Mining the mass mess: Intelligent use of signal complexity simplifies MS based metabolomics.

Corey Broeckling
Colorado State University
Proteomics and Metabolomics Facility

Non-targeted metabolomics

• Targeted: MRM based acquisition
• Non-targeted: Unbiased acquisition
  • Goal: see as much small molecule signal as possible
  • Strength: breadth of data acquired
    • Unforeseen results revealed
    • Valuable outside model species/samples and in plants/microbes (secondary metabolism not well conserved)
• Limitations:
  • Sacrifice sensitivity compared to targeted (though less now than historically)
  • Signal Annotation/Compound identification
A standard workflow for non-targeted metabolomics data analysis

1. Detect features - a mass and time specific signal (AMT)
2. Align features across samples
3. To group or not to group...
   a) assume features are all independent
   b) ‘deisotope’ or group features based on predictable fragmentation, adduction, dimerization
4. Statistically interrogate either individual features or feature groups
5. ID based on inferred molecular weight/formula from 3 or follow-up targeted MS/MS.
   a) Often an additional experiment
   b) MS/MS offers more confidence in annotation through use of multiple signals for a given compound

Drug-like compounds set our ESI expectations: “this is pretty easy...”
Most biological metabolites are not drug-like

- 73% of HMDB compounds have 1+ Lipinski failure(s)
- Diverse structure leads to diverse behavior

Diverse structures -> Diverse behavior:

- ~1200 authentic standards run under real acquisition conditions – LC-TOF, positive ionization mode
- Some signals are easily predicted, others less so
- High risk of mis-interpretation!
In-source fragmentation happens

Adduction can be complex

Rows = cmpds
Columns = adducts
Standard workflow for metabolomics data analysis

1. Detect features - a mass and time specific signal (AMT)
2. Align features across samples
3. To group or not to group…
   a) assume features are all independent
   b) ‘deisotope’ or group features based on predictable fragmentation, adduction, dimerization
4. Statistically interrogate either individual features or feature groups
5. ID based on follow-up MSMS or inferred molecular weight from 3.

Logical flaws in 3
a) Features are not independent
b) Spectra are often unpredictable

• Implications of complex spectra:
  • Overestimation of sample complexity
  • Reduced spectral quality (and confidence) for known compounds
  • Wasted ID effort for redundant (identified) signals
• SPECTRA ARE INFORMATIVE: diagnostic and interpretable!
• Feature grouping tools: AMDIS(NIST), MSClust(PRI), QUICS(Metabolon), Parafac2(University of Copenhagen), CAMERA (IPB-Halle, Germany)…

Data Independent (MSe, MSall,...)

• MS^n: CID fragmentation without precursor isolation

• Concurrent acquisition, high and low collision energy
  • MS and MS/MS for all signals in single LC-MS injection
  • Issue: assigning precursor/product relationships
idMS/MS and in-source complexity: two problems, one solution

- Fragmentation/Adduction predictability unnecessary
- Integration reproducibility important
- Two parameters:
  - Retention time similarity
  - Correlational similarity

RAMClustR: custom similarity matrix

- Similarity between two features is the product of two gaussian functions (σ is tunable in each)
  - Correlation (quantitative similarity \([r]\), MS vs MS, MS vs MS/MS, MS/MS vs MS/MS)
  - Retention time (temporal coelution)
  - No cutoffs!
- If either correlation or retention time is dissimilar, total similarity approaches zero
- Similarity calculated for all pairs of features
RAMClust: Hierarchical clustering

- The similarity matrix is a n x n of similarities between features (i.e. correlational r-values)
- HCA clusters features based on this matrix
- Dendrogram can be ‘pruned’ into groups
  - ‘DynamicTreeCut’ – unsupervised cutting of dendrograms, no need to predefine expected cluster number
  - Groups = Spectra

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>1</td>
<td>0.75</td>
<td>0.02</td>
<td>0.45</td>
<td>0.1</td>
</tr>
<tr>
<td>F2</td>
<td>0.75</td>
<td>1</td>
<td>0.32</td>
<td>0.56</td>
<td>0.82</td>
</tr>
<tr>
<td>F3</td>
<td>0.02</td>
<td>0.32</td>
<td>1</td>
<td>0.13</td>
<td>0.82</td>
</tr>
<tr>
<td>F4</td>
<td>0.45</td>
<td>0.56</td>
<td>0.13</td>
<td>1</td>
<td>0.09</td>
</tr>
<tr>
<td>F5</td>
<td>0.1</td>
<td>0.82</td>
<td>0.82</td>
<td>0.09</td>
<td>1</td>
</tr>
</tbody>
</table>

Example spectral matches: idMS/MS spectra match MS/MS of Lipids

- Top: RAMClust spectra
- Bottom: NIST MS/MS spectrum
RAMClust overview:

- Only input: dataset(s)
- Independent of predictability in:
  - Adduction
  - Fragmentation
  - Isotope pattern
- Output to .msp format
  - Both MS and idMS/MS spectra
  - Viewing and searching using MSSearch (NIST)
- Fast
  - Dataset(s) of 17,000 features to msp spectral library < 200 seconds
  - Easier downstream: 17,000 features ~ 2700 clusters

- Spectral searching
  - More reliable than MW – multiple signals!
  - No assumptions regarding MW of compound
  - Spectral searching offers a shallow learning curve compared to spectral interpretation
- Spectra more interpretable than features
- Dependent on reliable feature detection and integration

Bonus: reduced variance when using spectra

- Redundant measures of metabolite abundance results in a lower CV
- Lower analytical CV -> better sensitivity to biological changes
Reduction in:
• Dimensionality
• Analytical variance

RAMClustR: data driven feature clustering

RAMSearch: metabolomics-centered spectral search GUI
Compound # >>> Spectra #

- METLIN: 961,829 compounds
  - ~ 14,000 with authentic spectra (image format)
- LipidMaps: ~40,000 compounds
  - ~ 500 with spectra
- PubChem: 93,553,257 compounds
- NIST
  - GC-MS EI spectra 267,376 compounds
  - LC-MS/MS spectra 14,351 compounds

- Authentic standard spectral libraries will be incomplete for the foreseeable future
  - Can predicted analytical behavior help?

CSU PMF Spectral and Retention Time Library

- ~ 900 authentic standards
- UPLC c8 reverse phase MeOH gradient
  - Phenyl Hexyl ACN
  - HILIC
- MS and MSE data for each compound
- Retention time for each compound
Adduction is not random

patterns contain information diagnostic of structure
In-source patterns are predictable

- Structure database converted to retention time and MS level spectral library
- No MS/MS necessary (though will often be beneficial)
- Utilize MS signal ‘redundancy’ for efficient and confident annotation!
- Approach doesn’t scale well

In-source spectra are searchable

- Annotation from MS1-Spectrum and Time Predictions
- Use trained model to predict in-source spectra and retention time for all HMDB compounds (~40K)
- Search in RAMsearch
- ~50% of time the top ranking match is CORRECT
- ~85% of the time the CORRECT answer is in the top 5 matches.
- ONLY using in-source peak patterns and retention time
In-source spectra are interpretable
Summary: In-Source Phenomenon

• Biological metabolites often generate complex MS spectra
  • Features can be grouped without chemical assumptions
    • RAMClustR
  • In-source spectral complexity can be useful
    • In-source spectral matching using
      • NIST Search tools
      • RAMSearch
    • 1-SToP approach
      • Predicted in-source spectrum and retention time
      • Theoretical MS & RT signals from chemical structures
        • HMDB, LipidMaps....
      • RAMSearch
    • Interpretation of in-source delta mass to obtain more confident molecular weight – interpretMSSpectrum (called from RAMClustR)

Heterodimers !???!?! Ackkk!?!!!
Future directions

• Hardware: Continue to explore chromatography and source conditions to better understand (predict) chromatographic and in-source behavior
• Software:
  • Measured analytical properties predictable from structure:
    • Accurate mass
    • Isotope pattern
    • MS1 spectrum
    • Retention time
    • Collisional Cross Section
    • MS/MS
  • No informatics platform uses all the available data for annotation!
    • Predicted analytical behavior will enable efficient use of structure databases
    • MSFinder, Sirius – MS/MS interpretation. Currently incorporating into XCMS/RAMclustR workflow.

Conclusions

• Many biological compounds generate a collection of signals.
  • In-source fragments
  • Alternate adducts
  • Multimers
  • Hetromultimers!
• This collection of signals is a mass spectrum, not a single m/z value
• This complexity is underappreciated in the metabolomics community
• This complexity provides additional structurally relevant signal that can be used to improve confidence in identification
• Ignoring this complexity is bound to result
  • False positive identifications
  • Weaker statistical analysis
  • Misinterpreted biology
• Use tools that recognize this complexity please!
Acknowledgements

• CSU
  • Jessica Prenni (PMF)
  • PMF lab
  • Kevin Brown, Ben Sutton, Graham Peers, and Mark Layer
    • RAMSearch
• Broad Institute
  • Andrea Ganna
• Stanford
  • Erik Ingellson

RAMClust: custom similarity matrix

\[
S_{ij} = \frac{1}{\alpha} \left( \frac{\alpha_1 e^{-\left(1-c_{ij}^{MS1/MS1}\right)^2/2\sigma_1^2} + \alpha_2 e^{-\left(1-c_{ij}^{MS2/MS2}\right)^2/2\sigma_2^2} + \alpha_{12} e^{-\left(1-c_{ij}^{MS1/MS2}\right)^2/2\sigma_{12}^2}}{e^{-(t_i-t_j)^2/2\sigma_t^2}} \right)
\]

• Similarity between two features is the product of two Gaussian functions (\(\sigma\) is tunable in each)
  • Correlation (quantitative similarity, MS-MS, MS-MS/MS, MS/MS-MS/MS)
  • Retention time (temporal coelution)
  • No cutoffs!
• If either correlation or retention time is dissimilar, total similarity approaches zero
• Use Data Dependent MS/MS precursor-product relationships to examine quality of clustering
  • Response: average spectral similarity for all feature-mapped DDA spectra which have similarity > 0.5 for ANY combination of parameters
RAMClust: Parameter descriptions:

- **Sigma t**: platform dependent
  - sigma for retention time
  - Wider chromatographic peaks means wider retention time variation for features representing same compound.

- **Sigma r**: platform independent
  - sigma for correlational r value
  - r-value is independent of signal
    - Though higher variation at lower signal intensity

- **Hmax**: platform independent
  - Hierarchical clustering dendrogram
  - maximum cut height using dynamicTreeCut package in R

- All other parameters set at feature detection

Influence of sigma t (st) and sigma r (sr)
DDA MS/MS spectra validate RAMClust relationships:

- Use known precursor-product relationships as the benchmark
- Any sim > 0.5
  - n=311 DDA MS/MS
- 390 combinations
- Y-axis: average spectral similarity (from 0-1)
- Stability to hmax improves at higher sigma_r and sigma_t

Can we control in-source complexity?
Solvent to analyte ratio controls sodium adduction

3-oxocholic acid: nonlinearity of dimer to monomer
3-oxocholic acid: ranked intensity shows strong linear relationship