenylphosphatidylethanolamines	plasmenyl-PE	222	666	3
7	Phosphatidylserines	PS	5123	15369	3
8	Sphingomyelines	SM	168	336	2
9	Phosphatidic acids	PA	5476	16428	3
10	Phosphatidylinositols	PI	5476	5476	1
11	Phosphatidylglycerols	PG	5476	5476	1
12	Cardiolipins	CL	25426	50852	2
13	Ceramide-1-phosphates	CerP	168	336	2
14	Diacylglycerols	DAG	1764	1764	1
15	Triacylglycerols	TAG	2640	5280	2
16	Monogalactosyldiacylglycerols	MGDG	5476	21904	4
17	Digalactosyldiacylglycerols	DGDG	5476	10952	2
18	Sulfoquinovosyldiacylglycerols	SQDG	5476	5476	1
19	Diphosphorylated hexaacyl Lipid A	LipidA-PP	15625	15625	1
Total	**All libraries**		**95326**	**184164**	**40**

Covered adduct libraries

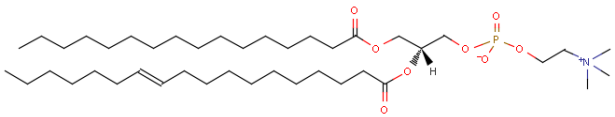
[M+H]+ [M+Na]+ [M+NH4]+ [M-H]-
[M-2H](2-) [M+NH4-CO]+ [M+Na2-H]+ [M]+ [M-H+Na]+

Example: ion trap mass spectrometer



Source: Agilent.com

Agilent Ion Trap SL/XCT



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

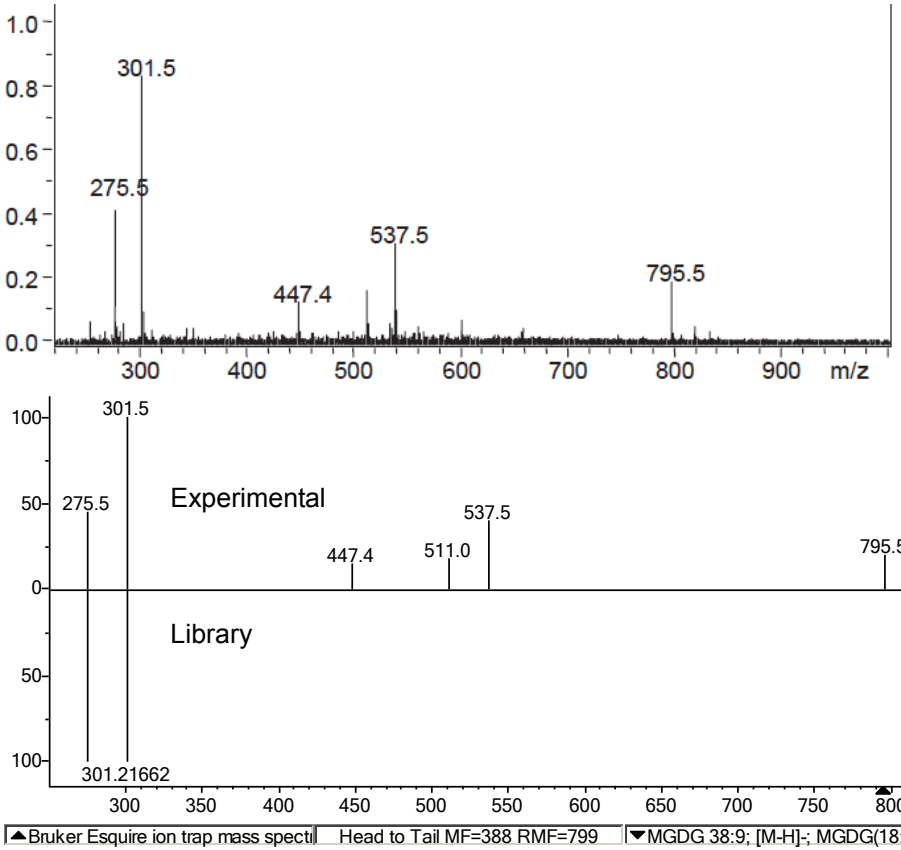
1st Hit group

PC 34:1
(42 candidates)

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

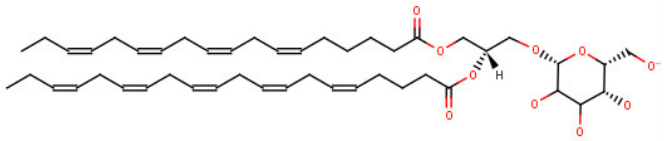
LC/MS Analysis of Bronchoalveolar Lavage Fluid Phospholipids as Biomarkers for Chronic Lung Inflammation; Agilent application note; 5989-1491EN; Barroso, Bischoff

Example: Electrospray-ion trap mass spectrometer



Source: www.bdal.com

Bruker Esquire Ion Trap



1st Hit

MGDG(20:5/18:4)
(4 candidates in database)
(512 double bond isomers)

Name: MGDG 38:9; [M-H]-; MGDG(18:4(6Z,9Z,12Z,15Z)/20:5(5Z,8Z,11Z,14Z,17Z))
MW: 795 ID#: 75218 DB: lipidblast-neg
Comment: Parent=795.50478 Mz_exact=795.50478 ; MGDG 38:9; [M-H]-;
MGDG(18:4(6Z,9Z,12Z,15Z)/20:5(5Z,8Z,11Z,14Z,17Z)); C47H72O10
2 largest peaks:

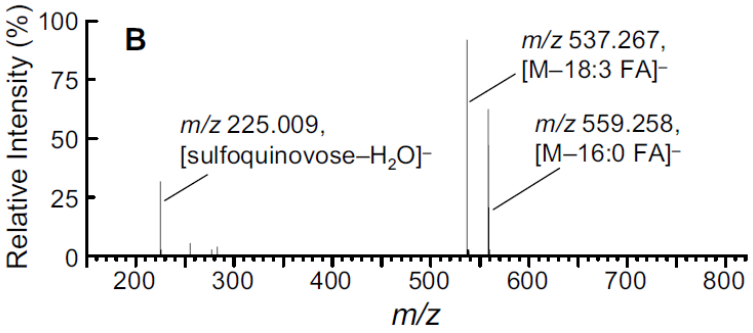
	301.21662	999.00 |	275.20098	999.00 |

2 m/z Values and Intensities:

301.21662	999.00	sn2 FA
275.20098	999.00	sn1 FA

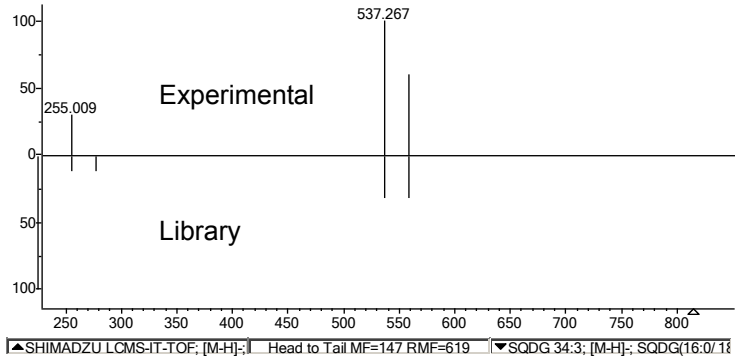
Source: Chimica e biologia a confronto: pigmenti e altri metaboliti secondari prodotti da dinoflagellati del Lago di Tovel; Studi Trent. Sci. Nat., Acta Biol., 81 (2004), Suppl. 2: 413-426;Rita FRASSANITO, Ines MANCINI & Graziano GUELLA

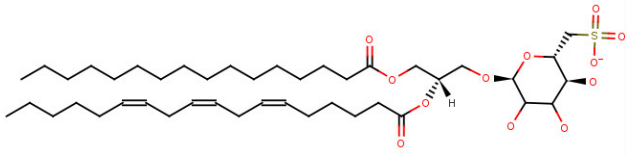
Example: Hybrid Ion-Trap (IT) and Time-of-Flight (TOF)



Source: schimadzu.com

Shimadzu's LCMS-IT-TOF





1st Hit

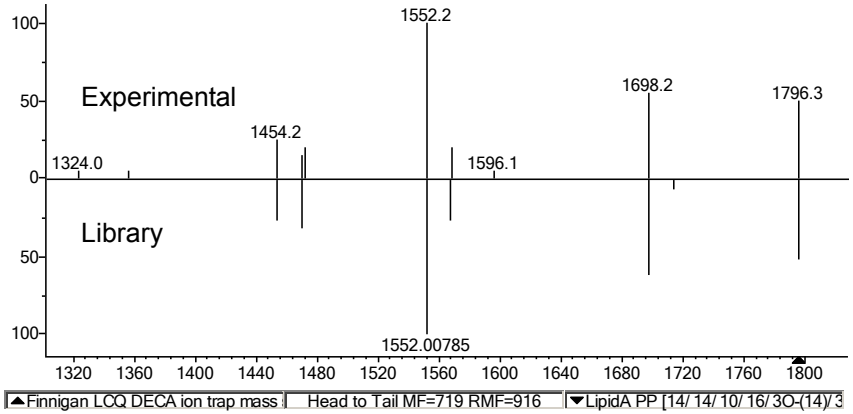
SQDG 34:3
(8 candidates)

Name: SQDG 34:3; [M-H]-; SQDG(16:0/18:3(6Z,9Z,12Z))
MW: 815 ID#: 106150 DB: lipidblast-neg
Comment: Parent=815.49792 Mz_exact=815.49792 ; SQDG 34:3; [M-H]-; SQDG(16:0/18:3(6Z,9Z,12Z)); C43H76O12S

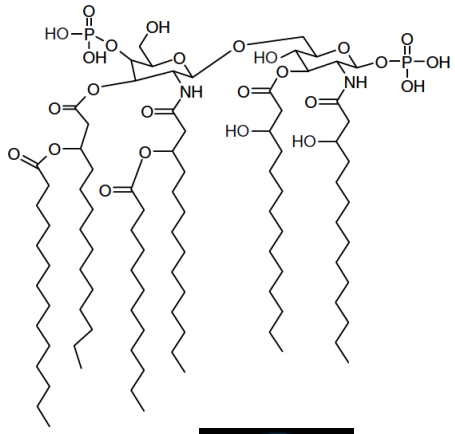
559.25784	300.00	[M-H]-sn1
537.27348	300.00	[M-H]-sn2
277.21662	100.00	sn2 FA
255.23226	100.00	sn1 FA
225.00690	999.00	fragment C6H9O7S

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



Thermo Finnigan LCQ/LTQ



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:

1552.00785	999.00 |	1698.23467	600.00 |
1796.21157	500.00 |	1498.05715	300.00 |
1470.02587	300.00 |		
1596.03405	250.00 |	1568.00277	250.00 |
1454.03095	250.00 |	1714.22959	50.00 |

9 m/z Values and Intensities:

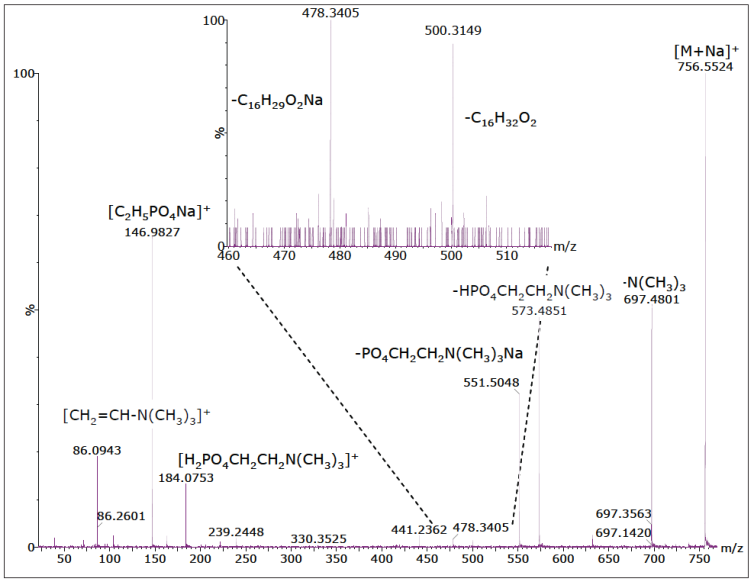
1796.21157	500.00	[M-H]-
1714.22959	50.00	[M-H]-PO3H
1698.23467	600.00	[M-H]-PO4H3
1596.03405	250.00	[M-H]-PO4H3-R2'-O-FA
1568.00277	250.00	[M-H]-PO4H3-R3'-O-FA
1552.00785	999.00	[M-H]-R2 acyl FA || [M-H]-R3 acyl FA
1498.05715	300.00	[M-H]-PO4H3-R2'-O-FA
1470.02587	300.00	[M-H]-PO4H3-R3'-O-FA
1454.03095	250.00	[M-H]-R2-PO4H3 || [M-H]-R3-PO4H3

2nd Hit

Lipid A (PP)
(16 candidates)

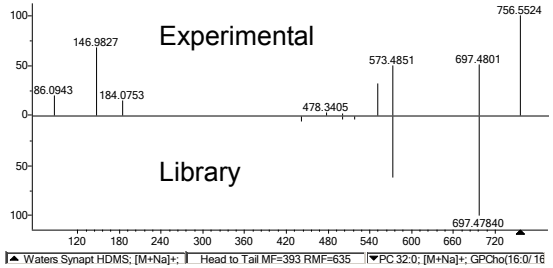
Structural analysis of lipid A from Escherichia coli O157:H7:K- using thin-layer chromatography and ion-trap mass spectrometry; Chang-Soo Lee, Yun-Gon Kim, Hwang-Soo Joo, Byung-Gee Kim; J Mass Spectrom. 2004 May;39(5):514-25.

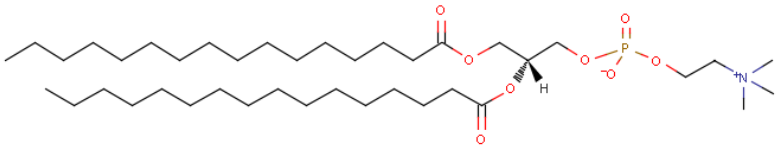
Example: hybrid quadrupole ion mobility spectrometry time-of-flight



Source: Waters.com

Waters HDMS Synapt





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

1st Hit

PC 32:0

Source: Direct Tissue Imaging and Characterization of Phospholipids Using MALDI SYNAPT HDMS System; Waters 2008; 720002444en
Emmanuelle Claude, Marten Snel, Therese McKenna, James Langridge;

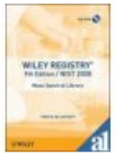
Library curation costs money

Wiley Registry of Mass Spectral Data, With Nist 2008
Ensure your lab has the most comprehensive **library**. **Wiley** combined the two ... 2008 **library**, including a new release of NIST **MS** Search software with AMDIS. ....
Add to Shopping List

$11,143.83 new
Alibris

Wiley Registry of Mass Spectral Data, 8th Edition
Ensure your lab has the most comprehensive **library**. **Wiley** combined the two ... 2008 **library**, including a new release of NIST **MS** Search software with AMDIS. ....
Add to Shopping List

$10,726.69 new
Alibris



Wiley Registry of Mass Spectral Data, with NIST 2008
... **Wiley** Registry/NIST **library** and the complete NIST/EPA/NIH **library** and ... 2008 **library**, including a new release of NIST **MS** Search software with AMDIS. ....
Add to Shopping List

$9,641.29 new
A1Books
★★★★☆ 6,669 seller ratings

Wiley Registry of Mass Spectral Data, 9th Ed. with NIST 2008 (U
Ensure Your Lab Has the Most Comprehensive **Library Wiley** combined the two ... 2008 **library**, including a new release of NIST **MS** Search software with AMDIS. ....
Add to Shopping List

$8,095.00 new
Free shipping
Amazon.com
★★★½☆ 4,159 seller ratings

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

Source: Steve Jurvetson FLICKR

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

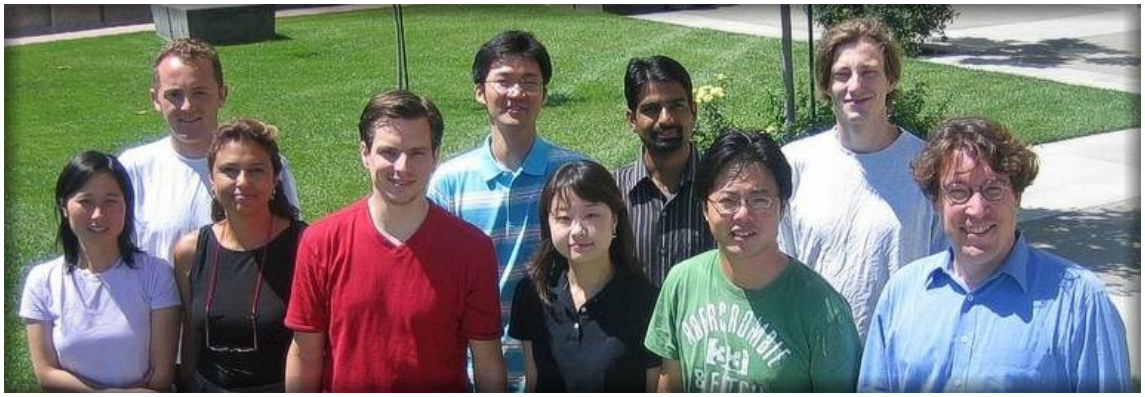
F1) Oxylipids and different oxygenated species for medical and age research require sensitive triple-quadrupole MS (QTRAP) or hybrids

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

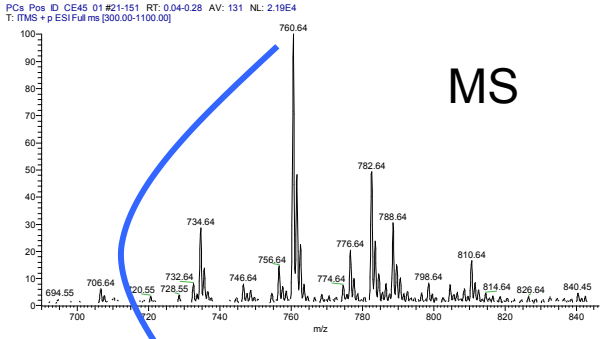
Dr. Oliver Fiehn (Principal Investigator)
Mine Palazoglu (Library, GC-MS, GCT)
Dr. Tobias Kind (Cheminformatics)
Dinesh Kumar Barupal (Bioinformatics)
Gert Wohlgemuth (BinBase)
Kirsten Skogerson (NMR, GCxGC)
Dr. Kwang-Hyeon Liu (LC, Pharma)
Sangeeta Kumari (GCT, GC-MS)
Sevini Shahbaz (Library)
Kristie Cloos (Lipids, MS, GC-MS)
Dr. Pierre Ayotte (Docking)
John Meissen (UPLC, LC)
Dr. Do Yup Lee (now LBNL Berkley)

Sponsors Fiehn Lab

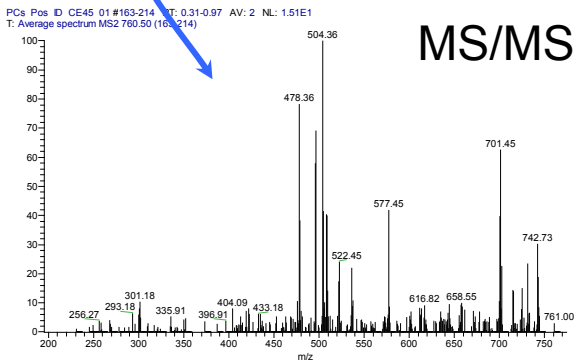
NIH R01 ES013932
NIH GM078233 & ARRA RC2
NIH R01 DK078328
NIH 1 R21 AI073323-01A1
UC Discovery itl07-10167
NSF MCB 0520140
EU FP7 Health-2007-2.1.4.1/Dupont
Agilent, LECO, Waters

Thanks to the useful LipidMaps service!
Please apply for beta-testing!

Tandem mass spectrometry (MS/MS)



MS

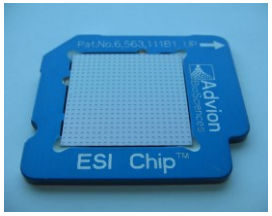


MS/MS

Iontrap MS/MS spectra creation

NanoMate nanoESI chip based infusion

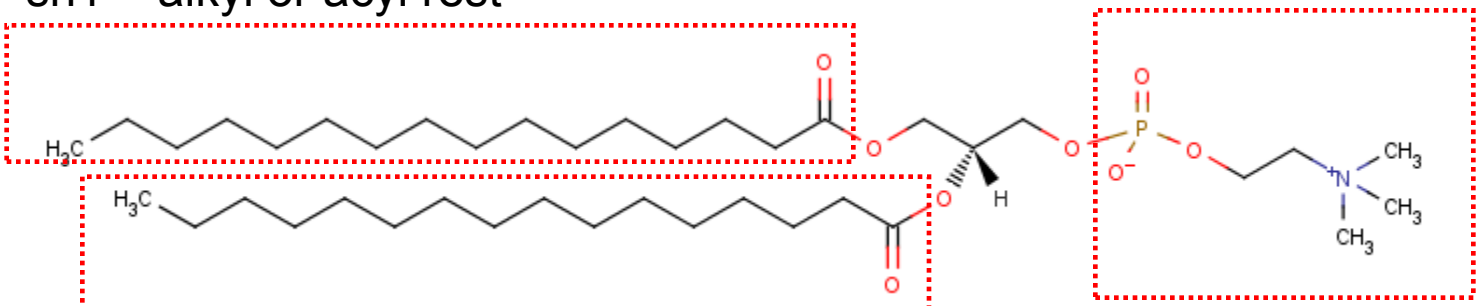
Low-resolution LTQ Ion Trap



nanoESI chip with 400 nozzles

High-resolution LTQ -FT

sn1 = alkyl or acyl rest



head group

sn2 = alkyl or acyl rest