Unraveling the chemical nature of biofluids by using reference materials, ARUS libraries and the hybrid search. Do it yourself

Yamil Simón-Manso (yamil.simon@nist.gov), Xinjian Yan (xinjian.yan@nist.gov), Kelly H. Telu (kelly.telu@nist.gov), Yuri Mirokhin (yuri.mirokhin@nist.gov), Yuxue Liang (yuxue.liang@nist.gov), and Stephen E. Stein (stephen.stein@nist.gov)

Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland

Keywords: ARUS, Hybrid search, SRM, Metabolomics

Introduction

Untargeted metabolomics approaches for characterizing and differentiating biofluid samples using liquid chromatography-mass spectrometry (LC-MS) are affected by the diversity of sampling targets, analytical methods, and instrumentation. Significant ambiguity is introduced in the analysis due to a variety of factors: lack of standards, matrix effects, retention times (RT), ionization efficiencies, in-source (IS) chemistry, etc. Although the major bottleneck is attributed to the large fraction of ions that remains unidentified [1].

In this presentation, reference materials, libraries of annotated recurrent unidentified spectra, (ARUS) [2] and recent developments of the hybrid search (HS) [3] are combined in a workflow to address these issues – for identifying as many components as possible in any material amenable to LC-MS analysis.

Methods

Tandem mass spectra from thousands LC-MS runs of biofluids, including human plasma and urine, and cell extracts are recorded on a Fusion-Lumos Orbitrap using different types of activation and charge states. Most data is acquired in DDA mode at seven different collision energies. Software is used to evaluate the quality of the LC-MS runs. ARUS libraries are built using in-house software for selecting, clustering, and annotating unidentified spectra. Hybrid-searching, which merges direct peak matching with neutral-loss matching, is used for annotating spectra. A variety of instruments, chromatography methods, sample type, and spiking are employed [4]. Post-identification filtering is performed by combining IS library searching, RT, MS1-info and prior-probability to ensure the quality of the identifications. The suite of programs is integrated in a simple workflow.

Results

A suite-of-programs and ARUS libraries whose uses range from quality control of LC-MS data, small-molecule identification, ID validation, to data comparison across experiments and instruments are integrated in a workflow that is used to characterize several biomaterials.

The workflow is divided in five steps: 1.- Data acquisition, optimization and preprocessing; 2.- Spectral similarity, clustering, and consensus spectrum building; 3.- Library building; 4.- Spectrum annotation (Chemical class determination); 5.- Validation.

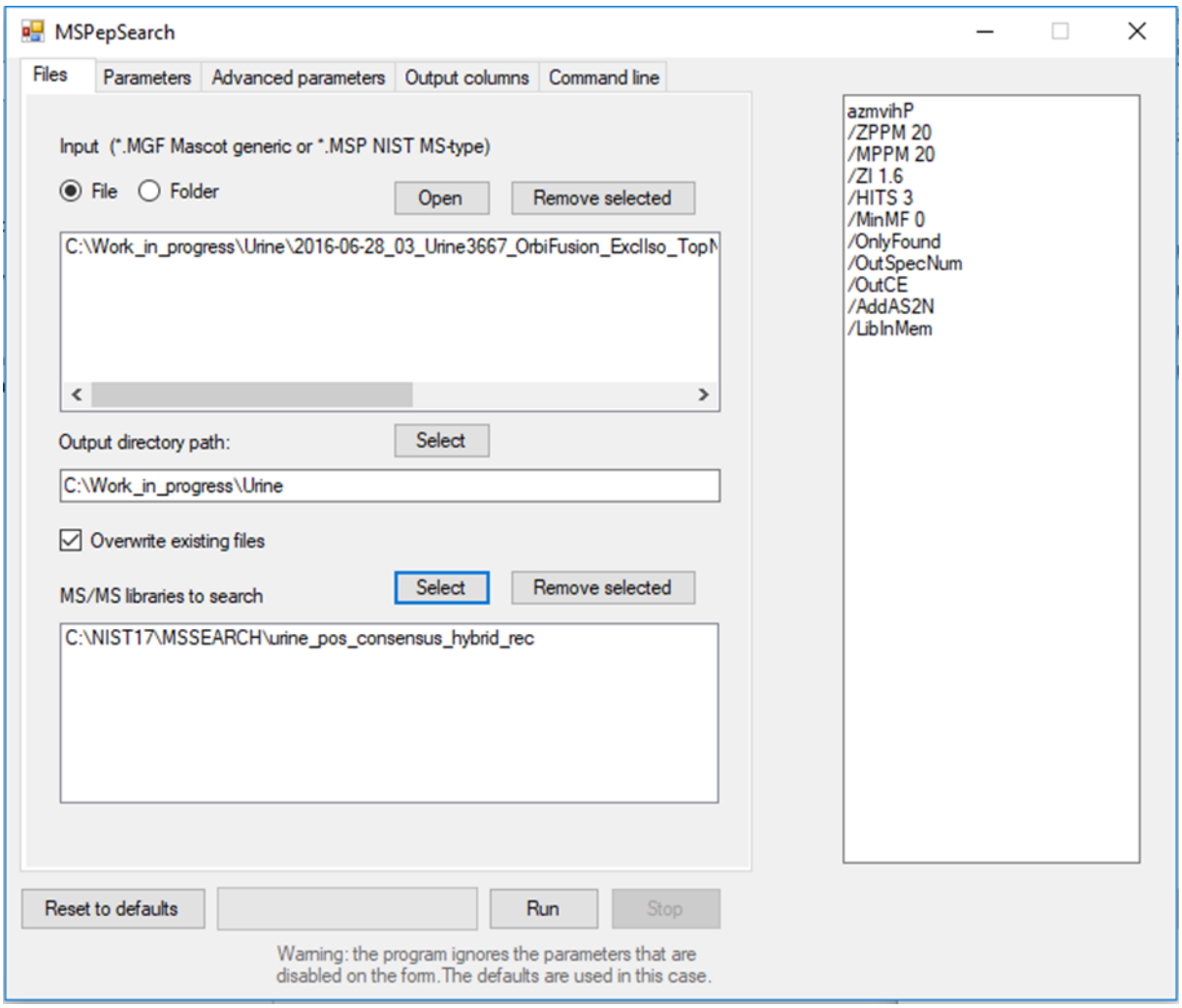
Data from samples of multiple SRMs are used as training sets to evaluate the performance of the suite of programs (see table 1).

Table 1. Biofluids and cell extracts used in this study.

Plasma/Serum (pooled)	
1950	Metabolites in Frozen Human Plasma
909c	Frozen Human Serum
967a	Creatinine in Frozen Human Serum
968e	Fat-Soluble Vitamins, Carotenoids, and Cholesterol in Human Serum
971	Hormones in Frozen Human Serum
972a	Vitamin D Metabolites in Frozen Human Serum
1951c	Lipids in Frozen Human Serum
3950	Vitamin B6 in Frozen Human Serum
956d	Electrolytes in Frozen Human Serum
Urine (pooled)	
3667	Creatinine in Frozen Human Urine
3671	Nicotine Metabolites in Human Urine (Frozen, 3 levels)
3672	Organic Contaminants in Smokers' Urine (Frozen)
3673	Organic Contaminants in Non-Smokers
CHO cells	
Culture media	ProCHO5
Cytosol	Extract from CHO cells
E. *coli*	
Cytosol	Extract from E. *coli* K12 cells

Tandem mass spectra are collected from multiple runs for each material and searched against the NIST tandem mass spectral library of pure compounds for completing a hit-list. Then, recurrent unidentified spectra are extracted, clustered, and ARUS libraries are built. The HS-technique is used for annotating consensus spectra. HS does not require precursor mass matching but increases scores for similar compounds by matching shifted product ion peaks and helps to identify chemical classes. The HS technique is been used successfully before in studying complex biofluids [2, 3, 5]. Most data processing tasks can be performed using freely available in-house Graphical Interface tools (Figure 2) developed by the NIST Mass Spectrometry Data Center.

Figure 2. Screen snapshot of the Graphical User Interface (GUI) tool to run MSPepSearch.



Then, information beyond spectral matching such as accurate m/z, retention times, IS searches, etc. is used to clean and validate the identifications.

Finally, how certified materials provide well-defined reference values and uncertainties in the context of metabolomics is discussed. The workflow is applicable to any biological fluid or extract. Libraries and software are developed as online resources for continuing community peer review and improvement. We will explain how users can manipulate raw data in ASCII text format to follow the workflow. Retention time collection and related issues will be briefly discussed.

Conclusions

An integrated suite of programs is used in a simple workflow to identify almost all ionizable chemicals present in a biofluid, thus allowing for effective data comparisons across experiments and instruments.

Novel Aspect

An integrated suite-of-programs and libraries for identifying most components in biofluids.

References

1.- Wishart. Bioanalysis **2011**, 3 (15), 1769.

2.- Simón-Manso *et al*. *Anal Chem* **2019**, 91 (18), 12021.

3.- Cooper *et al*. *Anal. Chem* **2019**, 91 (21), 13924.

4.- Bearden *et al*. *Metabolites* **2019**, 9(11):270.

5.- Blaženović *et al*. *Anal. Chem.* **2019**, 91 (3), 2155.

Contact: ysimon@nist.gov