De novo identification of small molecules with computer generated MS/MS libraries

Tobias Kind and Oliver Fiehn

UC Davis Genome Center - Metabolomics, Davis, CA

Introduction

The identification of small molecules using tandem mass spectrometry suffers from the non-existence of large MS/MS databases. Around 11.8 million chemicals are commercially available and approximately 100 million compounds are known in compound databases such as PubChem, Chemspider, CAS and CSLS. The largest MS/MS databases from NIST, MassBank, Metlin and ReSpect DB only cover around 10,000 compounds with a series of tandem mass spectra obtained under different voltages and ionization modes. Computer generated (in-silico) mass spectral databases can be created to fill that gap. Unlike in proteomics where MS/MS information can be deduced from large genomic sequence databases, such an approach is not directly applicable for diverse small molecules. We used a cheminformatics algorithm to analyze molecule classes with consistent fragmentation patterns and generated in-silico tandem mass spectral libraries for such small molecule compound classes.

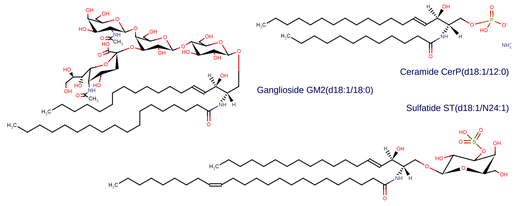
Methods

Compound structures were obtained from the LipidMaps compound database or were generated using combinatorial algorithms. Structures were transferred using ChemAxon Instant-JChem into EXCEL templates. The generation of in-silico mass spectral databases requires the modeling of fragmentation patterns, peak abundances and peak annotations. Our algorithm uses deterministic models for fragmentation patterns and heuristic models for peak abundance modeling. Visual Basic for Applications was used to export the data to an external exchange format. For MS/MS library search we used the freely available NIST algorithm (NIST MS Search Software 2010) with accurate mass pre-filter and dot-product matching. The library was validated with a three step process: 1) library search in itself 2) decoy database search with spectra that do not belong to the specific compound class and 3) library search of independent experimental spectra. Tandem mass spectra for validation purposes were collected from existing research publications. Experimental tandem mass spectra from triple quadrupole, Orbitrap, FT-ICR-MS and ion trap mass spectrometers covering ESI and MALDI ionization were investigated.

Results

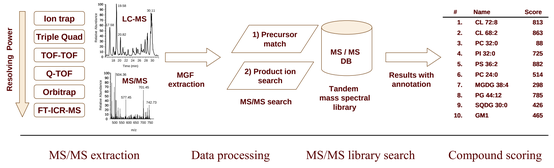
An accurate mass MS/MS library of 1384 tandem mass spectra with different adducts covering 168 ceramide-1-phosphates (CerP), 168 sulfatides (ST) and 880 gangliosides ([glycan]-Cer) was computationally created. Validation spectra from ceramide-1-phosphates, a series of sulfatides and the glycans of gangliosides including GM1, GM2 and GD1a are shown here. The use of MS/MS fragmentations, compared to simple accurate mass lookup, introduces an additional level of confidence because not only the accurate precursor mass is used, but additional molecule fragmentation data is included for comparison of spectra from different structural isomers. MS/MS database search uses two selective filters. The first filter is the precursor filter that searches an accurate mass within a certain m/z window. The second filter is a classic mass spectral database matching algorithm that generates a search score. The broad availability of instruments that generate MS/MS spectra allows the fast annotation of tandem mass spectra with MS/MS library search using experimentally and in-silico generated tandem mass spectra.

Gangliosides, ceramides and sulfatides



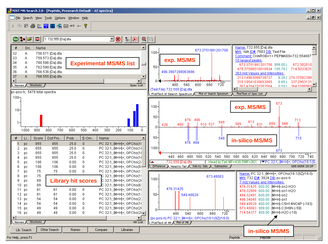
Glycosphingolipids and their related synthases are important regulators in signal transduction processes. They are ubiquitous in almost all vertebrates and bacterial cells. They recently have been overexpressed in various types of cancer such as breast, bladder and lung cancer but are also important in age research, stem cell research, for the investigation of autoimmune diseases and nutritional research.

MS/MS search for faster compound ID



Due to the lack of large MS/MS libraries (>100k spectra) tandem mass spectral library search for small molecules did not yet evolve to its true potential. The approach uses data-dependent scan extraction, the algorithm then filters all precursor ions with a preset search window (0.4-0.001 Da) and scores all remaining product ion spectra with a dot-product library search algorithm. The availability large MS/MS libraries from reference compounds or in-silico generated MS/MS spectra will enable faster and automated compound annotations. All tandem mass spectra originating from ion traps, triple quads, TOF-TOFs, Q-TOFs, Orbitraps and FT-ICR-MS can be utilized.

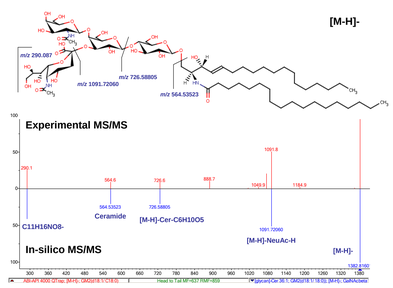
MS/MS search with NIST MS Search GUI



Traditional MS library search allows the deeper inspection and comparison of tandem mass spectra. This approach is superior compared to a manual inspection with simple fragment tables only or black box algorithms. Unknown spectra can be moved to a user library for later inspection. The freely available NIST MS Search program GUI can be used for visual inspection of the library hits. It also can be integrated into existing LC-MS software.

Precursor and product ion m/z tolerances can be adjusted according to high resolution (0.001 m/z unit window) or low resolution instruments (0.4-0.8 m/z unit window) used. After the precursor filter is applied, a dot product hit score and a probability match factor (PBM) and is calculated for each spectrum.

Gangliosides in-silico MS/MS library match

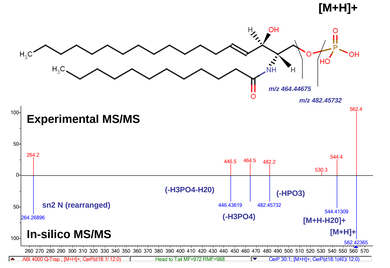


An MS/MS library search yielded two possible hit candidates. The first hit revealed the correct glyco-sphingolipid [glycan-Cer] GM2(d18:1/18:0).

Relevance: E. coli O157:H7 (EHEC) and their related bacterial proteins (Shiga toxins) utilize gangliosides GT1, GD1b, GM1 and GM2 as receptors on host cells.

Experimental spectrum: ABI 4000 QTrap; AnL Chem., 2008, 80 (8), pp 2780-2788

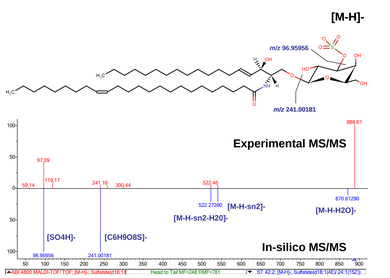
Ceramide-phosphates MS/MS library match



The library search against the in-silico library revealed three possible hits with high library scores. The first hit was the correct CerP(d18:1/12:0) with library scores higher than 900 (maximum is 1000). The ceramide-1-phosphate library covers positive mode [M+H]+ and negative mode spectra [M-H]-.

Experimental spectrum: ABI 4000 QTrap; LipidMaps LMSP02050001

Sulfatides in-silico MS/MS library match



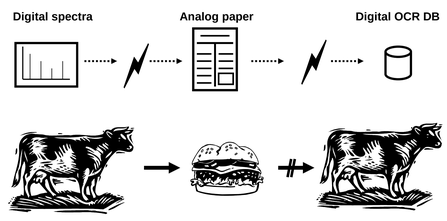
MS/MS library search with a 0.4 Da precursor window revealed two hit candidates with reverse scores > 800. The first hit correctly identified Sulfatide(d18:1/N24:1). The second sulfatide with the same accurate precursor mass was not matching because of different sn2 acyl chain losses.

Experimental spectrum: ABI 4800 MALDI-TOF/TOF; J Lipid Res. 2010 Jun; 51(6):1599-609

Instrument coverage of MS/MS search

Compound name	Instrument	Mode	Library score Dot-product	Library score Reverse dot	External MS/MS sources
CerP(d18:1/12:0)	4000 QTRAP	[M-H]-	761	762	LipidMaps
CerP(d18:1/12:0)	4000 QTRAP	[M+H]+	988	991	LipidMaps
Ganglioside GM2(d18:1/C18:0)	4000 QTRAP MALDI	[M-H]-	859	985	Chen Y. et al. Anal. Chem., 2008, 80 (8), pp 2780-2788
Ganglioside GM2(d20:1/C18:0)	4000 QTRAP MALDI	[M-H]-	622	793	Chen Y. et al. Anal. Chem., 2008, 80 (8), pp 2780-2788
Sulfatide(d18:1/N24:1)	4800 TOF/TOF MALDI	[M-H]-	781	901	Cheng H. et al. JLR, Vol. 51, 1599-1609, 2010
Sulfatide (d18:1/C22:0)	QSTAR-XL QTOF	[M-H]-	802	809	Sullards et al. Keystone 2007
Sulfatide(d18:1/16:0)	QSTAR-XL QTOF	[M-H]-	626	818	Liu et al. Molecular Cancer 2010, 9:186
Ganglioside GM3	JMS-HX110A/110A	[M-H]-	374	824	Jhon et al. Chem Pharm Bull (Tokyo). 1999 Jan;47(1):123-7.
Ganglioside GM2-alpha	LTQ-FT	[M-H]-	436	617	He et al. Anal Chem. 2007 15;79(22):8423
Ganglioside GM1(d18:1/18:0)	Orbitrap Velos	[M-H]-	651	938	Colsch et al. ASMS 2010
Ganglioside GM1	VG Quattro II	[M-H]-	659	895	Whitfield et al. Acta Neuropathol. 2000 Oct;100(4):409-14.
[glycan]-Cer(d18:1, C24:1)	QTOF Premier	[M-H]-	806	834	Mülthing et al. Mass Spectrom Rev. 2010; 29(3):425-79

MS/MS data sharing paradigm shift



Publishing MS/MS spectra on paper or bitmap-PDF does not hold up to current technological standards. The re-capturing process is error prone (hamburger-to-cow algorithm). Electronic submission of mass spectra in open exchange formats with each publication will allow thorough validation of published claims and speed-up theoretical research and development of better algorithms. Such requirements are common standard in crystallography and genomics.

Conclusions

The generation of in-silico tandem mass spectra can be a fast-lane for successful structural annotation of complex molecules by using automated MS/MS annotations similar to X!Tandem, Mascot, OMSSA or Sequest.

High precursor mass accuracy usually yields fewer result candidates. However fragment rich MS/MS spectra from unit resolution mass spectrometers are also sufficient because the subsequent dot-product algorithm generates higher hit scores when fragment rich low resolution spectra are provided. Chromatographic or ion mobility separation is needed in case of very similar structures or stereoisomers to allow compound annotations with the highest level of confidence.

The development of in-silico mass spectra is hindered by ancient publishing strategies of MS/MS spectra or publications about tandem mass spectrometry without any spectra at all. New data sharing principles have to be established and mandated. Community efforts are needed to enhance compound coverage and rapidly increase the number of MS/MS spectra.

The in-silico MS/MS library is freely available for commercial and academic research. Support: NSF MCB0520140, NIH GM092729 and NIH DK078328.