Building the NIST Tandem Mass Spectral Library 2014

Xiaoyu (Sara) Yang
Mass Spectrometry Data Center

Biomolecular Measurement Division Seminar
September 9, 2014
Outline

• Methods of building the NIST tandem mass spectral library

• Quality Control
  o Peak annotation
  o Noise Removal
  o Chemical information consistency

• Major types of mass spectra

• Major types of compounds
Mass Spectral Library Searching

Experimental mass spectrum

Searching

Reference Spectra in MS Library

Matched Reference

MS Search
Rationale and Objectives:

• Chemical identification through Electrospray ionization (ESI) tandem mass spectrometry (MS/MS) is becoming a routine technique in metabolomics, proteomics and other fields.

• The identification can be aided by matching the acquired tandem mass spectra against reference library spectra.

• We are developing a comprehensive library of high quality reference ESI tandem mass spectra for the identification of compounds through the ion fragmentations.

Goals:

• Develop a tandem mass spectral library of all biologically relevant metabolite ions.

• Provide the library in a form that is easily searchable using software tools.
Steps of Building the NIST Tandem Mass Spectral Library

Authentic samples

- metabolites, drugs, peptides
- lipids, pesticides, surfactants, glycans, sugars

LC/MS/MS

- Ion Trap (LTQ, IT/FTMS)
- Collision Cell (HCD, QQQ, QTOF)

Cluster MS² spectra using precursor m/z

- count-based clustering algorithm
- cluster spectra of the similar precursor m/z values

Create consensus spectra of MS², MS³ and MS⁴

- adjusted dot product-based clustering algorithm
- cluster spectra of the similar fragmentations

Precursor type identification

- precursor purity
- mass accuracy
- peak annotation
- noise removal

Manual inspection

Chemical structure, formula, name, synonyms, and CAS are consistent

MSMS library with multiple precursor types
Count-based Algorithm for Clustering Precursors

Steps:
1. Count the number of precursors (m/z count) within 0.1 m/z;
2. Sort the precursors by the m/z count in descending order;
3. Group similar precursors into the same cluster by using the precursor with the highest m/z count as the cluster center;
4. Repeat step 3 until all the precursors are clustered.
Clustering Algorithm for Generating Consensus Spectra

Steps:
1. Group similar spectra into the same cluster;
2. Generate one consensus spectrum from each cluster;
3. Pick the best consensus spectrum for the library.

Dot Product (DP) = \[ \sum \sqrt{I_1 \times I_2} / \sqrt{\sum I_1 \times \sum I_2} \]

Clusters the same ion at the same collision energy

MS/MS spectra

Consensus spectrum: Median peak intensity
Median m/z

1. Calculate adjusted dot product (DP)
2. Find a cluster center with the highest average DP (>0.7)

Consensus spectrum 1
Consensus spectrum 2

best consensus spectrum

Cluster 1
Cluster 2

1. 2. 3. 4. 5.

bin size

bin = 700 x 10ppm x 10^{-6} = 0.0070 m/z when m/z=700
Using Consensus Spectrum in the Library

- Eliminated low quality spectra by spectral clustering.
- Improved the spectrum quality by using the median of the m/z and intensity values.
- Realistically represented the characteristic fragmentations.

Same compound
Same precursor type
Same instrument
Same energy (cone, collision)
Same mode (+/-)
Same spectrum type (MS², MS³, MS⁴)

10-20 spectra / consensus spectrum
10-20 energy levels
## What Precursor Types are in the NIST MSMS Library?

<table>
<thead>
<tr>
<th>Compound Type</th>
<th>ESI Product</th>
<th>Ionic Species</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organic Salt Cations</strong></td>
<td>Positive Ions (1,751; <strong>1%</strong>)</td>
<td>[Cat]⁺, [Cat+H]²⁺, [Cat-H₂O]⁺, [Cat-NH₃]⁺, [Cat+H₂O]⁺</td>
</tr>
<tr>
<td></td>
<td>Negative Ions (40; &lt;<strong>0.1%</strong>)</td>
<td>[Cat-2H]⁻, [Cat-2H-H₂O]⁻, [Cat-2H-NH₃]⁻, [Cat-2H+H₂O]⁻, [Cat-2H+NH₃]⁻</td>
</tr>
</tbody>
</table>
Multiple Precursor Ions for More Flexible Identification

inosine 5'-diphosphate acquired on Orbitrap HCD
Structure dependent losses:

$2\text{H}_2\text{O}$, $3\text{H}_2\text{O}$, $\text{NH}_3+\text{H}_2\text{O}$, $\text{H}_2\text{S}$, $\text{HCl}$, $\text{H}_3\text{PO}_4$, $\text{HCN}$, $\text{H}_2$, $\text{CO}$, $\text{CO}_2$, $\text{HCOOH}$, $\text{CH}_4$, $\text{CH}_3$, $\text{CH}_3\text{OH}$, $\text{CH}_3\text{SH}$, $\text{C}_2\text{H}_5\text{OH}$, ...

What Precursor Types are in the NIST MSMS Library?
In Source Fragmentation Confirms Metabolite Identification

MS2 spectra acquired on Orbitrap HCD
Estra-1,3,5(10),7-tetraene-3,17β-diol [M+H]^+
Estra-1,3,5(10),7-tetraene-3,17β-diol [M+H]^+
1. Satellite peaks in Orbitrap HCD spectra: due to the Fourier transform ringing artifacts

![Graph showing satellite peaks and their intensity relative to the main peak.]

<1% intensity of the main peak within ± 2 m/z

thiencarbazone -methyl [M+H]^+
2. Tailing peaks in QTOF spectra: due to the imperfect centroiding

1-Eicosatrienoyl-sn-glycero-3-phosphoethanolamine [M+H]^+

<20% intensity of the main peak after 0.3 m/z
3. Random noise peaks: in all mass spectra due to unstable instrument, impurity...

The occurrence < 25%

ocurrence = 1/3

= 33.3%

ocurrence = 2/3

= 66.7%
Noise Removal – an Example of Voting Algorithm

<table>
<thead>
<tr>
<th>m/z</th>
<th>intensity</th>
<th>occurrence</th>
<th>m/z</th>
<th>intensity</th>
<th>occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.6951</td>
<td>3.10</td>
<td>2/12</td>
<td>91.0543</td>
<td>803.20</td>
<td>12/12</td>
</tr>
<tr>
<td>58.2664</td>
<td>2.30</td>
<td>2/12</td>
<td>91.1262</td>
<td>2.10</td>
<td>1/12</td>
</tr>
<tr>
<td>60.8356</td>
<td>2.10</td>
<td>1/12</td>
<td>91.6671</td>
<td>2.10</td>
<td>1/12</td>
</tr>
<tr>
<td>66.5419</td>
<td>2.00</td>
<td>1/12</td>
<td>91.9893</td>
<td>43.16</td>
<td>12/12</td>
</tr>
<tr>
<td>66.8604</td>
<td>2.00</td>
<td>1/12</td>
<td>95.6621</td>
<td>2.00</td>
<td>1/12</td>
</tr>
<tr>
<td>72.1006</td>
<td>2.10</td>
<td>1/12</td>
<td>100.4113</td>
<td>3.40</td>
<td>2/12</td>
</tr>
<tr>
<td>72.1606</td>
<td>3.10</td>
<td>2/12</td>
<td>102.0466</td>
<td>203.30</td>
<td>12/12</td>
</tr>
<tr>
<td>74.0151</td>
<td>298.60</td>
<td>12/12</td>
<td>105.0454</td>
<td>12.89</td>
<td>6/12</td>
</tr>
<tr>
<td>75.0229</td>
<td>355.04</td>
<td>12/12</td>
<td>105.4044</td>
<td>2.00</td>
<td>1/12</td>
</tr>
<tr>
<td>76.0307</td>
<td>77.12</td>
<td>12/12</td>
<td>110.2882</td>
<td>2.00</td>
<td>1/12</td>
</tr>
<tr>
<td>77.0386</td>
<td>11.69</td>
<td>5/12</td>
<td>115.0430</td>
<td>4.20</td>
<td>2/12</td>
</tr>
<tr>
<td>78.0464</td>
<td>34.27</td>
<td>10/12</td>
<td>115.0545</td>
<td>999.00</td>
<td>12/12</td>
</tr>
<tr>
<td>79.2907</td>
<td>2.00</td>
<td>1/12</td>
<td>115.9932</td>
<td>2.00</td>
<td>1/12</td>
</tr>
<tr>
<td>80.5368</td>
<td>2.00</td>
<td>1/12</td>
<td>116.0622</td>
<td>8.49</td>
<td>4/12</td>
</tr>
<tr>
<td>81.0335</td>
<td>13.09</td>
<td>6/12</td>
<td>116.2749</td>
<td>2.00</td>
<td>1/12</td>
</tr>
<tr>
<td>81.3809</td>
<td>2.00</td>
<td>1/12</td>
<td>118.1334</td>
<td>2.00</td>
<td>1/12</td>
</tr>
<tr>
<td>83.4835</td>
<td>3.30</td>
<td>2/12</td>
<td>129.0451</td>
<td>2.00</td>
<td>1/12</td>
</tr>
<tr>
<td>86.0151</td>
<td>15.98</td>
<td>7/12</td>
<td>134.7912</td>
<td>2.00</td>
<td>1/12</td>
</tr>
</tbody>
</table>

4-methylcinnamic acid [M+H]^+  
HCD
Peak Annotation – Peptide
(for low and high resolution MS/MS spectra)

- \( p, y, b, a, \) internal fragments, neutral losses(-\( \text{H}_2\text{O}, \text{-NH}_3, \text{-CO} \))
- \( y+10(\text{CO-}\text{H}_2\text{O}), a2-45(\text{CONH}_3) \)
- 52 immonium ions (e.g. IHA: \( \text{C}_6\text{H}_7\text{N}_3\text{O} + \text{H} - \text{CO} \rightarrow \text{C}_5\text{H}_8\text{N}_3 \))
Peak Annotation - Small Molecule
(for high resolution MS/MS spectra)

- Peaks were annotated with the most probable chemical formula consistent with the precursor formula

  \[ \text{formula valence} = \sum \text{Count} \times (\text{valence} - 2) + 2 + \text{charge} \]

  ↑

  count of each element

  \[ \text{formula valence} \leq \# \text{ of H}; \ H:C \text{ ratio} \geq 0.125 \]

  Accuracy (ppm) = \( \frac{|(\text{Observed m/z} - \text{Theoretical m/z})|}{\text{Observed m/z}} \times 10^6 \)

  Unassigned (%) = \( \frac{\text{Sum of intensities of unassigned peaks}}{\text{Sum of intensities of all peaks}} \times 100 \)

  10 ppm

Peak Annotation - Small Molecule
(for high resolution MS/MS spectra)

C_{10}H_{12}O H^+
Cuminaldehyde [M+H]^+
theoretical precursor m/z= 149.0961
HCD energy=19eV

Unassigned=0.9%
Quality Control for Building Libraries from Electrospray Ionization Tandem Mass Spectra

Xiaoyu Yang, * Pedatsur Neta, and Stephen E. Stein

Mass Spectrometry Data Center, National Institute of Standards and Technology, Mail Stop 8362, Gaithersburg, Maryland 20899, United States

Supporting Information

ABSTRACT: Electrospray ionization (ESI) tandem mass spectrometry coupled with liquid chromatography is a routine technique for identifying and quantifying compounds in complex mixtures. The identification step can be aided by matching acquired tandem mass spectra (MS²) against reference library spectra as is routine for electron ionization (EI) spectra from gas chromatography/mass spectrometry (GC/MS). However, unlike the latter spectra, ESI MS² spectra are likely to originate from various precursor ions for a given target molecule and may be acquired at varying energies and resolutions and have characteristic noise signatures, requiring processing methods very different from EI to obtain complete and high quality reference spectra for individual analytes. This paper presents procedures developed for creating a tandem mass spectral library that addresses these factors. Library building begins by acquiring MS² spectra for all major MS¹ peaks in an infusion run, followed by assigning MS³ spectra to clusters and creating a consensus spectrum for each. Intensity-based constraints for cluster membership were developed, as well as peak testing to recognize and eliminate suspect peaks and reduce noise. Consensus spectra were then examined by a human evaluator using a number of criteria, including a fraction of annotated peaks and consistency of spectra for a given ion at different energies. These methods have been developed and used to build a library from >9000 compounds, yielding 230,000 spectra.

Mass spectral reference libraries of electron ionization (EI) spectra are used extensively and routinely to identify compounds separated by gas chromatography.¹ For example, the current NIST/EPA/NIH Mass Spectral EI Library contains spectra for over 200,000 compounds and is a common, tightly integrated component in many gas chromatography/mass spectrometry (GC/MS) data systems.²,³ Such use of reference libraries for the identification of electrospray ionization (ESI) tandem mass spectra (MS²) has, however, been far more limited. While certain MS² reference libraries are available for specific applications, such as MFLIN⁴ for metabolomics, they are often limited to specific platforms or not integrated with an instrument database system.⁵-¹⁰ Also unlike EI libraries,² NIST has undertaken the production of a comprehensive ESI MS² library for a wide range of molecules,¹¹,¹² intended for use on a variety of platforms and in a range of applications. Different methods are required for development of an ESI MS² library in comparison to those used for the odd-electron, positive ion, unit mass resolution MS¹ EI library. ESI MS² spectra are the result of even-electron transfer of ionic charge to neutral molecules in solutions at atmospheric pressure, frequently resulting in simple spectra with sparse fragmentation. The presence of multiple precursor ions and charge states for a single analyte is a necessary consequence of the ESI experiment in which protons or other cations or anions impart charge and form adducts with the neutral virus. Additionally, there is a
What Types of Mass Spectra are in the NIST MSMS Library?

**Instruments:**

Micromass Quattro Micro: Triple Quadrupole

Thermo Finnigan LTQ: IT/ion trap

Agilent QTOF 6530: Q-TOF

Thermo Finnigan Elite Orbitrap: HCD

IT-FT/ion trap with FTMS

IT/ion trap
What Types of Compounds are in the NIST MSMS Library?

**Metabolites ~50%**

Ureidosuccinic acid

\[ [M+H]^+ \]

HCD 14eV
What Types of Compounds are in the NIST MSMS Library?

Drugs ~20%

Acetaminophen
[M+H]+
QQQ 24V
What Types of Compounds are in the NIST MSMS Library?

Bioactive peptides ~10%

ARLDVASEFRKKWNKWLRS
[M+4H]^{4+}
QQQ 27V

IKD_{129.4}
b_{2-17} 211.5
y_{5-18} 307.9
y_3 374.5
y_4 445.5
y_5 631.6
b_{16} 673.1
b_{14} 852.1
y_6 759.9
916.5
b_9 989.5 y_{17} 1060.9

Life Extension
Bioactive Milk Peptides
Dietary Supplement 20 Capsules
What Types of Compounds are in the NIST MSMS Library?

All amino acids (20)
All dipeptides (400)
All tryptic tripeptides (800)
What Types of Compounds are in the NIST MSMS Library?

Lipids ~ 5%

Cholesterol
[M+H-H2O]+ QTOF 20V
What Types of Compounds are in the NIST MSMS Library?

Glycans ~ 2%

Glycan A1F-MIX
[M+2Na]^{2+}
HCD 78eV

Graphic showing molecular structures and mass spectrum data.
Like sugar? So does diabetes...

D-(-)-Galactose
[M+H]^+
QQQ 12V

Sugars ~ 2%
What Types of Compounds are in the NIST MSMS Library?

Surfactants and Contaminants

Sorbitan monopalmitate

\([\text{M+H}]^+\)

QQQ 14V
What Types of Compounds are in the NIST MSMS Library?

Pesticides

Atrazine

\[ [M+H]^+ \]

HCD 32eV

Pesticides
NIST Tandem Mass Spectral Library 2014

- 9,345 Compounds
- 45,298 Precursor Ions
- 234,284 Spectra
- ~90% Positive Ion Spectra
- ~10% Negative Ion Spectra

<table>
<thead>
<tr>
<th>Instrument Type</th>
<th>Precursor Ions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ion Trap</td>
<td>&gt;40,000</td>
</tr>
<tr>
<td>Collision Cell (QTOF, QQQ, HCD)</td>
<td>&gt;14,000</td>
</tr>
</tbody>
</table>
Conclusions

• Two level clustering algorithms (counter-based and distance-based) were developed and provide a robust means of generating consensus spectra of multiple precursor types for the NIST Tandem Mass Spectral Library.

• Quality control programs such as peak annotation and noise removal methods were developed and are used in building the reference quality NIST Tandem Mass Spectral Library.

• *NIST Tandem Mass Spectral Library 2014* can be applied in chemical identification in Metabolomics, Proteomics and other fields.

**Plans for the future:**

• Improve peak annotation program by using compound’s structure and physical chemical properties.

• Develop more quality control methods to eliminate low quality spectra for the library.
Acknowledgements