CASMI2016から学ぶ化合物構造推定

理研CSRS（リーダー・有田正規）・理研IMS（リーダー・有田誠）
津川 裕司
CASMI 2017 will focus on testing current mass spectrometric approaches for identification of natural products. This year, CASMI will again have 3 categories based on (subsets of the) 243 challenges. These are:

**Category 1**
- Manual approaches: Challenges 1-45

**Category 2**
- Automated approaches (spectral data only): Challenges 1-243

**Category 3**
- Automated approaches (all metadata in): Challenges 1-243

If you are unsure in which category your approach fits, do not hesitate to contact the organisers (email link at the bottom). Challenge data and details on the experiments are available from the navigation menu above.

Update 20170522: We have fixed MS1 data for challenges 29, 42 and improved challenges 71, 89, 105, 106 and 144.
このMS/MSスペクトルの元構造は何でしょう？

1. Negative, <10ppm, Human
   - 78.9597
   - 158.9249
   - 256.9616
   - 274.9731
   - 368.00
   - Dibromophakellin
     - C_{11}H_{11}Br_{2}N_{5}O
     - PubChem: 42636938
     - ChemSpider: 9592765

3. Negative, <10ppm, Human
   - 79.9588
   - 80.9665
   - 132.046
   - Oroidin
     - C_{11}H_{11}Br_{2}N_{5}O
     - PubChem: 6312649
     - ChemSpider: 4880362

4. Positive, <3ppm, Plant (Onion)
   - 386.0155

2. Positive, <3ppm, Drug
   - 96.9693
   - 274.9731
   - 368.003
   - Cytochalasin B
     - C_{29}H_{37}NO_{5}
     - PubChem: 5311281
     - ChemSpider: 4470791
Computational MS と データベース
話の流れ

1. 組成式の決定方法
2. データベースの活用
3. MS/MSに対する部分構造アサインメント
4. 構造候補の検証
1. Negative, <10ppm, Human

\[ m/z \ 386.0155 \pm 0.005 \text{ as } [M-H]^- \]

\[ C_xH_yN зO_uS_vP_wF_aCl_bBr_cI_d \]

C: 12.0000000  O: 15.9949146  F: 18.99840322
H: 1.00782503  S: 31.9720710  Cl: 34.96885268
N: 14.0030740  P: 30.9737616  Br: 78.9183371
\( e^- \): 0.00054857  I: 126.904473
組成式の絞り込み：ルールの活用

✓原子価測の確認
例）C₈H₇NO₄S
TA: 8+7+1+4+1 = 21
OV: 7+1 = 8
SV: 32+7+3+8+6=56

原子価が奇数の元素の総数をOV
原子価の総数をSVとしたとき,

OV * SV = even
SV >= 2* (TA - 1)

ただし, N(3), O(2), P(5), S(6)

✓経験的閾値

<table>
<thead>
<tr>
<th></th>
<th>Mass</th>
<th>H/C</th>
<th>N/C</th>
<th>O/C</th>
<th>P/C</th>
<th>S/C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>1999.90</td>
<td>8.00</td>
<td>4.00</td>
<td>10.00</td>
<td>3.00</td>
<td>6.00</td>
</tr>
<tr>
<td>Min</td>
<td>50.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean</td>
<td>601.03</td>
<td>1.38</td>
<td>0.10</td>
<td>0.32</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Stdev</td>
<td>349.29</td>
<td>0.43</td>
<td>0.15</td>
<td>0.27</td>
<td>0.04</td>
<td>0.09</td>
</tr>
<tr>
<td>Median</td>
<td>512.32</td>
<td>1.39</td>
<td>0.04</td>
<td>0.27</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Percentile 0.15%</td>
<td>72.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Percentile 0.005%</td>
<td>52.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Percentile 99.85%</td>
<td>1939.74</td>
<td>3.33</td>
<td>1.20</td>
<td>2.20</td>
<td>0.40</td>
<td>1.00</td>
</tr>
<tr>
<td>Percentile 99.995%</td>
<td>1998.10</td>
<td>6.00</td>
<td>4.00</td>
<td>6.00</td>
<td>1.83</td>
<td>3.00</td>
</tr>
</tbody>
</table>
組成式の推定・決定 by Computational MS

1. Negative, <10ppm, Human
   - \( \text{C}_9\text{H}_{15}\text{N}_3\text{O}_{10}\text{P}_2 \)

2. Positive, <3ppm, Drug
   - \( \text{C}_{33}\text{H}_{40}\text{N}_2\text{O}_9 \)

3. Negative, <10ppm, Human
   - \( \text{C}_8\text{H}_7\text{NO}_4\text{S} \)

4. Positive, <3ppm, Plant (Onion)
   - \( \text{C}_{13}\text{H}_{23}\text{N}_9\text{O}_4\text{S}_2 \)
Computational MSとデータベース

保持時間 同位体比 精密質量 MS/MSスペクトル

組成式計算 → C_{11}H_{18}N_{2}O_{5}S

構造検索

代謝物データベース (HMDBなど)

In silico フラグメントと
実測 MS/MSの比較

ランキング

Rank  Candidate
1  
2  
3  
4  
5  

妥当性の検証

m/z
組成式に基づく化合物構造の検索

4. Positive, <3ppm, Plant (Onion)

\[ C_{13}H_{23}N_9O_4S_2 \]

KNApSAcK (45,852)  UNPD (166,995)  PlantCyc (3,817)
FooDB (21,943)  NAPNDB (3,882)  STOFF-IDENT (10,231)
LipidMAPS (34,466)  ChEBI (53,746)  DrugBank (6,273)
SMPDB (1,490)  YMDB (1,840)  ECMDB (1,342)
T3DB (2,499)  BMDB (7,232)  HMDB (95,924)  Urine (3,929)
Saliva (1,129)  Fecal (1,073)  CSF (360)  Serum (23,138)
Primary the compound database ‘STOFF-IDENT’ is used for identification purposes, more exactly for non-target screening (‘Hidden Targets’ and/or ‘Known Unknown’) and suspected-target screening analysis in different parts of water research. This is why just water relevant organic molecules, their transformation products and metabolites that occur in the environment are listed.²

Besides common compound parameters as Name, CAS Number, Formula, InChi key, IUPAC Name, SMILES Code, monoisotopic mass, also physicochemical indications as the logP and the logD value for four different pH values are included.²
MS-FINDERに考えさせた結果の比較
化合物同定に必要な行程まとめ

(a) Eliminate false positive peaks
1. Deisotoping
2. Adduct picking by mass shifts
3. Checking peak shape correlations
4. MS/MS matching (if available)
5. Abundance correlations (in large-scale assay)

(b) Search spectral libraries
- MetaboBASE
- NIST
- GNPS
- CASMI spectra
- ReSpect
- HMDB
- Metlin
- Internal DB
- MassBank
- Theoretical MS/MS (like LipidBlast)

(c) Predict molecular formula
- Ultrahigh resolution spectrum (if available)
- MS/MS
- Formula assignments to product ions
- Ranking
- C_{10}H_{15}N_{2}O_{9}P
- C_{9}H_{15}N_{2}O_{9}P
- C_{9}H_{11}O_{7}
- C_{10}H_{11}O_{7}P

(d) Retrieve structures followed by their ranking
- Computational fragment assignments
- Rule N2
- Rule N1
- MS spectrum
- [M-H]^{-}

(e) Expand the chemical spaces & Predict the molecular scaffold
- Molecular-spectrum networking
- Methylation? 14 Da difference
- In silico bioreactions
- Ranking candidates & standard synthesis
- Computational fragment assignments
- Rule N2
- Rule N1
- MS spectrum
- [M-H]^{-}

Check hydrogen rearrangement rules
Evaluate fragment stabilities
Utilize machine learning approaches

No candidates

Reference
RIKEN PRIME website

Platform for RIKEN Metabolomics: PRIME

About PRIME, PRIME, the Platform for RIKEN Metabolomics, is a Web-based service for metabolomics.

Software tools & Databases

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABF converter</td>
<td>ABF binary format file can be used in MS-DIAL and MRMPROBS, and the converter is freely available at the RIKEN Converter Website. The supported vendor- and common MS format is listed in their homepage.</td>
</tr>
<tr>
<td>MS-DIAL</td>
<td>MS-DIAL is a universal program for untargeted metabolomics and lipidomics supporting any type of chromatography/mass spectrometry methods (GC/MS, GC-MS/MS, LC/MS, and LC-MS/MS etc.).</td>
</tr>
<tr>
<td>MS-FINDER</td>
<td>MS-FINDER is a universal program for compound annotation supporting EI-MS (GC/MS) and ES-MS/MS spectra. It aims to provide solutions for 1) spectral searching, 2) formula prediction, and 3) structure elucidation for unknown spectra.</td>
</tr>
<tr>
<td>EI-MS and MS/MS library</td>
<td>We released several MSP files including both EI- and MS/MS spectra as a 'start-up kit'. Nowadays, MS-DIAL/MS-FINDER internally leave a version of Fiehn Lab's GC/MS database, and in silico RF- and MS/MS database for lipidomics.</td>
</tr>
<tr>
<td>MassBank to MSP</td>
<td>This program helps you to prepare NIST MS format libraries applicable to a lot of programs (MS DIAL, NIST MS Search, AMDIS) from MassBank records.</td>
</tr>
<tr>
<td>MRMPROBES/MRMDIFF</td>
<td>MRMPROBES is launched as a universal program for targeted metabolomics supporting not only multiple reaction monitoring (MRM) but also SCAN and data independent MS/MS acquisition (DIA) data.</td>
</tr>
<tr>
<td>MRM database</td>
<td>Several conditions/reference libraries for large scaled MRM assay are available.</td>
</tr>
<tr>
<td>Smetsearch</td>
<td>The purpose of this project is to develop the methodology for formula identification by ultrahigh resolution MS instrument in combination with labeled experiments.</td>
</tr>
</tbody>
</table>