An Optimized Data Compression Strategy for Efficient Storage and Analysis for Multidimensional Separations and Mass Spectrometry

Nathaniel Beagley, Chad Scherrer, Yan Shi, Brian H. Clowers, William F. Danielson, Anuj R. Shah, Anoop M. Mayampurath, Gordon A. Anderson, and Richard D. Smith
Pacific Northwest National Laboratory, Richland, WA

Pacific Northwest NATIONAL LABORATORY

Overview

- Storage and analysis of the large volumes of data sets produced by multidimensional separations coupled with mass spectrometry experiments is a growing problem.
- We present a novel compression methodology which reduces data storage size while increasing data storage speed and efficiency of data analysis.

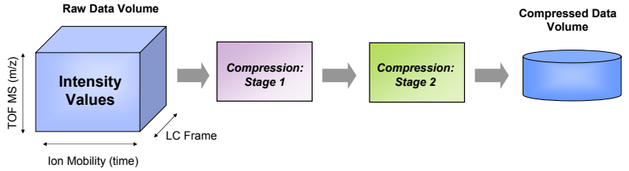
Introduction

High-throughput, multidimensional separations coupled with mass spectrometry (MS) play a vital role in scientific discovery; however, new developments in MS technology have increased data acquisition rates producing a significant rise in the quantities of data which must be stored, analyzed, and archived. The increased rate of data generation also poses a significant challenge to the real time analysis of data dependent modes of experimental operation.

We have developed a compression methodology to meet these challenges, which provides the benefits of size and speed from storage in compressed files, but also provides efficient data access in an analysis pipeline by allowing extraction of segments of uncompressed data from a file without first having to uncompress the entire file.

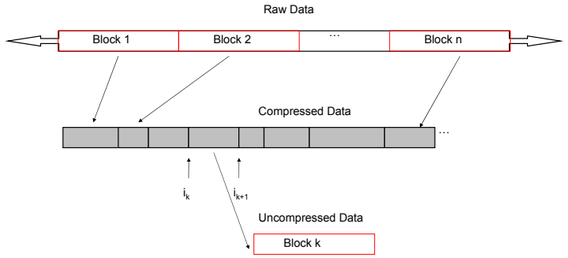
Algorithm

Hybrid, Multi-Pass Compression Algorithm



- Stage 1: Pre-processing Compression:
  Takes advantage of data characteristics (e.g., sparsity of intensity measurements)
- Stage 2: LZF Compression Algorithm:
  Achieves an optimal balance of ***compression speed*** and ***compression size***
- Compression is 100% lossless
- Optimized for integer intensity data values

Intelligent Compression / Data Organization



- Data is compressed in segments (optimally of a size needed for an analysis task).
- Compressed segments are stored in a file with start and stop locations recorded by indices
- A specific segment needed for analysis can then be extracted individually and uncompressed without having to first uncompress the entire file.

Metrics

Compression Size

	Raw Size	Current Storage Size	Compressed Size	Compression Percentage:
Low Density Data Frame	240 MB	8.2 MB	3.1 MB	37.8%
High Density Data Frame	240 MB	89.9 MB	35.3 MB	39.2%

Compression Speed

	Number of Spectra	Number of Bins	Compressed Size	Compression Time	Storage Time (to DB)	Total Time
Test 1	630	88,000	39.6 MB	0.58 s	0.27s	0.84 s
Test 2	630	88,000	39.7 MB	0.72 s	0.27s	0.98 s
Test 3	630	88,000	41.7 MB	0.72 s	0.27s	0.98 s
Test 4	630	88,000	40.0 MB	0.72 s	0.28 s	1.0 s

- Testing performed in an operational, real time, environment.
- Compression and storage of 600+ spectra to a database (> 200 MB of raw data) in ***~1 s!***

Extraction Speed

Benchmark: Extraction of a total of 20 MB (uncompressed) of separate data segments distributed throughout 9 GB of compressed data in ***1.5 s***.

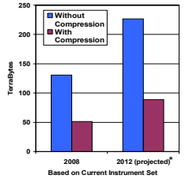
Application

This algorithm is an integral part of the Unified Ion Mobility Format (.UIMF) data format and the Next Generation Proteomics Measurement Platform developed by PNNL.

For additional information, see ASMS 2009 Poster #TP213: **A New Unified Data Format for Ion-Mobility-Time-of-Flight Mass Spectrometry,** Danielson, et. al.

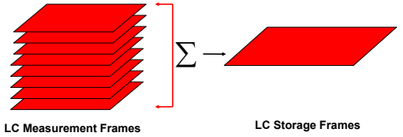
Compression Benefits

1) Reduction of Disk Storage Size



(*) Projected future size does not consider addition of new instruments or capture of higher resolution data from instrument advances.

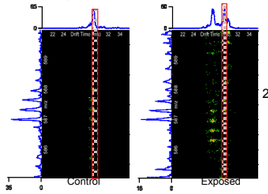
2) Storage of Increased Data Resolution



Due to disk write speed and PC data bus constraints in a real time environment, a number of LC Frames are aggregated before writing to disk thus losing information. Compression increases the quantity of data that can be stored and thus allows capture of greater data resolution.

3) Increased Analysis Efficiency

Peptide Biomarker Validation Experiment



400 Samples
X
250 LC Frames = 9 Terrabytes
X
90 MB

Analysis: Extract 5 Ion Mobility time columns (2 MB) from each LC frame

Without Intelligent Compression:
- Parse through 9 TB to extract 200 GB
- Extraction time > 24 h!

With Intelligent Compression:
- Direct Extraction of 200 GB
- **Extraction in ~20 min !!!**

Conclusions

By providing a 60% reduction in data storage size, we obtain the ability to:

- compress and store a 240 MB IMS-MS data frame in < 1 s
- extract an uncompressed segment of data from a compressed file without having to uncompress the entire file

As a result, the compression methodology provides significant benefits in capture, storage, and data analysis for mass spectrometry and many other multi-dimensional separation data volumes.

Acknowledgements

This work was funded by The Data Intensive Computing Initiative at Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is operated for the Department of Energy (DOE) by Battelle under contract DE-AC05-76RLO-1830.

Operational testing of the algorithm was performed in the Environmental Molecular Science Laboratory, a DOE/BER national scientific user facility at PNNL.

CONTACT: Nathaniel Beagley
Computational Mathematics, K7-90
Pacific Northwest National Laboratory
P.O. Box 999, Richland, WA 99352
E-mail: Nathaniel.Beagley@pnl.gov