Just how many unknowns are there in a metabolomics data set? Really?

Chris Beecher[1], Alexander Raskind[1], Casey Chamberlain[2], Joy Guingab[2], Rick Yost[2], Felice de Jong[1], Tim Garrett[2]
[1] IROA Technologies LLC, Bolton, MA, USA , [2] University of Florida, Gainsville, FL, USA

IROA TECHNOLOGIES

UF Southeast Center for Integrated Metabolomics
Clinical and Translational Science Institute

ABSTRACT

The question of unknowns has become probably the single most common question in metabolomics. It is generally suggested that there are hundreds or even thousands of unknowns in most datasets and that the number of unknowns is larger than the number of known (named) compounds. We have constructed a very specific dataset using IROA materials, and performed an in-depth LC-MS analysis of the peaks in these samples. Because of the IROA[1,2] patterning we can easily separate peaks of biological origin from artifacts, and using a specially written program we are annotating all of the biological peaks. The majority of unknown peaks of biological origins are fragment ions, adduct ions, or in-source polymeric ions. There appear to be few true unknowns.

INTRODUCTION

In every metabolomic dataset I have ever seen, the number of unknown (unnamed) compounds is larger than the number of known (named) compounds. Why is this? How can it be that there are so many unknowns when biochemistry is such a mature science? Can it really be true that there are two to three times as many unknown metabolites as there are known metabolites? This is possibly the single most important question in metabolomics today, and this study is an attempt to directly look at this problem. There are many clues that something else is going on; 1) we often see the same compound more than once, 2) there are sometimes visible patterns of peaks in a 2-D overview, and 3) too many unknowns have masses that do not easily translate to molecular formula. With regard to 1), we know this is in defiance of the principles of chromatography and is therefore dismissed as an artifact of the "look-it-up-in-a-database" method of naming things that is the basis of most non-targeted platforms, but many of them are structures unique to their molecular formula. When we see 2) we know that chromatography isn't perfect so they are all real, when we see 3) we tend to dismiss it because it is to unreal to accept.

This study has adopted an experimental design in which 1) all molecules of biological origin may be distinguished from molecules not of biological origin (See Fig. 1), and 2) because it is IROA we always know the number of carbons and the formula of any molecule. It starts by creating an IROA standard mixture that has reasonable biochemical complexity, and in which every molecule is specifically isotopically labeled. If all biological peaks are labeled we know all unlabeled peaks are non-biological.

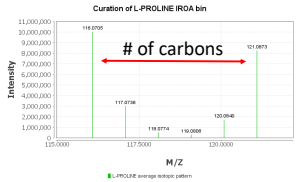


Fig. 1 – The IROA pattern for L-proline is that of a five carbon molecule with a C12 base mass at 116.0705. Note the spacing of peaks is exactly a neutron mass apart, and the peaks are symmetrical.

METHODS

Yeast was grown in either 5%- or 95%-U^{13}C IROA media in increasingly larger fermenters to yield yeast cells that were either purely 5%- or 95% -U^{13}C IROA-labeled. Equal aliquots of cells were pooled, extracted and subjected to LC-MS to yield near perfect mirror-image IROA patterns. This material, referred to as "Matrix", was analyzed in triplicate each day for 8 days. The experiment was repeated twice. Standardized peak picking was performed to determine how many peaks would be found in a "normal" metabolomic dataset. This was followed by a deep annotation of all of the IROA peaks (the only ones of actual biological origin) using ClusterFinder™ (CF) software with an additional module specifically designed for this study (now part of CF). This experiment was repeated on multiple instruments.

RESULTS

A. A standard peak-picking process, using mzMine, found 3,200 peaks on average, in each of the "Matrix" files when analyzed with a minimum peak height considered to be 100,000. The ClusterFinder program at the same setting (minimum = 100,000) finds only 781 IROA peaks. Since all of the biological peaks will yield an IROA pattern, of the 3200 peak picked peaks larger than 100,000, 2,419 (74%) are not of biological origin and are therefore non-biological artifacts.

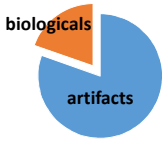


Fig. 2 – Peak-picking yielded peaks of biological (Orange), and non-biological (Blue) origin. In this study, only 24% of the peaks found came from the Matrix sample being analyzed. The rest were natural abundance and artifacts.

B. Within the biological peaks there are still a lot of unknown peaks. To see if these could be in-source, or instrument-induced fragmentation we created a software package that allowed us to select any peak and find all peaks that were highly correlated (as fragments, adducts, isotopomers, satellites, or polymeric derivatives will necessarily be). For any selected peak, the software found all correlated peaks and then created a matrixed table in which the degree of correlation is observed below the diagonal, and the mass differences above the diagonal (see Fig. 3A). From this table and the IROA-derived data it was generally possible to determine the exact fragmentation sequence (see Fig. 3B), and once it became obvious that there was a tremendous amount of post-LC fragmentation it became easy to recognize this as columns of peaks in a 2-D representation of the peaks (see Fig. 3C).

Fig. 3 – The correlation table exhibits the mass difference between any two correlated peaks above the diagonal. The fragmentation sequence may be read from the connections seen here.

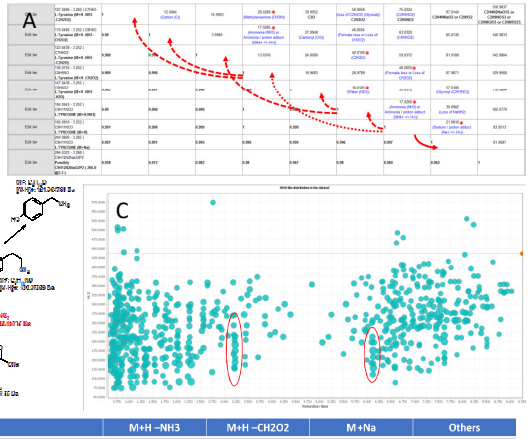


Fig. 4 – The positive-mode fragmentation of a few of the amino acids illustrate how many fragments, adducts, or polymerics may be seen for some molecules. Those peaks in this table are all in addition to the normal M+H.

	M+H –NH3	M+H –CH2O2	M+Na	Others
L-arginine	✓			4
L-aspartate				
L-glutamic acid				1
L-glutamine				
L-histidine			✓	1
L-isoleucine		✓	✓	
L-lysine	✓	✓		3
L-methionine	✓	✓	✓	1
L-methionine[34S]	✓	✓	✓	1
L-phenylalanine	✓	✓	✓	5
L-proline		✓	✓	2
L-threonine				
L-tryptophan	✓	✓	✓	8
L-tyrosine	✓	✓	✓	4
L-valine		✓	✓	

THEORY AND DISCUSSION

It was easily determined that a large percentage of peaks found in a basic peak-picking routine were artifacts since all of the peaks in the biological sample would exhibit IROA patterning, and therefore, if a peak did not have IROA patterning it could not have come from the biological sample. In this experiment, the IROA patterning was also useful because it allows a chemical formula to be accurately determined for each peak. Of the IROA-patterned peaks, many compounds were identified by comparison to authentic standards from the MSMLS[3]. Compounds for which standards were not available were identified tentatively by reference to the ClusterFinder™ embedded databases, and formulae calculations were attempted for all others. Among the biologically derived peaks were 1) known authenticated, and tentatively known compounds, and 2) unknowns either with or without formulae. Many compounds occurred multiple times. These observations are typically reported for metabolic untargeted analysis, and is completely "normal" for a metabolomic dataset.

We have observed that:

1) Many of the compounds occurring multiple times are common fragments of named compounds. They make chemical sense when the fragmentation path is examined.
2) Most of the unknowns examined to date are either fragments, adducts or polymerics. The formulae calculated from their IROA parameters are generally correct. They are often "named by lookup" and are simply not recognized as a fragment, for instance, indoline will show up for all tryptophan related compounds but it is usually a fragment.
3) A very limited number of the IROA peaks found could not be assigned relative to a known peak.

These findings will have significant implications for the metabolomics community. While the work reported here was done on Thermo Orbi Q Exactive, the extensive fragmentation seen is likely to be observed in all of the mass spectrometers in use today.

REFERENCES

1. de Jong F, Beecher C, "Addressing the current bottlenecks of metabolomics: Isotopic Ratio Outlier Analysis (IROA®), an isotopic-labeling technique for accurate biochemical profiling", *Bioanalysis* 2012, 4(18), 2303-14.
2. Stupp GS, Clendinen CS, Ajredini R, Szewc MA, Garrett T, Menger RF, Yost RA, Beecher C, Edison AS. "Isotopic Ratio Outlier Analysis Global Metabolomics of Caenorhabditis elegans." *Analytical Chemistry* 2013 *85*(24), 11858-11865. doi: 10.1021/ac4025413.
3. Mass Spectrometry Metabolite Library of Standards.

SUMMARY

- At least 75% of the peaks found in a standard peak-pick, the usual method of choice in metabolomics, are found to be non-biological artifacts.
- Peaks of biological origins that are "unknowns" are most often found to be fragments, adducts, isotopomers, satellites, or polymerics of other known peaks.
- A small percentage of peaks remained unknown after this analysis.
- We believe this demonstrates that most modern mass spectrometers are fragmenting their input molecules at a rate generally not realized.

Acknowledgements
We are thankful for support of the Southeast Center for Integrated Metabolomics (SECIM). IROA-labeled materials were produced for IROA Technologies (under agreement) by Cambridge Isotope Laboratories (CIL).