U.S. FOOD & DRUG ADMINISTRATION

Using the R Programming Language to Determine Rarity Scores for High Resolution LC-MS Data From TDS Food Items to Find Possible Untargeted Contaminants

Robert A. Levine

Kansas City Laboratory, U.S. Food and Drug Administration, 10749 West 84th Terrace, Lenexa, KS 66214
Robert.Levine@fda.hhs.gov

Introduction

The U.S. Food and Drug Administration has used Compound Discoverer™ (CD) software to evaluate its usefulness to conduct non-targeted analysis (NTA) using raw data from high resolution LC-MS Q-Exactive™ analysis of TDS samples.

When QuEChERS extracts of multiple market baskets for a single TDS food item are analyzed by LC-MS, the processing by CD presents a table of compounds detected having rows with m/z-RT pairs and peak areas for each market basket.

This can potentially provide a database of compounds normally found in food items due to the wide coverage of TDS sampling that encompasses a large geographical range as well as seasonal variations.

One use for this database could be to prioritize compounds for additional analysis and identification by selecting those with the widest range of variability of peak area and lowest detection frequency as rarity might be characteristic of unexpected contaminants not found by routine targeted methods.

Recently, a paper by Krauss (1) has proposed that site-specific compounds in bodies of water can be found by NTA using a rarity score (RS). This is not intended to replace targeted analysis, nor does it mean commonly found chemicals can always be ignored.

Objectives

To implement the rarity score calculation in Compound Discoverer, it was necessary to write a script in the R programming language that could be executed by the Scripting Node as the last step of post-processing.

The question to be answered here is whether this approach is feasible for prioritizing compounds from a pilot study of the NTA approach for TDS.

Materials and Methods

See details in reference (2).

<u>Samples:</u> Red apples from the 2018 TDS regional market baskets, collected monthly and composited from 3 cities, to give 12 samples representing 36 sites.

<u>QuEChERS:</u> 5g sample extracted with 25 mL acetonitrile, then after centrifuging phase separation with 6 g $MgSO_4$ and 1.5 g NaCl, aliquot of acetonitrile diluted 50:50 with water and filtered into autosampler vial.

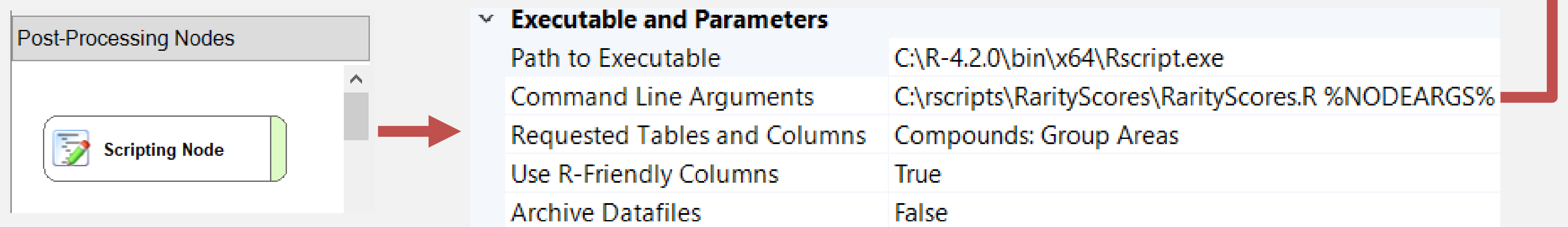
<u>LC generic method:</u> C18 column, Solvent A: 0.1% formic acid in water; Solvent B: 0.1% formic acid in acetonitrile, gradient: 2% B to 98% B in 35 min

<u>Mass Spectrometry:</u> Thermo Fisher Scientific Q-Exactive, ESI+ and ESI- modes, Full MS spectra were acquired from m/z 100-1500, resolution: 140,000 at m/z 200. Each sample and a blank was run in quadruplicate.

<u>Data analysis:</u> CD 3.3 SP1 was used with the workflow Untargeted Metabolomics, with minimum peak intensity lowered to 2000 from 10000. The 4 replicate areas for each sample were averaged into a group area. The data was filtered to select compounds with at least one group area 5x times bigger than the blank. Degenerate compounds were found by visual inspection and reduced to a single m/z-RT. Peaks consisting of noise were eliminated. Peak ratings were not used.

Scripting Node

The Scripting Node supports any programming language that can run code from the command line; see reference(3). The requested tables and columns are exported to a temporary folder as tab-separated values (TSV) text files.



Executable and Parameters

Path to Executable	C:\R-4.2.0\bin\x64\Rscript.exe
Command Line Arguments	C:\rscripts\RarityScores\RarityScores.R %NODEARGS%
Requested Tables and Columns	Compounds: Group Areas
Use R-Friendly Columns	True
Archive Datafiles	False

Rarity Score

Rarity Score of each compound according to Krauss et al. (1).

$$RS = \frac{\textit{maximum intensity}}{\textit{median intensity}} \times \frac{\textit{total number of samples}}{\textit{number of positive detects}}$$

It is necessary to have a column of RS values added to CD to perform filtering and sorting within an NTA workflow, which motivates use of a scripting language.

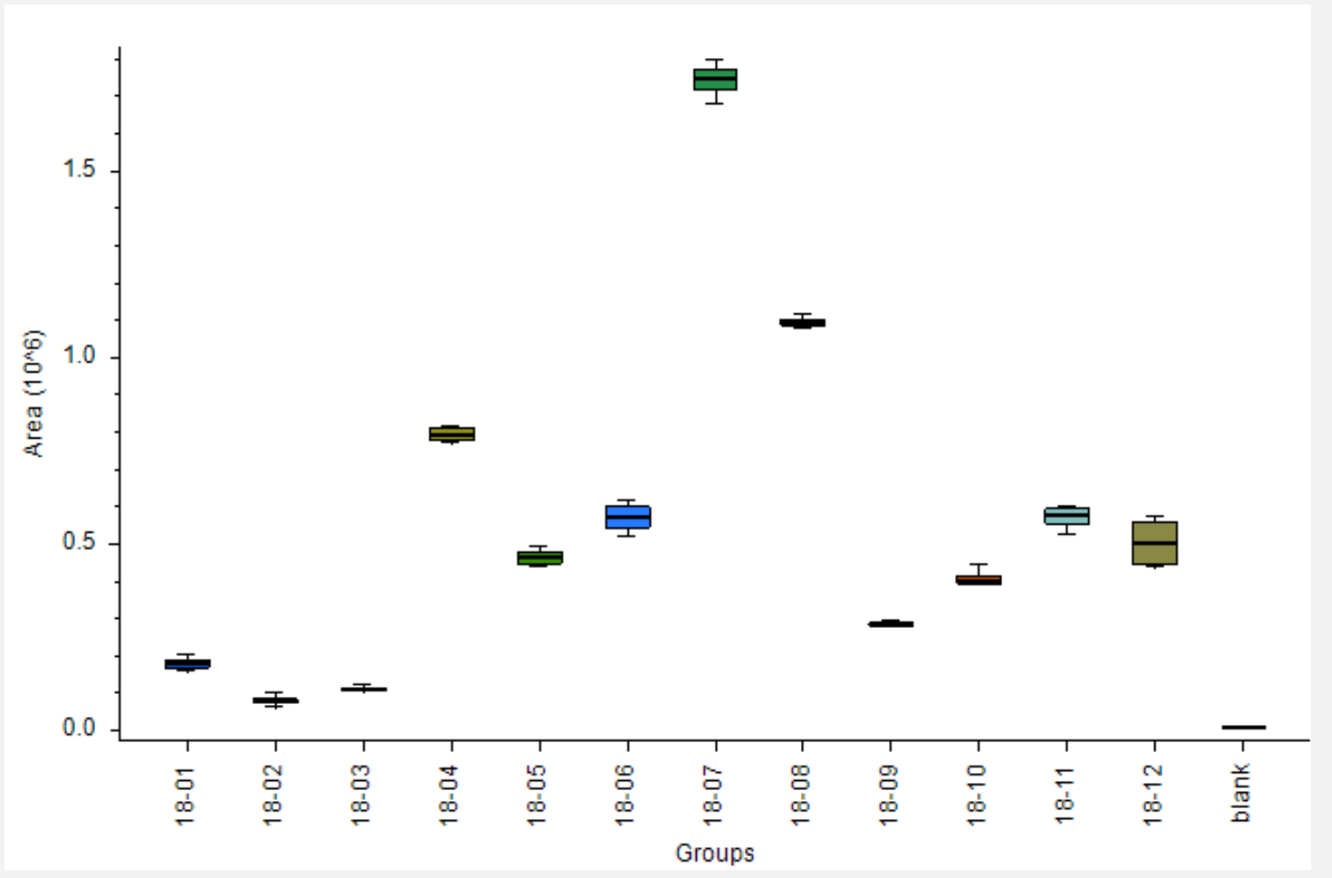
Results and Discussion

The value of RS depends on the method used for gap filling which effects the median intensity; in this case the CD default Fill Gaps node was used. The selection of a minimum detect level effects the number of positive detects; in this case it was observed that peak areas of 35000 could be reliably found and integrated, although at that level there were false positives consisting of noise that had to be excluded by visual inspection.

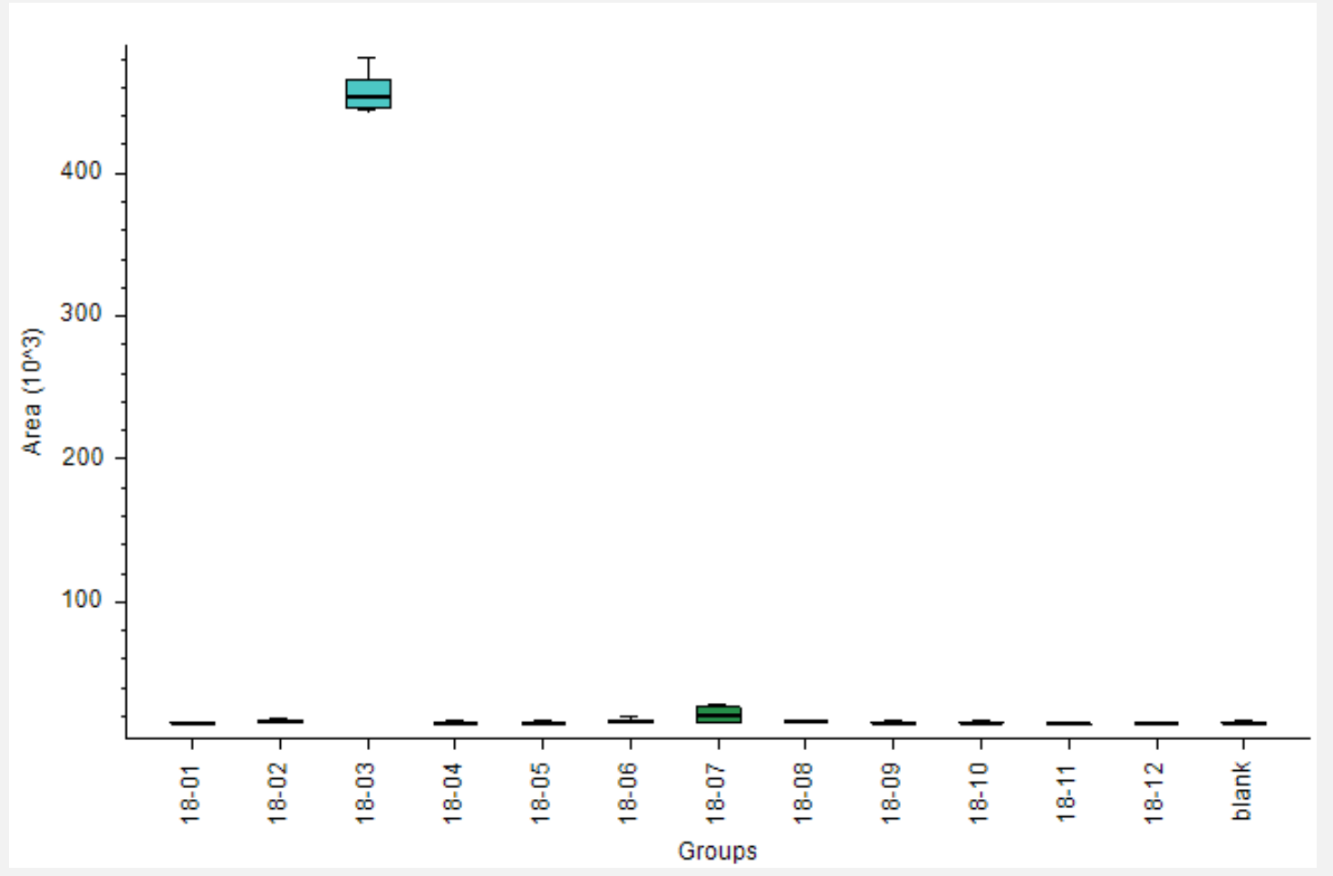
After blank subtraction of the ESI+ data, the number of compounds found was reduced from about 46000 to 3800, then filtering by RS>40 reduced the number to 500. Therefore, the RS reduced the number of compounds that needed to be visually inspected for degeneracy and noise, making the task manageable. After those additional cleaning steps only 270 compounds remained. The RS threshold can be made higher or lower if needed for a particular situation.

An example of low RS in spite of high variability is tryptophan in red apples in the 12 market baskets because of the high number of detects. A more important example is the high RS for the incurred residue of fluxapyroxad, in spite of the lower maximum peak area, which ranks in the top 10 of compounds passing the filters, indicating that RS has good sensitivity.

Tryptophan: RS = 4
max peak area: 1,800,000



Fluxapyroxad: RS = 348
max peak area: 454,000



R Script

The first part of the script (not shown) reads the command line arguments to get the full path to the JSON file that has metadata for the accompanying TSV file. The JSON file is read and the full path to the TSV file located and is assigned to `datafile`.

<u>R code to read input data from TSV file, perform calculations and save results:</u>

```
# create data frame from TSV file
    CD.input <- read.table(datafile, header=TRUE, check.names = FALSE)
# create data frame with columns of Group Areas only
# exclude “Compound ID” (col 1) and columns with “blank” in the name
    M <- CD.input[-c(1, grep("blank", colnames(CD.input), ignore.case = TRUE))]
# equation values and RS, functions “apply” and “rowSums” repeat for every row
    Mx <- apply(M, 1, max)     # max group area
    Md <- apply(M, 1, median)  # median group area
    Ns <- apply(M, 1, length)  # total samples
    Nd <- rowSums(M >= 35000)  # number of group areas >= 35000 detect limit
    RS <- Mx / Md * Ns / Nd
# in case number of detected peaks is 0 then set the rarity score to zero
    RS[!is.finite(RS)] <- 0
# create output file by extracting “Compound ID” column from input and bind RS
    data.output <- cbind(CD.input[1], RS)
# Write data output to temporary folder where CD will find it
    resultout <- gsub(".txt", ".out.txt", datafile)
    write.table(data.output, file = resultout, sep='\t', row.names = FALSE)
```

The remainder of the script writes a response JSON file that CD reads after the script ends, which provides metadata and path for `data.output` where CD finds the new RS column. (Copy of full script available from the author.)

Conclusions

An obstacle to utilizing high-resolution mass spectrometry for non-targeted analysis as part of TDS is the challenge to prioritize the huge number of signals, especially when trying to attain greater sensitivity by use of a low minimum peak threshold. Some proposed solutions are not built into the Compound Discoverer software used in many labs, so the ability to program with a scripting language is increasingly necessary.

Assuming that the rarity of detected signals can be used to prioritize ones for the difficult task of identification and confirmation is somewhat arbitrary and will lead to some false negatives, but it does eliminate natural products existing in all or most TDS food samples to enable focusing on chemicals that are not natural.

References

(1) Krauss, M., Hug, C., Bloch, R., Schulze, T. and Brack, W., 2019. Prioritising site-specific micropollutants in surface water from LC-HRMS non-target screening data using a rarity score. Environmental Sciences Europe, 31(1), pp.1-12.

(2) Bakota, E.L. and Levine, R.A., 2020. Untargeted Screening in a Case Control Study Using Apples as a Matrix. Journal of Agricultural and Food Chemistry, 68(37), pp.10232-10246.

(3) https://mycompounddiscoverer.com/resources/resources-scripting-node/

Disclaimer

The information in these materials is not a formal dissemination of information by FDA and does not represent agency position or policy.