# Using Isotopic Ratio Outlier Analysis (IROA™) and Non-negative Matrix Factorization (NMF) to sort mixtures in metabolomics analyses.

Chris Beecher[1,3] , Timothy J. Garrett[2] , Elizabeth S. Dhummakupt[1], Vanessa Y. Rubio[1]
[1] University of Florida, Gainesville, FL; [2] Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, FL
[3] IROA Technologies, Ann Arbor, MI

## Overview

This study investigates the use of an IROA-based workflow to generate a high-quality dataset, and the use of NMF to characterize mixtures and determine the chemical composition of the parts.

## Introduction

Mixtures are inevitable. In fact, most metabolomic samples are mixtures of one kind or another, e.g. most tissue is a collection of different cell-types, most biopsies include both “normal” and disease, blood and plasma are mixtures of bodily functions, a cell population may include cells at different growth states, and even a single cell may represent a mixture of physiological states. Thus, the data we generate, rather than representing a single state, almost always represents the sum of data for all of the underlying tissue or cell types. Most of the time we attempt to solve this problem by trying to acquire tissues as pure as possible, and do not deal with the inherent mixed nature, assuming it will “all come out in the wash”. This study attempts to take a direct approach, namely to factorize the experimental dataset in an attempt to determine what the original mixtures were.

Non-negative Matrix Factorization (NMF) is a factorization technique that analyzes the correlation and partial correlation structure of the entire dataset to determine these underlying factors. For NMF to be successful it is necessary to have a number of samples; each one of which represents a slightly different mixture. It does not need to know the percentages of the mixed ingredients, nor does it have to have a pure species, i.e. any samples that are “unmixed”. This technique has received a great deal of use in image analysis, and text analysis, but has been rarely used in biology [see [1,2] for an overview]. In 2006, Golub[3] pioneered the use of a NMF-based approach to create “Connectivity Maps” of gene expression data to great effect, but it has not yet received much use in metabolomics. In an effort to demonstrate the utility of NMF in metabolomics, this study created a number of artificial, and unknown to the researchers, mixtures of two meats, beef and pork. The question was “could NMF sort the resulting dataset and determine what a metabolomics signature for pure pork or pure beef would look like?”. If it can do this, then when presented with a collection of liver biopsies it can just as easily determine what a pure tumor metabolomics profile looks like even though the biopsies all represent mixtures of tumorous and normal cells.

IROA[4] is a collection of protocols that dramatically strengthens the identification and quantification of metabolites. The IROA-based workflow (see below) inserts a mixture of 400+ internal standard compounds into a sample in order to measure 400+ compounds more accurately and reproducibly than otherwise possible. The IROA-based workflow is essentially noise-free, and therefore could generate a dataset of sufficient quality to be sorted by NMF.

Here we introduce a novel approach based on the use of IROA to generate very clean complex targeted dataset devoid of artifacts and noise, and NMF to deconvolute mixtures of any kind into their pure component mixture. This is a general algorithm that may be successfully applied in most cases of such problems.

## NMF in Theory

NMF assumes that every dataset (matrix of compounds by samples) may be seen as the sum of k submatrices (A), and

(A) Experimental Dataset = NMF 1 + NMF 2 + . . . + noise

that every submatrix can be factorized into factors representing the participation of compounds and samples (B).

(B) Experimental Dataset = NMF 1 + NMF 2 + . . . + noise

Since all information in the original dataset is retained then one may examine the sample factor or the compound factor to determine their role in each submatrix, or one may regress the factors against the experimental design to determine their significance.

## IROA-based workflow

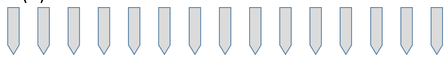
To each sample (C),

(C)



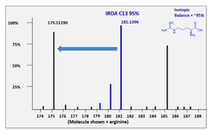
an IROA Internal Standard (IS) which contains over 400 compounds is added. (D)

(D)

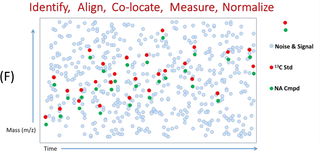


Each compound in the IROA Standard demonstrates a unique isotopic pattern in the MS (E), will co-elute with its natural abundance isotopologue, be easily identified, and will provide a means for high quality method for both identification and quantification. The IS has over 400 compounds so it allows metabolomics comparison of samples on over 400 phenotypic (metabolic) points (F).

(E)




(F) Identify, Align, Co-locate, Measure, Normalize




## Results

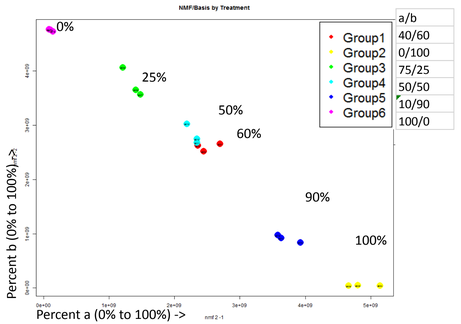



**Figure 1**: Percent of mixture as made up and as determined by NMF with a K=2.

## Experimental

Mixtures of meats, pork and beef, were made and prepared as blind samples. Metabolites were, homogenized in distilled water, and extracted using an 8:1:1 methanol:acetonitrile:acetone (MAA) protein precipitation method. An IROA-labeled yeast extract was added as a recovery Standard, and was the basis of the IROA-based workflow. The extracted samples were analyzed by reverse phase UHPLC-HRMS using an Ace Excel PFP-C 18 2.1x100 mm, 2.0 µm column. The UHPLC pump was a Dionex Ultimate 3000, and the mass spectrometer was a Thermo Fisher Q-Exactive Orbitrap. Solvent A was 0.1% formic acid in water, and solvent B was acetonitrile. The collected data were initially processed by Xcalibur Workstation software (version 3.0). Peak picking and compound identification was performed using IROA ClusterFinder™ software.

The resulting dataset was submitted to the IROA data portal for statistical analysis, which routinely includes NMF using the “NMF” R package. All values of k between k=2 and k=6 were examined.

## Conclusions

For this study, mixtures of similar meats, beef and pork, were made and prepared as blind samples. A normal IROA phenotypic analysis was performed with an IROA yeast extract, entirely made of uniform 95% $^{13}C$ labeled material, being introduced into each sample as an internal Standard. Using an IROA-labeled internal Standard, all peaks are easily identified according to their presence of characteristic M-1, M-2 etc. peaks and the exact mass and position of experimental natural abundance peaks are established relative to each standard peak. With the IROA Phenotypic approach the overall pattern of deviations from the Standard will define phenotype by difference from the Standard. The IROA Standard provided a mechanism to assure all peaks were of biological origin and assured accurate quantitation, i.e. false data point or random features which would confuse the interpretation were completely avoided. The resulting dataset was analyzed using NMF to deconvolute the nature of the underlying mixtures. The technique resulted in closely identifying the percentages of each meat in each mixture as well as determining what a 100% pure sample of each of the unmixed meats would look like.

NMF was run with k values from 2 to 6. All of these k values achieved statistically significant fits, and all had at least one plot similar to that shown in Figure 1. The sample factors for the K=2 were determined to have a p value of $1.1X10^{-12}$, indicating a very tight fit. K=3 and k=4 also had good p-values, but were all in the range of $1X10^{-4}$ and $1X10^{-2}$. Since they all found plots similar to Figure 1, it is likely that they found some additional dimension that was actually present in the meat samples, i.e. grizzle, or fat.

## Future Directions

- The IROA workflow significantly improved data quality in this study. It will be refined and applied to additional experimental designs.
- Investigate other applications of the NMF algorithms.

## References

[1] Fogel *et al.* Bioinformatics (epub 2006) 23: 44-49 (2006).
[2] Lamb *et al.*, Science 313: 1929-1935 (2006).
[3] Fogel *et al.*, American Statistician 67:207-218 (2013).
[4] de Jong F and Beecher C, Bioanalysis 4: 2303-2314 (2012).

## Acknowledgments

- Southeast Center for Integrated Metabolomics Grant #U24 DK097209