# Annotated Recurrent Unidentified Spectra (ARUS)

To assist in the compound identification problem, the NIST Mass Spectrometry Data Center has developed a novel type of mass spectral library, one that includes all recurrent unidentified mass spectra in a material. Unlike traditional spectral libraries, which consist of reference spectra of known compounds derived from neat standards, these libraries are derived from recurring spectra of unknown identity in the target material itself, where spectra are extracted, clustered, and where possible annotated prior to entry into a library. Building the library itself follows a similar methodological procedure to the one described for libraries of neat compounds, though with a different set of spectrum and measurement annotation.

In general, this type of library can be useful in many usual tasks of 'omics studies (i) answering where, how often, and in what conditions certain ions are observed, (ii) assigning class ID for compounds not in current tandem mass spectral libraries or not commercially available, (iii) connecting samples in an unambiguous way for control-case studies or interlaboratory comparisons (each molecular feature is represented by a spectrum in the library).

# **Getting Started**

To run the ARUS library, download the installation program for the NIST Search Software and the library from the NIST Website (https://chemdata.nist.gov/). A copy of the library must be present in the folder "MSSEARCH" before opening the browser.

Most datasets were generated using a Fusion Lumos Orbitrap. Eventually, data from other instruments such as Agilent and Waters QTOF was also collected. Although no systematic comparison has been made between data from different instruments, the library coverage for all instruments is similar. However, mass accuracy and ranges need to be adjusted accordingly in order to yield similar library scores for the same ions.

Figure 1 shows an example of a library spectrum that was annotated using the hybrid search technique:

NIST MS S	earch 2.4	- [IDs r	ange 500	-1000, 501	.01 spectra]	- 0
File Searc	h <u>V</u> iew	Tools	Options	Window	w Help	
🗈 🖻 🖗	s 😗 I		■ m/z	← ?		
				-		
👫 ID Number 🗸 1. IDs range 500-1000 🗸 🕅 🖫 🔍 🍸				× 1.1	l IDs range 500-1000 V 🔀 🤤 🍸	
#	Lib. ur	RI	Syn 3	DBs 0	Name Cluster 000500 delta mass.(Oxidation+H2O)* SI 0023( 100-	312.217
2	ur	-	0	0	Cluster_000501 NS=612 NR=86 Ch=1 [M+H]+ 20% P=	
3	ur	-	0	0	Cluster_000502_delta_mass.(Na[I]+H2O) NS=611 NR=55	
4	ur	-	5	0	1,4,7,10,13,16-Hexaoxacyclooctadecane/[M+H+NH3]+ De	
5	ur	-	5	0	Pentaethylene glycol/[M+H]+ Delta mass=m0000.0001 Sc	
6	-	-	5	0		
5	ur	-	5 5	-	Pentaethylene glycol/[M+H]+ Delta_mass=0000.0001 Sco	
8	ur	-	5 5	0 0	Pentaethylene glycol/[M+H]+ Delta_mass=0000.0001 Sco	
8	ur	-	5 5	0	7-Methyluric acid/[M+H]+ Delta_mass=m0000.9840 Score	
	ur	-	•	-	Pentaethylene glycol/[M+H]+ Delta_mass=0000.0002 Sco	253.1438
10	ur	-	0	0	Cluster_000509_delta_mass,(Na[I]+H2O) NS=607 NR=2! 85.0287 123.1171	
11	ur	-	5	0	Uric acid/[M+H]+ Delta_mass=0000.0001 Score=945 NCE 60.0810 .107.0857 151.1120	
12	ur	-	5	0	1,4,7,10,13,16-Hexaoxacyclooctadecane/[M+H-C4H8O2]	······································
13	ur	-	5	0	2-Pyridylacetic acid/[M+H-H2O]+ Delta_mass=0000.9841 60 80 100 120 140 160 180 200	220 240 260 280 300 3
14	ur	-	5	0	4-Methyl-2',4',6'-trimethoxychalcone/[M+H]+ Delta_mass=( (urine_hcd_pos_rec) Hexanoyl-L-camitine/[M+H]+ Delta_mass=0052.0314 Score	
15	ur	-	0	0	Cluster_000514_delta_mass.(Na[I]+H2O) NS=599 NR=1 Name: HexanoyL-carnitine/[M+H]+ Delta_mass=0052.0314 Score=840 NCE=20 N	NISTNO=1474668
16		-	3	0	Cluster_000515_delta_mass,(Na[I]+H2O)_SI_002369_0( Collision energy: 20%	
17		-		0	Hexanoy-L-carnitine/[M+H]+ Delta_mass=0052.0314 Sco Precursor m/z; 312.217	
18	ur	-	0	0	Cluster_000517_delta_mass,Na[i]* NS=597 NR=22 Ch=1 MW: NA Exact Mass: 311.2097 ID#: 516 DB: urine_hcd_pos_rec	
19	ur	-	0	0	Cluster_000518 NS=596 NR=84 Ch=1 [M+H]+ 20% P= Destantiation of the second sec	
20	ur	-	5	0	Pentaethylene giycov[M+H]+ Delta_mass=m0000.0001 Sc	
21	ur	-	5	0	N-Isopropylbenzenesulfonamide/[M+H-C3H9N]+ Delta_ma	
22	ur	-	0	0	Cluster_000521 NS=595 NR=67 Ch=1 [M+H]+ 20% P=	
23	ur	-	0	0	Cluster_000522_delta_mass,Na[i] NS=594 NR=16 Ch=1 I.Delta_mass=0052.0314 Hexanoyl-L-carnitine/[M+H]+ Score=840 NCE=20 NISTN	0-1474668
24	ur	-	5	0	Pentaethylene glycol/[M+H]+ Delta_mass=0000.0001 Scol 2.Cluster_000516 delta_mass.ris_SL_00051.0009 mz_253.144 / NL_000514 (otta_mass.ris_SL_000515, 0009 mz_253.144 / NL_000514 (otta_mass.ris_SL_000516) (otta_mass.ris_SL_00516) (otta_mass.ris_SL	
25	ur	-	0	0	Cluster_000524_delta_mass,(Na[I]+H2O) NS=592 NR=7; 3.HY_000694_0003_Cluster_000516_delta_mass,Tris_Si_000051_0009_mz_253	3.144_NL_000514_0003_loss_59.073
26	ur	-	0	0	Cluster_000525_delta_mass,(Na[i]+H2O) NS=591 NR=3( 4.NL_000514_0003_loss_59.073_Cluster_000516_delta_mass,Tris_SL_000051_000500000000	009_mz_253.144
27	ur	-	5	0	N-Acetyl-L-phenylalanine/[M+H]+ Delta_mass=0004.0062 5.SI_000051_0009_mz_253.144_Cluster_000516_delta_mass,Tris	
28	ur	-	0	0	Cluster_000527_delta_mass,Na[I] NS=591 NR=28 Ch=1	
29	ur	-	5	0	7-Methylimidazo[1,2-a]pyridine/[M+H]+ Delta_mass=0006.	
< 20			c	^	4 4 7 40 42 46 Upurana uklaastada ana (BA) U INU21: Da	
	ructures ,	/			PlotText ( Plot /	
/ maines // or						

Figure 1. Example of an annotated compound from the urine\_hcd\_pos\_rec library using the NIST Hybrid Search and the NIST Tandem Library.

This method is called a "hybrid" because it combines matching both ion m/z and mass losses from the precursor ion (neutral losses for singly charged ions). It is used to identify tandem mass spectra of compounds that differ from library compounds by a single chemical group. Peaks containing this group will, of course, be shifted by the mass of this group, as will the mass of the molecule that contains this group. This difference, termed DeltaMass, is used to shift the product ions in the library spectrum that contain the modification, thereby allowing library product ions that contain the unexpected modification to match the query spectrum. Peaks that match before or after shifting are treated equally, and if a single peak matches both before and after shifting, the abundance is partitioned. The hybrid search is implemented in the MS Search Program and can be used interactively or in a batch run.

Compounds that do not match a library spectrum using the NIST Hybrid Search are true unknown and named after the cluster number of the consensus spectrum (see Figure 2).

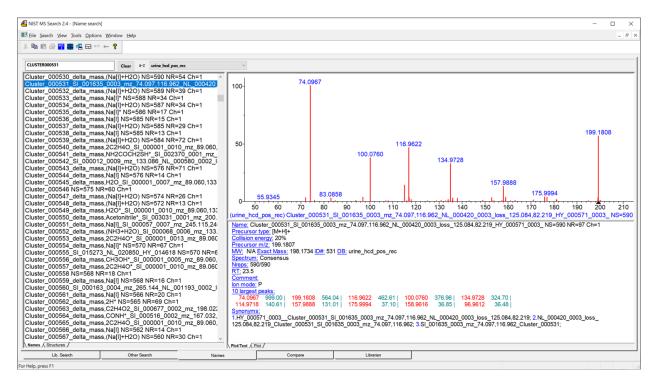


Figure 2. Example of an unknown compound from the urine\_hcd\_pos\_rec library.

The following nomenclature was included within compound names in the library: Name\_Adduct Type\_Score\_DeltaMasss\_Formula\_LibID, where Name, DeltaMass, Score and Formula are derived from the best hit in a hybrid search. LibID is a sequential number assigned to spectra in the archive. In case of not matching, known-unknown "Names" were simply given as cluster numbers (see Figure 2). Relevant information about the spectrum, such as name, retention time, the number of spectra used for building the consensus spectrum and others can be retrieved by using the Options tab > Comments Field Display (see Figure 3).

🕌 NIST MS Search 2.4 - [Name sear	ch]			- 🗆 X						
Eile <u>S</u> earch <u>V</u> iew <u>T</u> ools <u>O</u> ptic	ons <u>W</u> indow <u>H</u> elp			_ 8 ×						
	<b>?</b>									
11ALPHAHYDROXYPROG	Clear a-z	urine_hcd_pos_rec	~							
1-(1,4'-Bipiperidin-3-yl)-N,N-dim 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g	lucuronide lucuronide lucuronide	281.17	25							
11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g	llucul Display comment fi llucul	ield options	×							
11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g	Hydroxyprogesterone β-D-glucu Hydroxyprogesterone β-D-glucu Hydroxyprogesterone β-D-glucu Hydroxyprogesterone β-D-glucu									
11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g	lucu Spectrum lucu Nreps RT		<u>_</u>	457.2048						
11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g	α-Hydroxyprogesterone β-D-glucu       For example, if RI=132.1 is embedded in Comments, it can be displayed as a separate line in the text by entering the line:         α-Hydroxyprogesterone β-D-glucu       separate line in the text by entering the line:         α-Hydroxyprogesterone β-D-glucu       BI									
	11α-Hydroxyprogesterone β-D-glucu 11α-Hydroxyprogesterone β-D-glucu in the above field. Multiple lines will display multiple tag=value pairs in the order entered.									
11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g	lucui ⊠ Show tag=value o Iucui			57.2049 MS_mode="+" Pr						
11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g	lucui	Cancel	Help							
11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g	lucuronide 282.17 lucuronide 299.17 lucuronide 323.18	765 2.42 10 10 703 2.94 10 10 333 1.09 10 10								
11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g 11α-Hydroxyprogesterone β-D-g	lucuronide lucuronide lucuronide lucuronide lucuronide	048 120.07 10 10								
Names / Structures /	Other Search	tt_{Plot_/ Names	Compare	Librarian						

Figure 3. Comment field options.

#### Prerequisites

Downloaded the NIST Search Software and libraries from the NIST Website (<u>https://chemdata.nist.gov/</u>).

#### Installing

The installation of the search program is straightforward; however, a detailed manual can be found on the website accompanying the software.

### **Running the tests**

For testing the installation and library performance take the following these steps:

1.- Download and install the NIST20 browser.

2.- Download one or more raw datafiles. A human plasma data-dependent dataset can be found at <u>https://nvlpubs.nist.gov/nistpubs/jres/121/jres.121.022.pdf</u> and used for this purpose.

3.- Use ProteoWizard or proprietary software to extract the tandem mass spectra in MGF format.

4.- Import the MGF files into the NIST20 browser by using the FILE tab > Open.

5.- Adjust searching options (Figure 3) by using the Options tab > Library searching options.

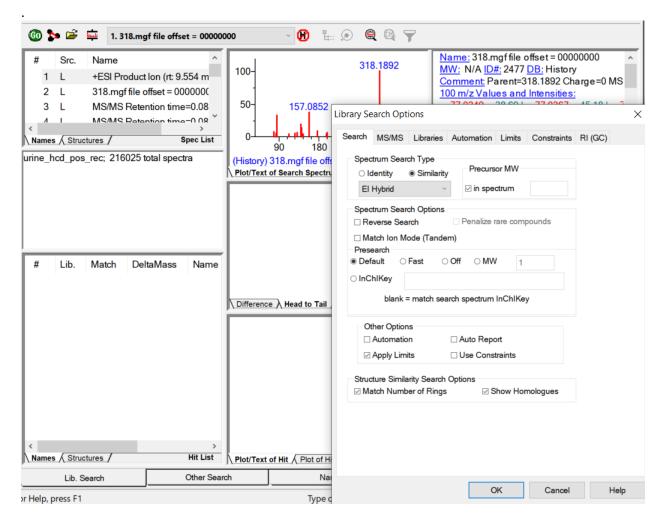


Figure 3. Library searching options

6.- For searching a single query spectrum, simply double-click on its name. For searching batches using MSPepSearch in a command line, consult the documentation of the NIST browser and the MSPepSearch program (help tab). A MS Excel file containing the library identifications in a single run of a plasma sample can be found with the ARUS libraries. The first cell of the second raw contains the command line used for the library searching.

# **Built With**

- NIST pipeline
- NISTms and MSpepSearch

# Contributions

All software and datasets were developed at the NIST Mass Spectrometry Data Center. For submitting requests or external contributions, please contact:

massspec@nist.gov

## Versioning

Versions 3.0, other versions are available upon request.

## **Authors**

NIST Mass Spectrometry Data Center.

#### License

<u>NIST Licensing Policy</u>. The library is offered without warranty and will be in continuous development in the future (<u>http://chmdatafnx.nist.gov/dokuwiki/doku.php?id=chemdata:arus</u>).