Next generation data exchange format for mass spectrometry

Anuj R. Shah, Matthew E. Monroe, Yan Shi, Brian LaMarche, Kevin Crowell, Gordon S. Slysz, Gordon A. Anderson and Richard D. Smith
Pacific Northwest National Laboratory, Richland, WA 99352

Pacific Northwest
NATIONAL LABORATORY

Overview

- The next generation proteomics instrumentation and analysis workflows call for a data format that facilitates analysis, maintenance, integration and exchange of both experimental and processed data.
- Recognizing the inefficiencies of the XML based formats, the proteomics community has entertained alternate strategies for data exchange, e.g., using a common application programming interface or a database-derived format.
- The ever increasing numbers of spectra produced from instruments call for a change in our existing data formats (XML scales poorly) and data management procedures.

Introduction

- XML formats are significantly redundant because of the tag based nature.
- Spectra are non-readable, and data processing pushes the current limits of both software and hardware.
- Data compression techniques are applied to make file sizes more manageable for routine operations with little success.
- Our approach is to create an over-arching data format based on standard database principles that offers multiple benefits over existing formats in terms of storage size, ease of processing, data retrieval times and extensibility.
- The format, termed YAFMS for "yet another format for mass spectrometry," also accommodates for updates from multiple analysis tools and can easily be extended for multi-dimensional separation systems.

CONTACT: Anuj R. Shah, Ph.D.
Biological Sciences Division, K8-98
Pacific Northwest National Laboratory
P.O. Box 999, Richland, WA 99352
E-mail: anuj.shah@pnl.gov

Methods

- Using SQLite, a relational database is created that facilitates large data management and supports efficient retrieval of spectra.
- Spectra are compressed and stored as binary large objects (blobs) within the database tables [1].
- Fast decompression algorithms facilitate retrieval of extracted ion chromatograms from the spectra in addition to supporting range queries. The compression / decompression time is less than the disk I/O time, resulting in a net savings in data access times.
- Additional tables can be created that store deisotoped results, clusters of deisotoped features as well as details of peptide identification and roll up to proteins. These extensions do not create any compatibility problems, another advantage of using a relational schema.

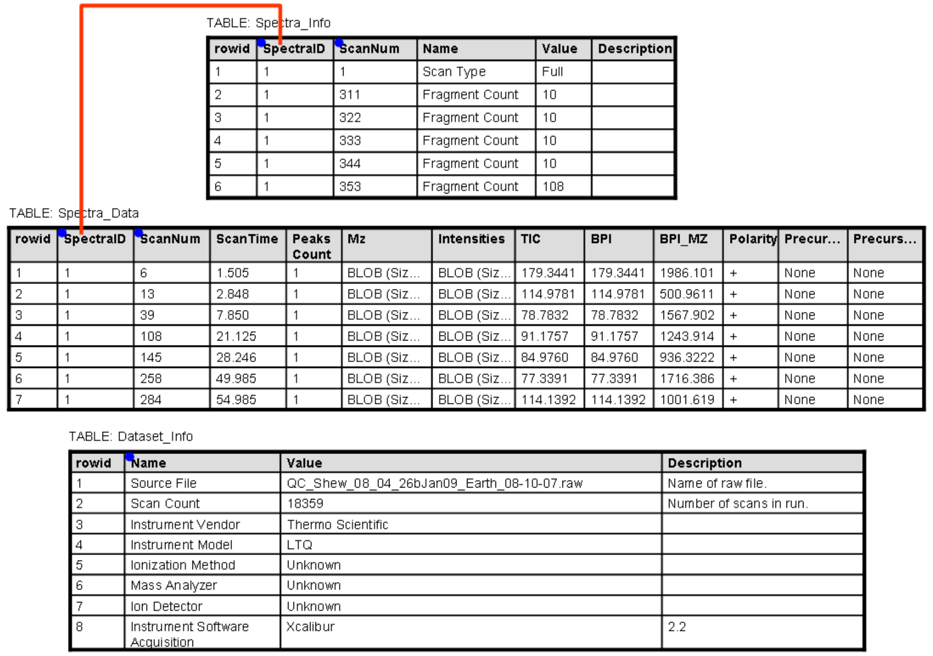

TABLE: Spectra_Info

rowid	SpectralD	ScanNum	Name	Value	Description
1	1	1	Scan Type	Full	
2	1	311	Fragment Count	10	
3	1	322	Fragment Count	10	
4	1	333	Fragment Count	10	
5	1	344	Fragment Count	10	
6	1	353	Fragment Count	108	

TABLE: Spectra_Data

rowid	SpectralD	ScanNum	ScanTime	Peaks Count	Mz	Intensities	TIC	BPI	BPI_MZ	Polarity	Precur...	Precurs...
1	1	6	1.505	1	BLOB (Siz...	BLOB (Siz...	179.3441	179.3441	1986.101	+	None	None
2	1	13	2.848	1	BLOB (Siz...	BLOB (Siz...	114.9781	114.9781	500.9611	+	None	None
3	1	39	7.850	1	BLOB (Siz...	BLOB (Siz...	78.7832	78.7832	1567.902	+	None	None
4	1	108	21.125	1	BLOB (Siz...	BLOB (Siz...	91.1757	91.1757	1243.914	+	None	None
5	1	145	28.246	1	BLOB (Siz...	BLOB (Siz...	84.9760	84.9760	936.3222	+	None	None
6	1	258	49.985	1	BLOB (Siz...	BLOB (Siz...	77.3391	77.3391	1716.386	+	None	None
7	1	284	54.985	1	BLOB (Siz...	BLOB (Siz...	114.1392	114.1392	1001.619	+	None	None

TABLE: Dataset_Info

rowid	Name	Value	Description
1	Source File	QC_Shew_08_04_26bJan09_Earth_08-10-07.raw	Name of raw file.
2	Scan Count	18359	Number of scans in run.
3	Instrument Vendor	Thermo Scientific	
4	Instrument Model	LTQ	
5	Ionization Method	Unknown	
6	Mass Analyzer	Unknown	
7	Ion Detector	Unknown	
8	Instrument Software Acquisition	Xcalibur	2.2

Figure 1. The Dataset_info table stores experimental setup details, the instrument used and other metadata. The Spectra_Info table stores sparse information related to spectra that are not stored as columns under the Spectra_Data table. The Spectra_Data table stores the mass/charge ratios and intensities as two identical length binary large objects (blobs) in the database, in conjunction with the total ion count, base peak intensity, base peak mass/charge, precursor mass/charge, scan time etc. The red line indicates the link between Spectra_Info and Spectra_Data table. The blue dot alongside a column indicates database indexes built on those columns for fast retrieval.

Performance Measures

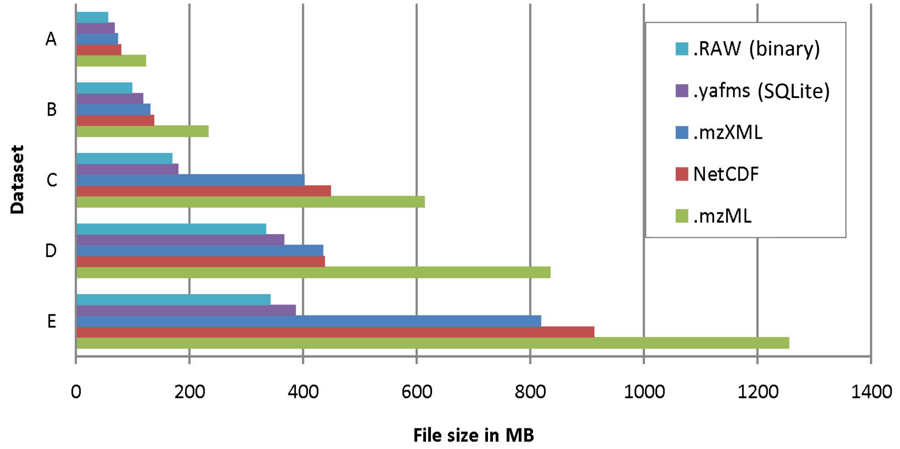

Figure 2. File size comparison for different data file formats. The YAFMS (SQLite) file sizes are comparable to the .RAW data formats and always significantly smaller than the mzXML and mzML data files (as much as 25-30% samples with dense spectra and more than 50% in cases of sparse spectral density). The NetCDF files were created using XCalibur 2.1.0 as distributed by Thermo.

Table 1. Data access rate comparison for different file formats

Dataset	Scans	Spectra Density (KB/scan)	YAFMS (ms)	RAW (ms)	mzXML (ms)
A	6878	24	0.106	6.432	0.570
B	18359	5.18	0.869	4.825	5.129
C	5548	60.26	2.891	40.062	41.873
D	13844	23.58	0.704	4.317	4.490
E	90766	0.60	2.390	48.130	49.591

A – Read times for five different datasets were calculated using the average of one-thousand randomly generated scan numbers and m/z ranges to the nearest microsecond using high resolution timing calls.

B – Spectra density is calculated by dividing the total file size by the total number of scans. The average spectra density shows that C and D are highly complex samples.

C, D, E – The Decon2LS [2] application was used to read all data formats. Decon2LS uses the RAMP mzXML parser to read mzXML files, ThermoFinnigan's proprietary libraries for .RAW files and our custom dynamic link library to read YAFMS files.

Conclusions

- We have presented and have in use a novel file format based on the principles of relational database management systems that affords multiple improvements over existing formats.
- The ability of our file format to represent multi-dimensional experimental configurations, such as ion mobility separations, makes it a good candidate for future generation mass spectrometry systems.
- Additional information about deisotoped features, clusters of features and peptides or proteins identified can be incorporated via new tables.
- Such a file format should add high value to raw data repositories and archives as it saves significant amounts of disk space.
- Software downloads from http://omics.pnl.gov/software/YAFMS.php

Acknowledgements

This work was funded by NIH National Center for Research Resources (RR18522 and ARRA administrative supplement) and the National Institute Of Allergy And Infectious Diseases (U54AI081680).

Samples were analyzed using capabilities developed under the U.S. Department of Energy Biological and Environmental Research (DOE/BER). Work was performed in the Environmental Molecular Science Laboratory, a DOE/BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is operated for the DOE by Battelle under contract DE-AC05-76RLO-1830.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of Allergy And Infectious Diseases or the National Institutes of Health.

References

1. N Beagley, C Scherrer, Y Shi, BH Clowers, WF Danielson, AR Shah. "Increasing the Efficiency of Data Storage and Analysis Using Indexed Compression," e-Science and Grid Computing, International Conference on, pp. 66-71, 2009 Fifth IEEE International Conference on e-Science (2009).
2. N Jaitly, N, A Mayampurath, K Littlefield, JN Adkins, GA Anderson, and RD Smith. "Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data," *BMC Bioinformatics,* ***10:***87 (2009).