NEAR-LOSSLESS COMPRESSION OF MASS SPECTRA FOR PROTEOMICS

Agnieszka C. Miguel[†] *Martin Kearney-Fischer*[†] *John F. Keane*[‡]
Jeffrey Whiteaker[‡] *Li-Chia Feng*[‡] *Amanda Paulovich*[‡]

[†] Department of Electrical and Computer Engineering,
Seattle University, Seattle, WA 98122-1090
[‡] Fred Hutchinson Cancer Research Center,
1100 Fairview Ave. N., Seattle, WA 98109-1024

ABSTRACT

Recent improvements in mass spectrometry (MS) technology led to an explosive amount of MS data collected and shared. A typical liquid chromatography/mass spectrometry (LC/MS) "image" from the instrument used in this study consists of 4GB of data. To reduce the bit rate required to code the MS data below that of the authors' previous (lossless) algorithm, we introduce a technique for near-lossless compression. It guarantees that each decompressed sample differs from its original value by no more than a user-specified quantity defined as the target Maximum Absolute Distortion (MAD). We evaluate the proposed method by introducing feature-based metrics applied to the decompressed MS data and show that the MAD-based compression outperforms a traditional coding algorithm aimed at minimizing the mean squared error.

Index Terms— spectroscopy, distortion, image coding, data compression

1. INTRODUCTION

Research into the protein composition of biological samples is a critical endeavor in the life sciences, and its progress has been accelerated by the capability to simultaneously estimate peptide sequences and abundances via mass spectrometry (MS) for hundreds of proteins in a given sample [1]. Scientists need to share and aggregate this information, however the large files make storage, processing, visualization, and transmission very challenging. A data set produced by a typical proteomics experiment may consist of 500GB, which would require 1.4 days to transfer using a T3 line (at 43Mbps), 38 days using a T1 line (at 1.54Mbps), or 75 days using a DSL or cable connection (768kbps). Our previously reported lossless compression ratio of 25:1 [2] would allow a T1 line to deliver the data over a weekend rather than 6 weeks, which illustrates the enabling role of data compression in biological research.

In the research reported here, we investigated and developed a method for near-lossless compression of MS data. In near-lossless compression, every sample value in a reconstructed data is guaranteed to differ from the corresponding value in the original data by no more than a user-specified amount.

In Section 2, we review related background material. In Section 3 we introduce our algorithm for near-lossless coding of mass spectra. The results are presented in Section 4. Finally, we conclude in Section 5.

2. BACKGROUND

2.1. Proteomics

Although the Human Genome Project and the refinement of DNA microarrays have led to a recent growth in genomic-based research, it is *proteomics* that is ultimately expected to have a far greater impact on science and medicine [3]. The proteome contains important information that is not contained in gene sequences or mRNA abundances [1].

Proteins must be processed prior to analysis by MS. They are typically chemically digested into lower-mass peptides and stratified by biophysical properties in a high performance liquid chromatography (HPLC) system. The mass analysis process involves ionizing the sample, separating the ions via the mass-to-charge ratio (m/z), and measuring the result [4].

Storage and communication of minimally-processed LC/MS data is necessary so that scientists who wish to compare results (or re-analyze older data with updated tools or protein databases) may access and pass multiple data sets through a common analysis pipeline [5]. This is infeasible for the heavily reduced results files that are currently small enough to store and transmit. Lossless compression tools are emerging to address this need [6, 7, 2] with reported performance of up to $25:1$

The authors J.K., J.W., L-C.F., and A.P. would like to acknowledge support under National Cancer Institute contract 23XS144A. The author A.M. would like to acknowledge support under the 2006 Seattle University Summer Faculty Research Grant. The authors thank Dane Barney from the University of Washington for the bit plane coder source code. Contact author: Dr. Agnieszka Miguel: amiguel@seattleu.edu.

2.2. Data Characteristics

LC/MS data consists of a series of one-dimensional intensity measurements ("scans") over a discrete set of mass-to-charge ratio (m/z) values. We will use an electrospray ionization time-of-flight (ESI-TOF) LC/MS system for illustration. Consecutive scans contain similar peptides and therefore have similar spectra, however the spectrum gradually changes over the range of LC "retention times". For TOF MS, the m/z sampling rate is not constant, rather, it is a function of m/z (Fig. 1).

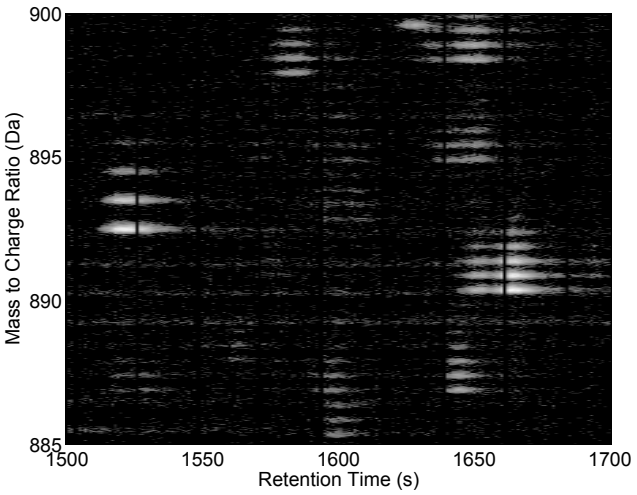

Fig. 1. Example of a region of an LC/MS image. The scale is from black (zero) to white (full scale).

For the TOF instrument, a single scan could contain 200k data points. However, spectra usually have sparsely populated regions, and contiguous blocks of zero intensity are not stored explicitly, making a typical scan 100k points. The spectral peak lobes extend over approximately 20 samples and are digitized to a resolution of approximately 12 bits. For a complex biological sample, a single LC/MS run could produce, for example, 7000 scans and 4GB of data.

2.3. Mass Spectrometry Data Analysis

The goal of feature detection is to quantify the abundance of each peptide in a sample via its ion intensities in the mass spectrum. Peptide information is spread over ions in multiple charge states (e.g., $z = 1, 2, 3...$), multiple masses (i.e., the monoisotope and successively heavier (stable) isotopes), and multiple retention times (scans). For each charge state of a peptide, the feature detector finds the m/z ratio of the monoisotopic peak and infers the charge state based on the $\Delta m/z$ between isotopes. Furthermore, it computes an intensity value for each charge state (e.g., the apex or integral (in one or both dimensions) of the monoisotopic peak, or a sum over all isotopic peaks). Peaks must be detected in the presence of measurement variability and background noise (e.g., due to other peptides and chemicals eluting from the chromatographic column, airborne contaminants, and degradation (i.e., mass reduction) of peptides in the mass analyzer). A typical algorithm for feature detection incorporates background estimation and subtraction, smoothing, individual peak detection, grouping of multiple co-eluting peaks into a set of isotopes of a common peptide and charge state (using knowledge of the expected isotopic distribution), and optionally combining estimates of a peptide seen in multiple charge states.

2.4. Near-Lossless Compression

Most standard lossy data compression algorithms minimize the mean squared error (MSE) between the original data and its decompressed version. Near-lossless coding schemes minimize the maximum absolute distortion (MAD) which is equivalent to the L_∞ norm of $x - \hat{x}$, where x is the original data and $\hat{x}$ is its decompressed version: $MAD(x, \hat{x}) = \max_{i,j} |x(i, j) - \hat{x}(i, j)|$. Every reconstructed sample is guaranteed to differ from its original value by up to a small and preset amount. The advantage of near-lossless coding over a standard MSE-based compression is in maintaining a uniform quality across the whole data set.

Near-lossless coding is used in imaging applications where strict control of the error is required such as medical imaging [8]. Methods used to achieve near-lossless compression include differential pulse code modulation [8], vector quantization [9], and wavelet transforms [10].

3. NEAR-LOSSLESS COMPRESSION USING BIT-PLANE CODING

In this section we describe the proposed method for near-lossless compression of MS data. The first step of the algorithm is universal gridding during which the MS data from all scans are translated onto the same grid. Then, the intensity values are encoded using a bit plane arithmetic coder optimized for satisfying the target MAD.

3.1. Universal Gridding

As the first step in the proposed near-lossless compression algorithm, we translate the MS data onto a universal grid defined as the union of all of the individual grid points in all scans. The original scan points are then mapped onto the universal grid. If the original scan does not have an intensity value for a particular grid point, the intensity is assumed to be zero. The process of generating a universal grid is shown in Fig. 2.

3.2. Bit Plane Coding (MAD-BPAC)

We use a near-lossless coder introduced in [11] that performs bit-plane coding directly on the intensity values. The bit-plane coder uses context-based adaptive binary arithmetic

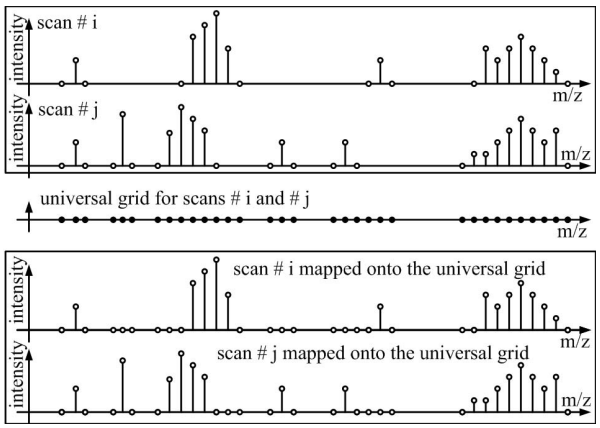

Fig. 2. Process of generating a universal grid for two scans.

coding on the bit-planes. It does not perform any type of transformation on the image pixels, and thus operates in pixel space.

The arithmetic coder uses contexts for coding of the significance pass bits as well as the sign bits. The significance pass contexts are based upon the number of significant pixels among the 8 spatially adjacent neighbors surrounding each pixel. We have found in our tests that limiting the number of contexts to significant neighbor counts of 0, 1, 2, ..., 6, and 7+ has generally produced excellent results.

The pixels are encoded in a priority-based ordering. The priority of a pixel is equal to the number of significant neighbors surrounding the pixel. To maintain synchronization between the encoder and the decoder, in determining this priority, the encoder is only able to count neighbors which have already been encoded. Therefore, neighbors which have not yet been encoded are simply assumed to be insignificant.

The data is encoded until the specified MAD is met. Each time a bit of a pixel is encoded, the encoder checks if the target MAD is satisfied for all pixels. If it is, the encoding stops. We will refer to this method as MAD-BPAC.

4. RESULTS

We show our results on a representative data file derived from a hand-mixed sample of rabbit aldolase (39 kDa) and bovine catalase (57 kDa) that were digested with trypsin. To manage the memory requirements of processing this large file (738 MB), we divide it into bands of 100 daltons (Da) which corresponds to approximately 10000 pixels.

We first encode the test file to target MAD values 0-10 with MAD-BPAC and plot the resulting bit rate in Fig. 3. The bit rates for MAD values within the groups 1-2, 3-6, and 7-10 are very similar because the bit plane encoder terminated within the same bit plane. The bit rate for an MAD of 1 is 0.35 bpp. In comparison, the bit rate for an MAD of 0 (corresponding to lossless compression) is 1.09 bpp. Thus near-lossless compression provides a three-fold decrease in the file size compared to a lossless coding.

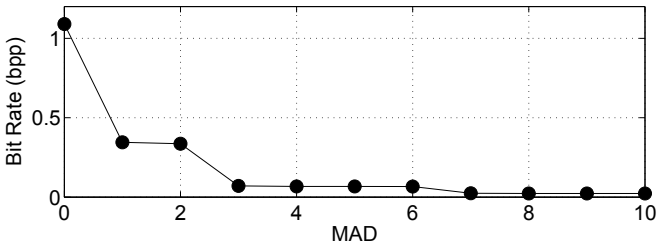

Fig. 3. Compression performance for MAD-BPAC.

To compare the proposed near-lossless algorithm with a traditional coding method based on minimizing the MSE, we encode the data to MAD values 0-10 with MAD-BPAC and compute the resulting bit rate. We then substitute the bit plane coder with JPEG2000 [12] and encode the data to the same bit rate as obtained with MAD-BPAC. Fig. 4 shows a scan after it was decompressed using the two methods at a bit rate of 0.35 bpp. As expected, the waveform obtained using the near-lossless codec is much closer to the original data shape than the waveform obtained using JPEG2000.

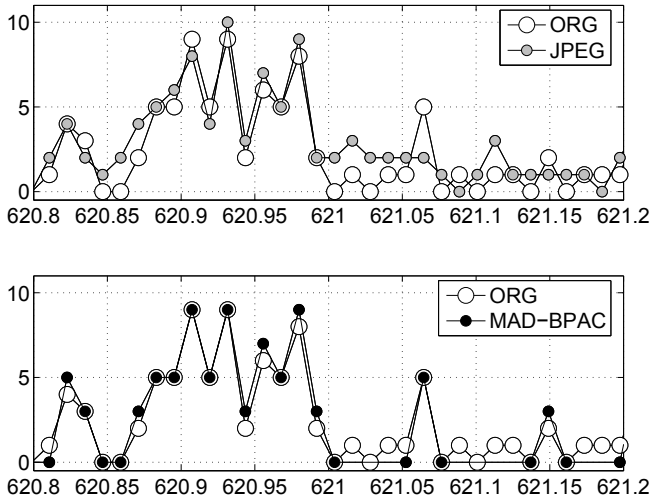

Fig. 4. Decompressed scans using JPEG and MAD-BPAC techniques vs. the original scan.

Next, we validate the proposed method by detecting and comparing features from the unprocessed file and data that was compressed and decompressed by the presented near-lossless coder. We used msInspect [13] (build 2395, strategy=FeatureStrategyPeakClusters) to detect peptide features (isotopic series for each charge state) for our analysis. Fig. 5 shows the percent of correctly matched features and the percent of falsely detected features (artifacts) for each bit rate corresponding to MAD values $1-10$. The impact of compression on the feature values (the m/z and intensity) is shown

in Fig. 6.

The MAD-BPAC outperforms JPEG2000 in the number of correctly identified features. It also tends to detect new features (mostly with low intensity) that were not found in the original data. The errors in feature values are largest in low intensity features. In general, errors increase with larger MAD. However, as MAD grows (e.g. from 6 to 7), many low intensity features fail to be detected and the relative intensity error falls. For m/z, even at MAD of 10 the m/z error is only on the order of the instrument resolution.

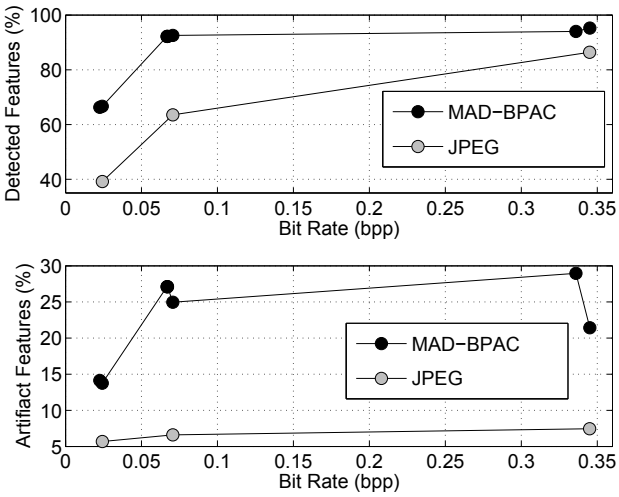

Fig. 5. Feature detection results for JPEG and MAD-BPAC techniques.

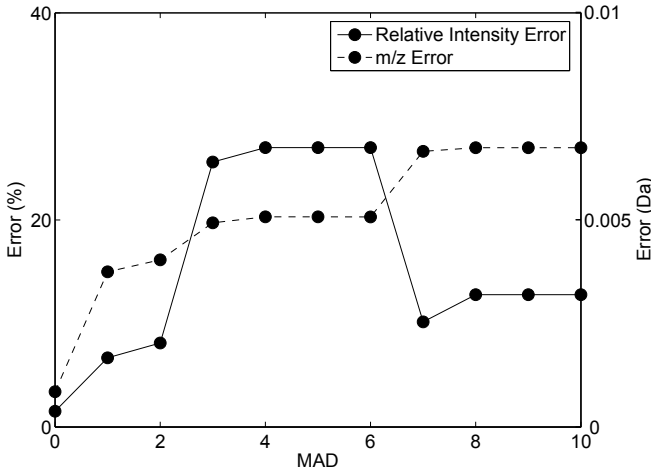

Fig. 6. Feature characteristics for the MAD-BPAC technique.

5. CONCLUSION

We have proposed a near-lossless bit plane coder that uses context-based adaptive binary arithmetic coding. The presented algorithm guarantees that the intensity of any sample in the decompressed mass spectrum differs from its original value by no more than a user-specified quantity. We have shown that from the perspective of a feature detection tool, the proposed method preserves data integrity at a much lower bit rate than a lossless compression algorithm. Therefore, the near-lossless compression technique is an attractive alternative to storing or transmitting raw or losslessly encoded data when storage space or transmission time need to be minimized. Future work includes the design of near-lossless coding methods that decrease the number of falsely detected features.

6. REFERENCES

[1] S. A. Russell, W. Old, K. A. Resing, and L. Hunter, "Proteomic informatics," *Int. Review of Neurobiology*, vol. 61, pp. 127–157, 2004.

[2] A. C. Miguel, J. F. Keane, J. Whiteaker, H. Zhang, and A. Paulovich, "Compression of LC/MS proteomic data," in *IEEE International Symposium on Computer-Based Medical Systems*, June 2006.

[3] N. C. VerBerkmoes, W. J. Hervey, M. Shah, M. Land, L. Hauser, F. W. Larimer, G. J. V. Berkel, and D. E. Goeringer, "Evaluation of "Shotgun" proteomics for identification of biological threat agents in complex environmental matrixes: Experimental simulations," *Analytical Chemistry*, vol. 77, pp. 923–932, February 2005.

[4] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, pp. 198–207, March 2003.

[5] L. Martens, A. I. Nesvizhskii, H. Hermjakob, M. Adamski, G. S. Omenn, J. Vandekerckhove, and K. Gevaert, "Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories," *Proteomics*, vol. 5, no. 13, pp. 3501–3505, 2005.

[6] P. G. Pedrioli, J. S. Eddes, J. K. Eng, N. L. King, B. Pratt, D. Shteynberg, J. M. Tasman, N. Zhang, and R. Aebersold, "The mzXML schema version 3.0," in *Proceedings of the 54th ASMS Conference on Mass Spectrometry*, Poster Number: 387, 2006.

[7] M. Xie and B. Ma, "MSPack - mass spectrometry data compression software," in *Proceedings of the 54th ASMS Conference on Mass Spectrometry*, Poster Number: 071, 2006.

[8] K. Chen and T. Ramabadran, "Near-lossless compression of medical images through entropy-coded DPCM," *IEEE Transactions on Medical Imaging*, vol. 13, no. 3, pp. 538–548, 1994.

[9] G. Motta, F. Rizzo, and J. A. Storer, "Compression of hyperspectral imagery," in *Proc. Data Compression Conference*, pp. 333–342, 2003.

[10] R. Ansari, N. Memon, and E. Ceran, "Near-lossless image compression techniques," *Journal of Electronic Imaging*, vol. 7, no. 3, pp. 486–494, 1998.

[11] A. Miguel, J. Liu, D. Barney, R. Ladner, and E. Riskin, "Near-lossless compression of hyperspectral images," in *Proceedings of the International Conference on Image Processing*, 2006.

[12] D. S. Taubman and M. W. Marcellin, *JPEG2000 : image compression fundamentals, standards, and practice*. Kluwer international series in engineering and computer science, Boston: Kluwer Academic Publishers, 2002.

[13] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. E. J, R. Fang, C. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, and M. McIntosh, "A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS," *Bioinformatics*, vol. 22, no. 15, pp. 1902–9.