COMPRESSION METHODS FOR MULTI-DIMENSIONAL MASS SPECTROMETRY DATA

Waters
THE SCIENCE OF WHAT'S POSSIBLE.™

Keith Richardson, Richard Denny, Darrell Williams, Stephen Platt*
Waters Corporation, Manchester UK

OVERVIEW

- **A number of methods for compression of continuum MS data are discussed.**
- **These methods are applied to proteomic datasets, and compression ratios of 6-10 are demonstrated.**
- **The compressed data are processed and searched as a simple test of data quality.**

INTRODUCTION

Increasing instrument sensitivity, resolution, dynamic range and the adoption of multi-dimensional separation techniques (such as chromatography and ion mobility coupled to mass spectrometry or LC-IMS-MS) all contribute to a continuing increase in the amount of data that is produced by modern mass spectrometers. This increase places severe demands on data transfer and storage requirements. We describe a series of compression algorithms that can be used in many combinations to reduce the size of the datasets produced whilst maintaining data fidelity. Application of these techniques at an early point in the data pipeline can allow acquisition of data that would otherwise be impossible owing to hardware constraints. Smaller datasets are obviously also more convenient for archiving, visualisation and post-acquisition processing

METHODS

100 ng of a cytosolic *E. coli* tryptic digest standard was injected using a nanoACQUITY system (Waters Corporation), equipped with a C_{18} 20 mm x 180 µm trap column and a C_{18} 15 cm x 75 µm analytical reversed phase column. The total gradient length was 120 minutes.

Data were acquired at a rate of 2 spectra per second using a Synapt G2-S HDMS mass spectrometer (Waters Corporation) operating at approximately 20,000 resolution (FWHM) over the m/z range 50-2000 Da/e. In both LC-MS and LC-IMS-MS experiments, the instrument was operated in a data-independent (MS^E) mode and alternate low and elevated collision energy data were collected.

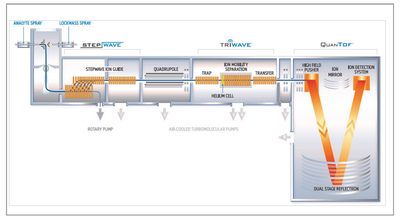

Figure 1. Schematic of the Synapt G2-S HDMS instrument used in this study.

Lossless Compression: Differentiation, packing and zipping

A mass spectrum can be regarded as a pair of lists of numbers (masses and intensities). In fact, due to the digital nature of most acquisition systems, in their raw form these numbers are usually integers and we shall refer to them as mass indices and intensities. Data points with zero intensity are usually discarded.

In a well-populated mass spectrum, consecutive mass indices often lie close together. In the limit of a fully populated spectrum, differences between consecutive mass indices are all unity. Similarly, in well-sampled data, intensities for consecutive points are often highly correlated because the data consist of a series of peaks.

These correlations can be exploited by storing differences between consecutive mass indices and intensities in records of reduced length. As the size of the records are reduced, difference values arise that cannot be stored using the allocated record size. These overflows are stored in separate tables of repair values utilizing larger record sizes (e.g. 4 bytes).

Figure 2 shows the total memory required to store all mass index and intensity differences and repairs arising in a 120 minute LC-MS^E proteomics experiment. For comparison, the memory required to store the original un-differenced values using the same scheme is also shown.

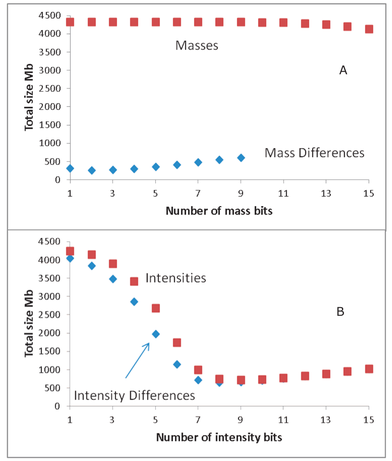

Figure 2. Total number of bits, including repairs, required to store mass and intensity information in the MS^E dataset. In this case an optimum record size of around 3 bits is observed for mass index differences (Fig.2A), while around 9 bits are required for intensity differences (Fig.2B). In this example, the difference between storing intensities and intensity differences is small.

As the number of bits allocated is reduced, the size of the repair tables increases, and these eventually dominate the overall size of the data. In this example, the optimum record size is under 3 bits for mass differences, and about 8 bits for intensity differences.

Finally, data that have been packed as described above can often be compressed further using general-purpose compression algorithms.

Adaptive Background Subtraction

Electrospray data often exhibit a background of broad peaks which repeat with a period of approximately 1Da. These may represent charged clusters of analyte and solvent molecules, but they do not generally yield useful information. However, the peak shape changes only slowly with m/z, and it is possible to use a moving window of the data (usually about 20 Da) to construct a model of the local background peak shape which can then be subtracted from the data. This algorithm, which has been described in more detail elsewhere [1], can remove interferences from low intensity peaks that would otherwise yield little or no information.

Another benefit of background subtraction is that it can substantially reduce the number of points with positive intensity in a dataset. Figure 3 shows a portion of a mass spectrum before and after adaptive background subtraction. In this small section of spectrum, the number of points with non-zero intensity is reduced by around 45%.

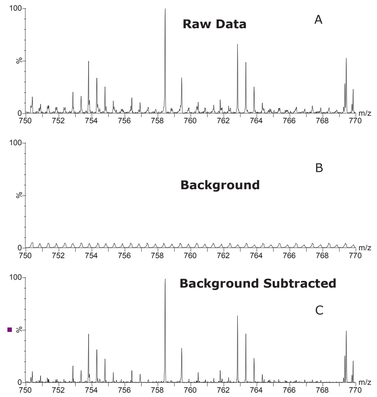

Figure 3. Part of a mass spectrum illustrating the effect of adaptive background subtraction. The original data (Fig.3A) consists of 1639 points with positive intensities, while the subtracted data (Fig.3C) has 899 points with positive intensity. The subtracted background is shown in Fig.3B.

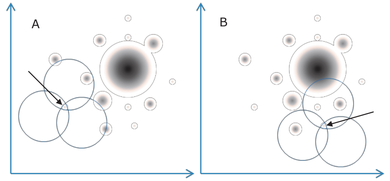

Figure 4. A schematic representation of part of a 2D dataset illustrating the "Data Sweep" method of data reduction. Spots of different sizes correspond to datapoints with different intensities. The point indicated in Fig.4A is discarded, as none of the possible peak positions (some examples of which are represented by the unfilled circles) correspond to peaks of above-threshold intensity. The point in Fig.4B is retained due to the higher local density of data.

Data Sweep

Thresholding is a simple way to reduce the size of a dataset in which points with intensities above a pre-determined threshold value are retained. However, molecular species are represented in continuous mass spectra as peaks spread out over many data points. Applying a flat threshold to the data will often cause points which lie on the edges of peaks, whose tops lie above the threshold, to be discarded. This effect is more severe in multi-dimensional data (in which peaks have a width in each dimension), and in data which are well sampled (having many points across a peak width).

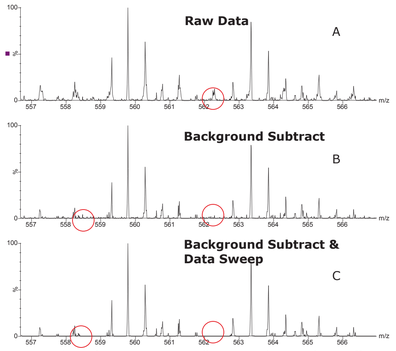

Figure 5. Part of a mass spectrum illustrating the cumulative effect of adaptive background subtraction and data sweep. The top spectrum shows the original data, the middle spectrum shows the data following background subtraction and the bottom spectrum shows the data following a one dimensional data sweep.

In the method described here, this problem is overcome using knowledge of local peak widths. Many methods can be used to estimate the intensity (or maximum possible intensity) of a hypothetical peak at a given position in a multi-dimensional dataset. These methods include simple summation, correlation with known peak shapes and more sophisticated probabilistic approaches.

This calculation is ideally performed at every position in the data and data points that contribute to a hypothetical peak exceeding some pre-determined local threshold intensity are labeled. Unlabeled peaks are then discarded. The local threshold intensity could vary with position in the data and might, for example, be set to achieve a minimum mass precision requirement for a particular application.

The operation of the sweep algorithm in two dimensions is illustrated schematically in Figure 4. A real one dimensional example is given in Figure 5. in which the instrument resolution was used to set the width of the sweep window, and data points contributing to putative peaks having over 10 ion counts were retained.

RESULTS

The original and compressed forms of the LC-MS^E dataset were processed and searched using ProteinLynx Global Server version 2.5.2. Ion detection thresholds were lowered for processing of background subtracted data, but otherwise processing parameters were identical. The requested false positive rate was 4%. The results are presented in Table 1. and Table 2. below. In both cases the "Original" size refers to the native raw file format produced by the instrument.

LC-MS^E	Original	ABS	+ Sweep	+ Lossless
Low Energy	5109 Mb	4669 Mb	2647 Mb	531 Mb
Elev. Energy	5033 Mb	4649 Mb	2184 Mb	406 Mb
Total	10142 Mb	9318 Mb	4831 Mb	937 Mb
Protein ID's	684	667	664	664*

*Table 1. The cumulative results obtained by applying adaptive background subtraction (ABS), Data Sweep and the lossless differentiation, packing and zipping methods to LC-MS^E data. The maximum compression ratio is over 10x, while the number of proteins identified has been reduced by less than 3%. *Lossless compression obviously has no effect on search results.*

LC-IMS-MS^E	Original	ABS	+ Lossless
Low Energy	9572 Mb	4856 Mb	1465 Mb
Elev. Energy	10514 Mb	5313 Mb	1617 Mb
Total	20086 Mb	10166 Mb	3082 Mb
Protein ID's	823	851	851

Table 2. The cumulative results obtained by applying adaptive background subtraction, and the lossless differentiation, packing and zipping methods to LC-IMS-MS^E data. The maximum compression ratio is over 6x. The number of proteins identified has increased by around 3%.

DISCUSSION

The results clearly illustrate that useful compression of electrospray time-of-flight MS datasets is possible without significant loss of data quality. In particular, over ten-fold compression of the LC-MS^E dataset is achieved. At the same time, no statistically significant decrease in the number of proteins identified is observed. Interestingly the final, lossless compression step delivers the largest compression ratio.

The ion mobility dataset compresses by less than seven-fold. However it should be noted that the original IMS-MS data format is already more efficient than the MS data format, and the Sweep algorithm has not yet been evaluated alongside the other compression methods for IMS-MS data. In this case a small, but statistically insignificant, increase in the number of proteins identified using the compressed data is actually observed.

It should be emphasized that all of the algorithms described here have intentionally been designed to consume and produce recognizable continuum datasets that can be processed using existing software. More aggressive lossy algorithms could yield further compression, and efficient storage of the data as lists of (one, two or three dimensional) peaks is also possible.

The lossless compression (and decompression) methods are simple and fast and have already been utilized in real-time LC-IMS-MS data collection.

Future work is likely to concentrate on acceleration of the lossy adaptive background subtraction and data sweep algorithms as both algorithms are well suited to parallelization, for example on multi-processor CPU or GPU-based devices.

Further assessment of compressed data quality is also required.

CONCLUSION

- **A variety of methods are presented for compression of continuum multi-dimensional LC-MS and LC-IMS-MS data.**
- **Using a combination of compression methods, 6-10 fold reductions of the size of proteomic data files were observed, when compared with the native raw data format.**
- **The number of proteins identified using the compressed data was within 4% of the original total in both cases, suggesting that data integrity has been maintained.**

References

1. A novel adaptive background subtraction algorithm for reduction of chemical noise in mass spectra. Green *et. al.* Poster ASMS 2006.

TO DOWNLOAD A COPY OF THIS POSTER, VISIT WWW.WATERS.COM/POSTERS