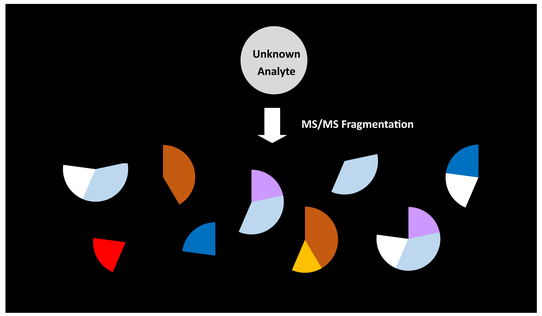
New Excel Add-In Worksheet for Substructure Searching from Accurate-Mass Fragmentation Spectra

When a molecule is fragmented in a mass spectrometer, we could visualize many of its fragment ions as being composed of smaller imaginary subfragments. A subfragment could be considered a block of atoms that tend to hold together during the fragmentation process. Our Excel Add-in is a tool to help identify an unknown analyte by "reassembling" that analyte structure from these subfragments.



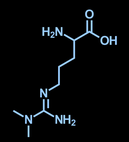
Our *Rational Numbers* Excel Add-in calculates the masses of possible subfragments and finds sets of these masses that sum up to the accurate mass of the whole analyte.

After finding these sets of subfragment masses, this Excel Add-in converts these subfragment masses into corresponding subformulas aligned in space as two-dimensional modular structures.

Many smaller subformulas can be directly related to only a few substructures. However, substructure searches in Pubchem previously could not be limited to a specific molecular formula or a limited accurate-mass slice, and therefore searching the entire database of over 100 million compounds for common small substructures (e.g., methoxy) would either time out or stop at 1 million hits.

Recently, scientists at the NCBI developed "fast substructure searching" which limits a substructure search to either a single molecular formula or a small accurate-mass slice of the massive Pubchem Compound database (1). Substructure searching of small substructures is now very practicable and it is available as a new worksheet in our Excel Add-in.

1. Kim S, Thiessen PA, Cheng T, Yu B, Bolton EE. NIH, National Center for Biotechnology Information. An update on PUG-REST: RESTful interface for programmatic access to PubChem. Nucleic Acids Res. 2018 Jul 2;46(W1):W563-W570. doi: 10.1093/nar/gky294. PMID: 29718389; PMCID: PMC6030920.



ACCESSION: MSBNK-MSSJ-MSJ00345

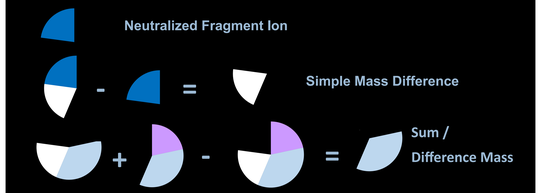
RECORD_TITLE: Asymmetric dimethylarginine; ESI-QTOF; MS2; POSITIVE; [M+H]+; CID; 20 V

DATE: 2020.11.12

AUTHOR: Atsushi Yamamoto, Faculty of Environmental Studies, Tottori University of Environmental Studies, 1-1, Wakabadai-kita, Tottori City, Tottori 689-1111, Japan.

This accurate-mass MS/MS spectrum of dimethylarginine is used here to illustrate substructure searching with the Excel Add-in.

Three Types of Subfragment Masses



Mathematical Partition

450572 + 420222 + 690582 + 460042 = 2021418

A "partition" is a set of integers that sum up to a larger integer. Here the masses are shown in units of tenths of millidaltons.

Partition of a Set

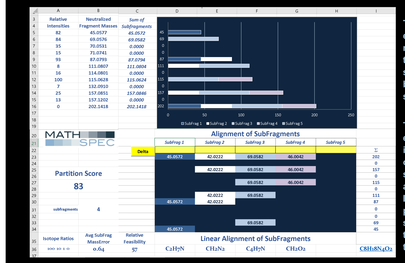
C_2H_7N + CH_2N_2 + C_4H_7N + CH_2O_2 = $C_8H_{18}N_4O_2$

A partition of a set of atoms (e.g., formula $C_8H_{18}N_4O_2$) is a grouping of the set's atoms into non-empty subformulas, in such a way that every atom is included in one and only one of the subformulas. The subformulas are therefore mutually exclusive.

Subformulas to SMARTS Substructure Representations

C_2H_7N	[s(N(C([H])([H])[H])C([H])([H])[H]),s(C(N)([H])([H])C([H])([H])[H])]
CH_2N_2	[s(C(~N)N),s(C~N~N)]
C_4H_7N	
CH_2O_2	[s(C(O)=O)&!s(C(=O)OC([H])([H])[H])]

Some small subformulas can be converted by the Excel Add-in into "SMARTS" representations of molecular substructures. Combining each of these representations together, separated by a period (dot), a small subset of Pubchem Compound (e.g., Formula = $C_8H_{18}N_4O_2$) can then be filtered for compounds with these mutually exclusive substructures by using "fast substructure searching".

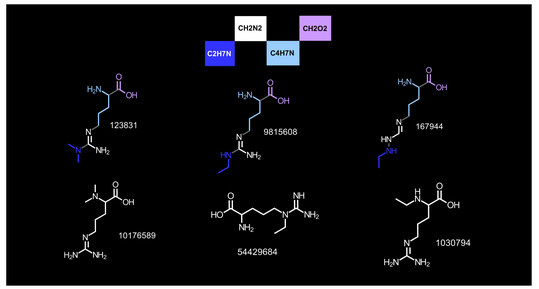


The Excel Add-in generates models, called modular structures, that view some mass spectral fragments as being composed of smaller subfragments.

The subfragment masses are then converted into one or more sets of subformulas. The subformulas are also aligned in space, here linear, based on comparing subfragment assignments with truth tables. This analysis took 13 seconds.

The Search_Within Worksheet

Three of the small subformulas in Row 36 of the Partition above and column D here were converted by the Excel Add-in into "SMARTS" representations of molecular substructures (column E). The Add-in also combined these smarts representations together, separated by a period (cell B4). The subset of Pubchem Compound having all 826 compounds with the formula $C_8H_{18}N_4O_2$ was then created and stored temporarily in a cache in Pubchem. Using the Text CacheKey (cell B3), this cache of 826 compounds was then filtered for compounds in the cache having all three of these mutually exclusive substructures by using "fast substructure searching". This produced a final listing of 38 compounds.



Above, six of the 38 molecules in Pubchem Compound that have both the formula $C_8H_{18}N_4O_2$ and the three SMARTS substructures are shown. The three molecules at the top have both the linear alignment and ordering of the substructures that are consistent with modular structure (top); the three molecules at the bottom have substructure connections that are inconsistent with that modular structure. The spectrum analyzed was generated from Pubchem CID 123831 (top left).

Mass spectrometry deals with combinations of atoms. Because combinations have no ordering, mass spectral interpretation is always somewhat ambiguous. This Excel Add-in illustrates the multiple possible interpretations of this spectrum. Definitive identifications can be obtained by matching an unknown analyte with an authentic sample.

MATH SPEC

Rational Numbers Excel Add-In (https://www.mathspec.com) This Excel Add-in rapidly transforms accurate-mass fragmentation data into modular structures.

Hyperlinks: Pubchem formula / EPA accurate mass
Special Worksheets: High-resolution Isotope patterns / Substructure Searching
Functions: Mass / RDBE / BaseRate / Add Formulas / Subtract Formulas / ET
Price: $249 USD equivalent + any applicable taxes

Similar Partition Screening (https://aws.amazon.com/marketplace)

This SPS software that screens for 240000 common compounds, subtracts blanks, annotates adducts, and summarizes all of your accurate-mass LCMS data for importing into Microsoft Access, is launched as a "server" (or EC2 on Amazon Web Services) from our AMI whenever you want to use SPS. There are no high upfront costs or subscriptions, and it comes with a Microsoft Access template and a free 14-day trial. This software also makes it easy to import spectra into our *Rational Numbers* Excel Add-In.

Excel Add-In

SPS on Amazon AWS

YouTube: Substructure Searching