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, ChemoSpider, CASt and CEISL. The largest MS/MS databases from NIST, MassBank, Merlin and RetinLab 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 LipiMaps compound database or were generated using combinatorial algorithms. Structures were transformed using ChemAxon Instant-JChem into EXCELL 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 library (NIST MS Search Software 2010) with accurate mass pre-filter and dot-product matching. This library was validated with a three step process: 1) library search in itself 2) decay 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 and from in-silico MS/MS spectra from triple quadrupole, Orbitrap, FT-ICR-MS and ion trap mass spectrometer experiments covering ESI and MALDI ionization were investigated.

Results

An accurate mass MS/MS library of 1384 tandem mass spectra with different adducts covering 146 ceramides (CerP), 368 sulfatides (ST) and 880 gangliosides (Glycan-Ceri) was computationally created. Using in-silico generated CerP, ST and sulfatide mass spectra, a series of sulfatides and the glycosyl ceramides including GM1, GD1a and GD2 are shown here. The use of MS/MS fragmentations, compared to simple tandem MS, offers 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

An MS/MS library search yielded two possible hit candidates. The first hit revealed the correct glycosphingolipid (Glycan-Ceri) by its retention index (RI). The second hit revealed a glycosphingolipid and their related synthetic analogs are important regulators in signal transduction processes. They are ubiquitous in almost all vertebrates and bacteria 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 neurological research.

Gangliosides in-silico 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 glycosphingolipid (Glycan-Ceri) with library scores higher than 900 (maximum: 1000). The ceramide-1 phospholipid library covers positive mode (M+) and negative mode spectra (M-). The second hit revealed the correct glycosphingolipid by its RI and its MS/MS spectrum. The third hit revealed a compound originating from in, triple quadrupole, Orbitrap, FT-ICR-MS and ion trap mass spectrometers covering ESI and MALDI ionization were investigated.

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 glycosphingolipid (Glycan-Ceri) with library scores higher than 900 (maximum: 1000). The ceramide-1 phospholipid library covers positive mode (M+) and negative mode spectra (M-). The second hit revealed the correct glycosphingolipid by its RI and its MS/MS spectrum. The third hit revealed a compound originating from in, triple quadrupole, Orbitrap, FT-ICR-MS and ion trap mass spectrometers covering ESI and MALDI ionization were investigated.

Conclusions

The generation of in-silico tandem mass spectra can be a fast lane for biochemical studies aiming at complex molecule identification. In-silico MS/MS annotations similar to X/Tandem, Mascot, OMSSA or Sequest. High precursor mass accuracy usually yields fewer ion candidates. However, fragment rich MS/MS spectra from unit resolution mass spectrometers are more abundant for 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 isomerism 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 maintained. Community efforts are needed to enhance compound coverage and rapidly increase the number of MS/MS spectra.

MS/MS data sharing paradigm shift

The in-silico MS/MS library is freely available for commercial and academic research. Support: NIDDK5821495, NH9052752 and NH401797528.