Publicly Available Metabolomics Data Alignment of Human Urine Measured by RPLC-MS

Hani Habra, Yamil Simon and Tytus Mak
Mass Spectrometry Data Center, National Institute of Standards and Technology, Gaithersburg, MD

NIST National Institute of Standards and Technology U.S. Department of Commerce

Mass Spectrometry Data Center

OVERVIEW

The lack of standardization in LC-HRMS protocols and instrumentation for metabolite profiling results in significant disparities in metabolome detection coverage, chromatographic retention times, and signal abundances across institutions. The goal of the Publicly Available Metabolomics Data Alignment (PAMDA) is to determine the comprehensive set of reproducible mass spectral features in experimental data sets deposited in public repositories pertaining to a commonly studied specimen (human urine). PAMDA leverages *metabCombiner*, which uses m/z, retention time warping, and relative abundance similarities to match features representing the same analytes. In addition, MS/MS spectral information is incorporated through direct inter-experimental similarity scoring and comparison of identities obtained through NIST20 Tandem MS Library searches. Together, these efforts aim to achieve data harmonization for metabolomics measurements of a complex biospecimen.

Chromatographic Retention Times

Experiment	#1	#2	#3	#4	#5
Creatinine	1.06	0.68	0.79	0.81	0.8
L-Kynurenine	2.63	1.72	2.88	1.54	2.3
L-Trytophan	4.74	3.22	5.36	2.54	4.07
Azelaic Acid	9.41	8.22	18.55		4.67
2-Acetamidooctanoic acid	11.38	10.44	20.78	8.11	5.07

Table 1. Measured retention times for 5 urinary compounds in five different experiments, showcasing the variability of liquid chromatography methods.

INTRODUCTION

Untargeted metabolomics is widely applied for the detection and quantitation of small molecules in biological specimens. Liquid Chromatography - Mass Spectrometry (LC-MS) is the most widely used method for metabolomics due to its sensitivity, versatility, and throughput. Despite these advantages, LC-MS has limited reproducibility within and between instruments, experiments, and laboratories. This problem is compounded by the lack of standardized LC-MS methods, with differences in analytical factors among laboratories, such as stationary phases, LC column dimensions, mobile phase solvents, gradients, sample processing, and mass spectrometry instrumentation. The most significant barrier to harmonization of metabolomics data is the alignment of non-identically acquired LC-MS features, most of which are not easily identified.

METHODS

Metabolomics study data from public data repositories (Metabolomics Workbench, Metabolights) if they meet the following criteria: 1) Human urine samples are analyzed; 2) Reversed Phase Liquid Chromatography (RPLC) utilized using H_2O and Acetonitrile mobile phases; 3) Positive ionization mode; 4) One data set per institution. Priority is given to data sets with available MS/MS spectra.

Experiment ID	Repository	Samples	LC Time	Mass Spectrometer	Full Scan MS?	MS/MS ?
MTBLS469	Metabolights	48	20 min	FT-IT (Thermo Fisher)	yes	yes
MTBLS1465	Metabolights	18	15 min	Q-Exactive (Thermo Fisher)	yes	yes
MTBLS1572	Metabolights	3	30 min	QTOF (Bruker)	yes	yes
ST000291	Workbench	15	16 min	Q-Exactive (Thermo Fisher)	yes	no
ST001874	Workbench	3	30 min	QTOF (Agilent)	no	yes
SRM	NIST	3	30 min	Lumos Orbitrap	no	yes

Table 2. Experimental details for selected urinary metabolomics studies for alignment in PAMDA

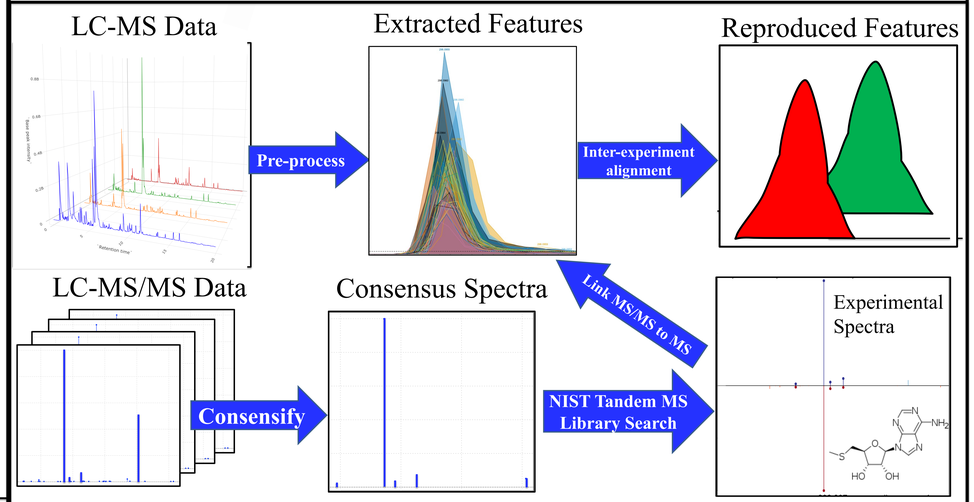


Figure 1. Workflow for individual experiment data processessing. LC-MS scans are conventionally pre-processed using MZmine3 to obtain feature lists. Meanwhile, MS/MS spectra (if provided) are merged into consensus spectra, compared to NIST Tandem MS library spectra, and mapped to MS1 features. The lists are then aligned to other experiments.

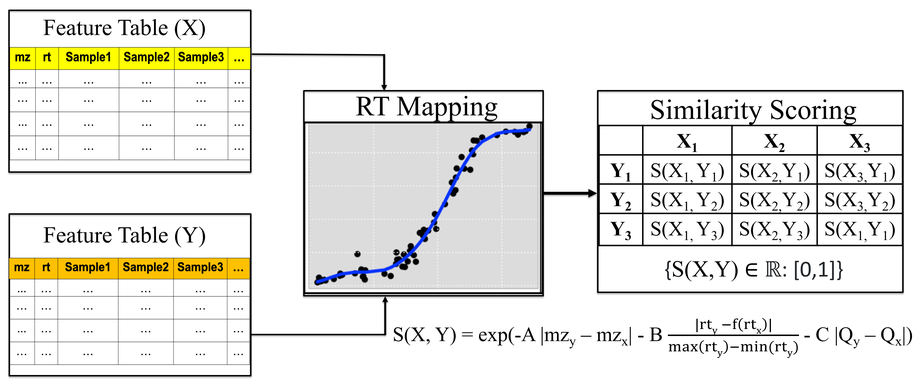


Figure 2. Inter-Experiment Alignment Workflow. Features from a pair of pre-processed experimental tables are grouped by similar m/z, followed by mapping of chromatographic retention times (RT) from one data set (X) to the other (Y), followed by similarity scoring based on m/z, RT, and Q (ranked relative abundance) distances.

RESULTS

For demonstration purposes, we illustrate the alignment of two LC-MS urinary metabolomics studies, MTBLS469 and MTBLS1572. LC-MS feature counts and LC-MS/MS spectral totals are listed in Table 3.

	MTBLS469	MTBLS1572
Total LC-MS features	34181	8308
Total LC-MS features (Post-filter)	14138	6332
Consensus MS/MS Spectra	15051	6075
Total features with 1+ Mapped MS/MS Spectra	2457	1963
Estimated Intersection Size	3479	

Table 3. Summary of MTBLS469 – MTBLS1572 inter-experimental alignment

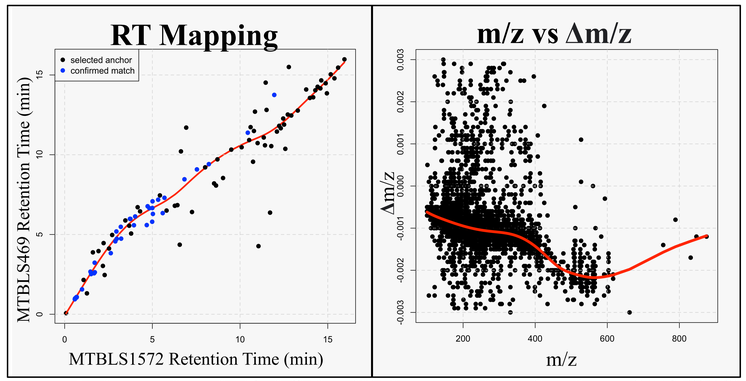
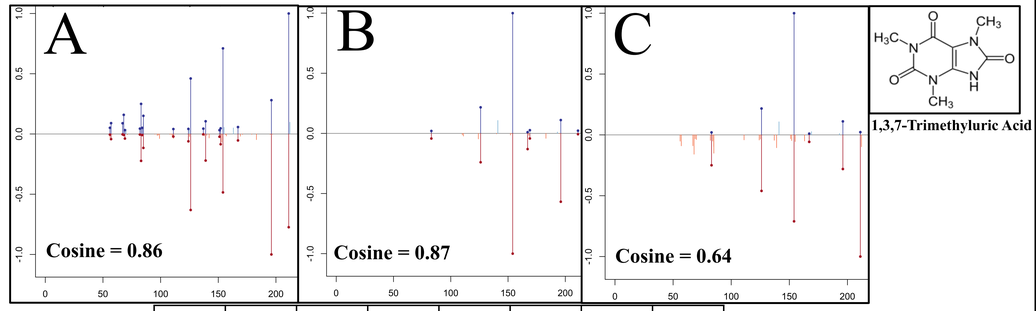


Figure 4. RT mapping and m/z correction using Basis Splines (A) The retention times of one data set (MTBLS1572) are mapped to the other (MTBLS469), using selected "anchor" points; identity matched features are incorporated for improved performance (B) The data sets exhibit a systematic bias in m/z distances, modeled with all points within <RT, m/z, Q> constraints



mzx	mzy	rtx	rty	rtProj	Qx	Qy	Score
211.0827	211.0817	2.95	5.19	4.8327	0.9793	0.9196	0.7826

Figure 5. Feature Match Feature X (m/z = 211.0827, RT = 2.95) from MTBLS1572 matched to Feature Y (m/z = 211.0817, RT = 5.19) from MTBLS469 using multiple lines of evidence: m/z, RT mapping, relative abundance similarity, EIC (not shown), and MS/MS match to **1,3,7-Trimethyluric Acid** in the NIST20 library. **A)** Feature X vs Library Match (20eV); **B)** Feature Y vs Library (IT-FT); **C)** Y vs X

CONCLUSIONS

Vast differences between LC-MS protocols are a major barrier to data inter-operability in the metabolomics field. Using study data from public repositories and disparate LC-MS alignment methods, PAMDA represents an effort to identify reproducible features in a commonly studied specimen across unrelated studies using RPLC-MS, a step towards data harmonization.

hani.habra@nist.gov