RAPID COMMUNICATIONS IN MASS SPECTROMETRY
Rapid Commun. Mass Spectrom. 2009; **23**: 1229–1233
Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/rcm.3962

Efficient encoding and rapid decoding for interactive visualization of large three-dimensional hyperspectral chemical images

Stephen E. Reichenbach[1]*, Alex Henderson[2], Robert Lindquist[1] and Qingping Tao[3]
[1]Department of Computer Science and Engineering, University of Nebraska – Lincoln, Lincoln, NE 68588-0115, USA
[2]Surface Analysis Research Centre, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester M1 7DN, UK
[3]GC Image, LLC, Lincoln, NE 68506, USA

Received 28 October 2008; Revised 23 January 2009; Accepted 23 January 2009

Interactive visualization of data from a new generation of chemical imaging systems requires coding that is efficient and accessible. New technologies for secondary ion mass spectrometry (SIMS) generate large three-dimensional, hyperspectral datasets with high spatial and spectral resolution. Interactive visualization is important for chemical analysis, but the raw dataset size exceeds the memory capacities of typical current computer systems and is a significant obstacle. This paper reports the development of a lossless coding method that is memory efficient, enabling large SIMS datasets to be held in fast memory, and supports quick access for interactive visualization. The approach provides pixel indexing, as required for chemical imaging applications, and is based on the statistical characteristics of the data. The method uses differential time-of-flight to effect mass-spectral run-length-encoding and uses a scheme for variable-length, byte-unit representations for both mass-spectral time-of-flight and intensity values. Experiments demonstrate high compression rates and fast access. Copyright © 2009 John Wiley & Sons, Ltd.

The lossless coding scheme described in this paper facilitates rapid visualization and analysis of large, multi-dimensional, hyperspectral datasets generated by a new generation of chemical imaging systems such as the time-of-flight secondary ion mass spectrometry (ToF-SIMS) instrument developed by Vickerman and co-workers at the Manchester Interdisciplinary Biocentre (MIB) in conjunction with Ionoptika (Chandler's Ford, UK).[1] In ToF-SIMS, a beam of primary ions is directed onto a target, eroding molecules and molecular fragments as neutral species and ions (i.e. secondary ions) from the target surface (as illustrated in Fig. 1). The secondary ions that are eroded from the target surface are electrostatically accelerated to a detector that measures their intensity as a function of flight time – data that can be converted into mass spectra.[2] The primary-ion beam can be directed in a raster pattern to create a mass-spectral image and the raster scanning can be repeated to generate a three-dimensional (3D) mass-spectral image, as illustrated in Fig. 2.

The Ionoptika J105 3D chemical imager has been detailed elsewhere.[1] Briefly, it combines several advances, including polyatomic primary-ion beams and an advanced buncher for secondary ions that facilitates a continuous-beam primary-ion probe. Polyatomic primary-ion beams (e.g. buckminsterfullerene, $[C_{60}]^+$) provide greater secondary-ion yield, more uniformity in the secondary-ion yield, and less damage to the substrate of the target than traditional primary-ion beams.[3] Greater yield improves the signal-to-noise ratio and sensitivity. Improved uniformity enhances effective resolution and allows more accurate mapping of chemical constituents. With reduced sub-surface degradation, as the surface is eroded, subsequent scans across the target yield more accurate depth profiling to improve 3D chemical imaging. An innovative secondary-ion buncher shapes the electric field that propels the secondary ions for the time-of-flight (ToF) mass spectrometer, thereby focusing the variable-sized, variable-positioned secondary ions. Time focusing obviates the need to pulse the primary-ion beam to limit the time range of the secondary ions, which allows quasi-continuous operation of the primary-ion beam. The continuous primary-ion beam provides faster analyses and increased spatial resolution. The system then uses a harmonic-field reflectron with the property that the time-of-flight in and out of the reflector depends on mass-to-charge only (and not on their variable energy). This creative design provides high-precision mass spectrometry even with continuous operation of the primary-ion beam.

The system's high spatial resolution, fine mass precision, and high-sensitivity surface and depth-profile characterizations of the molecular chemistry of heterogeneous materials, including biological tissues and cells, promise to

*Correspondence to: S. E. Reichenbach, 260 Avery Hall, University of Nebraska – Lincoln, Lincoln, NE 68588-0115, USA.
E-mail: reich@cse.unl.edu

Contract/grant sponsor: UK Engineering and Physical Sciences Research Council; contract/grant number: EP/FO12985.
Contract/grant sponsor: The Faculty Development Leave Program of the University of Nebraska – Lincoln.
Contract/grant sponsor: The USA National Science Foundation; contract/grant numbers: IIS-0431119 and IIP-0741027NER/O/S/2003/00667.

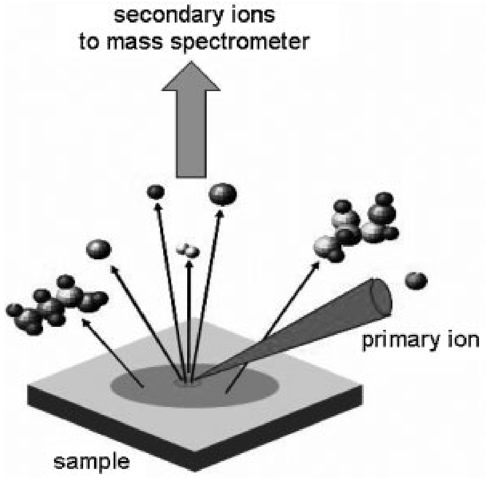
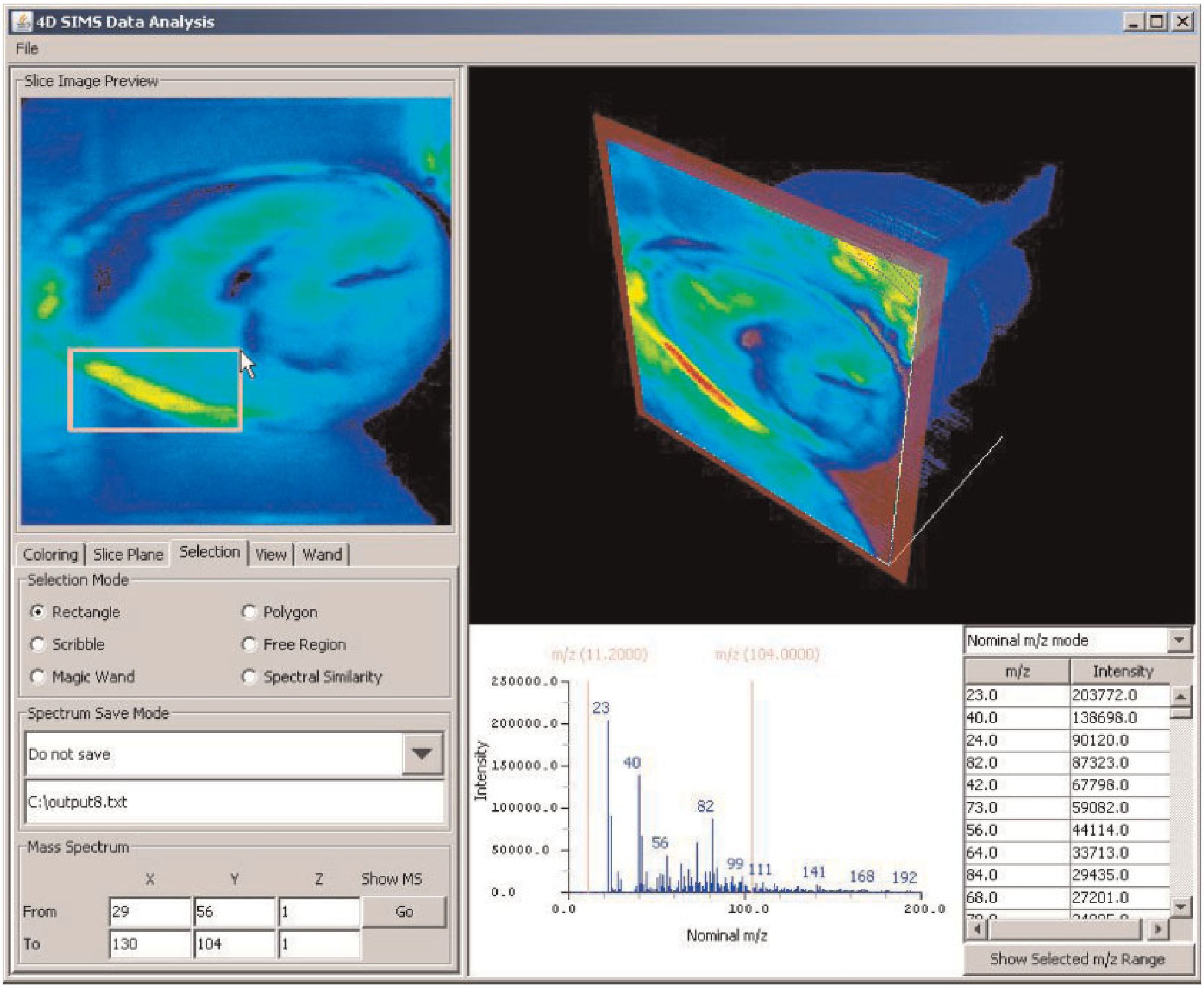


Figure 1. Primary ions directed at the target erode secondary ions for analysis by mass spectrometry.

provide an informational basis for important advances in a wide variety of applications, including cancer treatments. However, the volume of data produced poses a significant challenge for interactive visualization and analysis.

When fully operational, this ToF-SIMS instrument will produce 3D datasets of the order of 512^3 spectra with tens of thousands of ToF channels. In the example datasets presented here, individual mass spectra are sampled in up to 85 000 ToF channels at a rate of 1 ns per 8-bit intensity (a raw data rate of 1 gigabyte per second) and accumulated in hardware with an Ortec Fastflight-2^{TM} (Oak Ridge, TN, USA). In the example datasets, 200 to 1000 raw spectra are accumulated per pixel, but the number may be larger or smaller depending on the application. If no more than 256 spectra are accumulated, each accumulated intensity can be represented with a 16-bit unsigned integer in the raw data file, reducing the data rate to 10 megabytes per second (MB/s). However, a 128 × 128, 16-layer image with 85 000 2-byte ToF channels requires 45 gigabytes (GB) without compression and even a single two-dimensional slice with



Selection Mode

- ● Rectangle
- ○ Polygon
- ○ Scribble
- ○ Free Region
- ○ Magic Wand
- ○ Spectral Similarity

Spectrum Save Mode

Do not save

C:\output8.txt

Mass Spectrum

	X	Y	Z	Show MS
From	29	56	1	Go
To	130	104	1	

Nominal m/z mode

m/z	Intensity
23.0	203772.0
40.0	138698.0
24.0	90120.0
82.0	87323.0
42.0	67798.0
73.0	59082.0
56.0	44114.0
64.0	33713.0
84.0	29435.0
68.0	27201.0

Figure 2. A 3D SIMS image colorized for intensity in mass spectral ranges. (Data from Fletcher *et al*[3] Visualization and analysis software from GC Image, LLC.) This figure is available in color online at www.interscience.wiley.com/journal/rcm.

DOI: 10.1002/rcm

512×512 data points with 85,000 2-byte ToF channels requires 45GB. Datasets of tens to hundreds of gigabytes cannot be held in the fast memory of typical computer systems, which creates a bottleneck for interactive visualization and analysis with general-purpose imaging software. Real-time, interactive, 3D visualization and analysis require memory-efficient coding.

The most important access mode for interactive SIMS visualization is retrieving spectra by pixel (i.e. spatial position). Analysts determine chemical compositions on the basis of mass-spectral characteristics, so viewing of the mass spectra is fundamental. Common interactive operations are to view the mass spectrum at a point in the image space indicated by point-and-click and to view the mass spectrum summed over a spatial region indicated by drawing. An important operation is to generate a classification rule(s) based on mass spectra in two (or more) regions. Analysts also view a mass-spectral range (e.g. for a selected ion) across the image space, but such spectral-spatial viewing does not require immediate interactivity to the degree required in pixel-oriented spatial-spectral access. Moreover, operations to generate spatial maps from their spectra require pixel-by-pixel access to many or all ToF channels, e.g. to map regions that satisfy a classification rule.

Tretter *et al.*[4] cite two approaches among methods for lossless compression of hyperspectral images: predictive coding and reversible transforms, each followed by context modeling and coding. Both approaches can be applied either with respect to the spatial dimensions or to the spectral dimension (or both). Predictive coding has been the predominant approach for hyperspectral data. Lossless transform coding methods for hyperspectral data are newer and typically require greater computation than lossless predictive methods, but may achieve greater compression. Given the motivation of interactive visualization and analysis of SIMS data, low computational complexity is more important than optimal compression, so the more traditional approach of predictive coding may be better suited. Given the primary need for spatial access in SIMS analysis, each pixel spectrum should be compressed separately.

This paper describes a new approach that codes individual spectra, consistent with the predominant access mode for SIMS analysis, based on statistical and structural characteristics of SIMS spectra. Unlike some hyperspectral data generated by remote sensing satellites, for which many hyperspectral compression methods have been developed, SIMS spectra have many zero values and the probability distribution of the intensity values is skewed significantly, decreasing rapidly with magnitude. In addition, many of the non-zero values are in adjacent ToF channels, forming peaks in the mass spectra. Other technologies, such as matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) and confocal fluorescence or Raman microscopy, may produce data with similar statistical characteristics, in which case the approach described here would be applicable. As described in the next section, these statistical characteristics can be exploited to give highly compressed data that can be accessed quickly.

SIMS DATA CHARACTERISTICS

The most notable characteristic of SIMS hyperspectral data is that many of the intensity values are zero. As shown in Table 1, more than 96% of all intensity values in each of the example datasets are zero. (The datasets are divided into units smaller than 2GB to facilitate experimental processing.) For example, of the 1380 million spectral intensities for target 20071213z0 (128×128 with 85 000 ToF channels), only 38 million non-zero intensities are recorded. This characteristic of the data reflects the fact that the number of chemical constituents at a sample point of the target limits the number of secondary ions and therefore the number of peaks in each mass spectrum.

Another important characteristic of the data is that the probability distribution of the intensity values decreases with intensity. Many of the non-zero values are equal to one (many of which may be noise but must be coded in a lossless method) and most non-zero values are less than 256. Because of the large number of zero values and the long-tailed skewed probability distributions, least-squares predictors, which are effective for remote sensing data, do not perform well for SIMS data. For example, with the example dataset 20071213z0, the optimal least-squares predictor based on the previous intensity value in the spectrum reduces the variance in the residual (from that variance of the data), but increases the entropy to 0.45 bits/value compared with the original dataset entropy of 0.34 bits/value. (Entropy is computed as $\sum_i \mathrm{P}[i]\ \mathrm{lg}(\mathrm{P}[i])$, where i is each intensity value and $\mathrm{P}[i]$ is the probability of the intensity value.)

Another characteristic of the datasets is that peaks in the mass spectra may be wider than the ToF channels, so each mass-spectral peak may cause several non-zero values in adjacent ToF channels. The high probability of zero values and the clustering of non-zero values suggest that run length encoding may be used to effectively code the long runs of zeros between non-zero values. Commonly used sparse array representations of mass spectra (i.e. recording the mass and value for each non-zero value) similarly take advantage of the large number of zeros to efficiently represent MS data. If, instead of the ToF channel index, the differential of indexes of non-zero-valued channels is used (i.e. the

Table 1. Intensity probability distributions (%)

Dataset	Count(M)	$P(x=0)$	$P(x=1)$	$P(1<x<2^8)$	$P(2^8 \le x < 2^{16})$	$P(2^{16} \le x)$	Entropy
GridSpot	938	96.85	1.08	1.87	0.21	0.00	0.36
20071213z0	1380	97.22	0.31	2.48	0.00	0.00	0.34
20071213z1	1371	98.04	0.22	1.75	0.00	0.00	0.25
20071213z2	1360	98.65	0.15	1.20	0.00	0.00	0.18

DOI: 10.1002/rcm

Table 2. ToF differential probability distributions (%)

Dataset	Count(M)	$P(d' = 0)$	$P(0 < d' < 2^8)$	$P(2^8 \leq d' < 2^{16})$	$P(2^{16} \leq d')$	Entropy
GridSpot	30	54.18	43.78	2.04	0.00	4.36
20071213z0	38	65.00	31.60	3.40	0.00	3.60
20071213z1	27	64.64	30.44	4.92	0.00	3.76
20071213z2	18	64.34	28.77	6.94	0.00	3.94

difference between the index of the next channel with a non-zero value and one more than the index of the current non-zero channel), the result is a run length code. For example, if the channels with non-zero values are 7, 50, 89, 188, 189, 198, 199, 200, . . .; then the differentials of the indices are 6, 42, 38, 98, 0, 8, 0, 0, . . ., which are just the run lengths of the zeros.

The probability distribution and entropy of the example ToF differentials is shown in Table 2. Because the non-zero values tend to be clustered, more than half of the ToF differentials in each of the example datasets are equal to zero. Many of the other differentials are less than 256. For the dataset 20071213z0, the total entropy of the ToF differentials (entropy per value times number of values) is 17MB with only 38 million non-zero intensities to be coded. This approach is the basis of the method described in the next section.

SIMS DATA COMPRESSION

Based on the SIMS data characteristics, the method developed here separately codes the ToF differentials and non-zero intensity values. Because many of the ToF differentials are zero, it is important to code them most efficiently. Because the compressed data will be decoded for visualization, the approach uses representations of integer byte-lengths which do not require computation for decoding – just byte copies. Accordingly, the method uses 2-bit length codes to record the number of bytes for each ToF differential and zero bytes are used if the differential is zero. The length codes (in binary) are: 00 if the differential is zero, with no separate representation of the differential; 01 if the differential is in the range 1–255, with the differential coded in one byte; 10 if the differential is in the range 256–65535, with the differential is coded in two bytes; and 11 if the differential is 65536 or larger, with the differential coded in four bytes. Thus, only two bits are required for the ToF differentials that are equal to zero, 10 bits are required for the ToF differentials in the range 1–255, etc. For the example above, with ToF differentials 6, 42, 38, 98, 0, 8, 0, 0, . . ., the length codes would be 0101010100010000 (two bytes in binary) and the differential codes would be 062A266208 (five bytes in hexadecimal). The differential codes can be retrieved quickly using byte copies. For the dataset 20071213z0, this coding of the ToF differentials requires 24MB, compared with the total entropy of 17MB, but allows very rapid decoding (as documented in the next section) at the cost of a relatively small difference in compression.

The non-zero intensity values could be compressed by any method, but the scheme used for the ToF differentials can be used and is justified by the significant number of ones and small values. The integer byte-length scheme also allows quick retrieval of the intensity for a specific channel, decoding only the ToF differentials and the intensity byte-lengths to locate the byte(s) with the intensity value. Thus, here, the non-zero intensity values are reduced by one (so that the smallest value to be recorded, which is one, is mapped to zero) and then coded using the length-coding scheme described above. For the dataset 20071213z0, this coding of the non-zero values requires 44MB, compared with a total entropy of 27MB for the non-zero values, but here also decoding requires only byte copies and is very fast (as documented in the next section).

RESULTS

This section compares the speed and compression of the proposed SIMS compression method with GZIP (in java.util.zip[5]), arithmetic coding with an adaptive unigram model (in com.colloquial.arithcode[6]), and arithmetic coding with Prediction by Partial Matching[7,8] (PPM(3) in com.colloquial.arithcode). Each method was applied separately to each of the spectra in each of the datasets.

Table 3 summarizes the compression times and sizes. (The raw dataset type for each intensity is a four-byte integer.) PPM(3) with arithmetic coding requires more computation but achieves better compression than the adaptive unigram model with arithmetic coding, and the adaptive unigram model with arithmetic coding requires more computation but achieves better compression than GZIP. These results are as expected. The SIMS coding method achieved the greatest compression, but more importantly was very rapid. The encoding times for the SIMS method were less than 30% of

Table 3. Compression and decompression times and sizes

Dataset		GZIP			Adaptive Unigram			PPM(3)			SIMS		
Name	Size (MB)	Encode time(s)	Decode time(s)	Size (MB)	Encode time(s)	Decode time(s)	Size (MB)	Encode time(s)	Decode time(s)	Size (MB)	Encode time(s)	Decode time(s)	Size (MB)
Grid Spot	3750	88	28	76	1621	1846	56	1743	2546	54	26	1	50
20071213z0	5520	134	40	98	2386	2739	74	2760	4172	73	38	1	68
20071213z1	5486	126	40	72	2380	2735	55	2501	3669	53	37	1	48
20071213z2	5440	123	41	52	2414	2788	41	2347	3345	38	37	1	33

DOI: 10.1002/rcm

the times for GZIP. More importantly, the decompression times for the SIMS method were less than 3% of the times for GZIP. Decompression is especially fast because only the bytes for the non-zero intensities must be copied into the output array.

The compression rate for all methods would enable fairly large ToF-SIMS datasets to be stored in the memory of current desktop systems (typically 2–8 GB). Pixel mass spectra are compressed to an average of about 3 kB to 6 kB, so with any of the methods a typical memory could hold more than 1M spectra (e.g. a 1024 × 1024 slice or even a 64^3 or 128^3 visualization data cube). The performance of the SIMS method, both for compression and decoding speed, is excellent.

CONCLUSIONS

This paper describes the development of a new coding method for multi-dimensional hyperspectral data generated by advanced chemical imaging systems, such as ToF-SIMS. The method is designed based on data characteristics to provide indexed access to pixel spectra with very rapid decoding. Experimental results indicate that the method achieves memory-efficient compression and provides quick access. Although the method was developed and tested for ToF-SIMS data, it should be effective for other sparse hyperspectral data with predominantly small values and skewed probability distribution.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council's ''Collaborating for Success through People'' funding to John C. Vickerman (EP/FO12985), by the Faculty Development Leave program of the University of Nebraska – Lincoln, and by the USA National Science Foundation funding to S. E. Reichenbach (IIS-0431119) and Q. Tao (IIP-0741027). The authors gratefully acknowledge the support and data provided by John Vickerman and John Fletcher of the Surface Analysis Research Centre at the University of Manchester.

REFERENCES

1. Fletcher JS, Rabbani S, Henderson A, Blenkinsopp P, Thompson S, Lockyer NP, Vickerman JC. *Anal. Chem.* 2008; **80**: 9058.
2. Vickerman JC. In *ToF-SIMS: Surface Analysis by Mass Spectrometry*, Vickerman JC, Briggs D (eds). SurfaceSpectra: Manchester, UK, 2001.
3. Fletcher JS, Lockyer NP, Vaidyanathan S, Vickerman JC. *Anal. Chem.* 2007; **79**: 2199.
4. Tretter D, Memon N, Bouman CA. In *Handbook of Image and Video Processing*, Bovik A (ed). Academic Press: San Diego, 2000.
5. Sun Microsystems. *Package java.util.zip*. Available: http://java.sun.com/j2se/1.5.0/docs/api/java/util/zip/package-summary.html/.
6. Carpenter B. *Arithmetic Coding*. Available: http://www.colloquial.com/ArithmeticCoding/javadoc/index.html.
7. Cleary JG, Witten IH. *IEEE T. Comput.* 1984; **32**: 396.
8. Witten IH, Neal R, Cleary JG. *Commun. ACM* 1987; **30**: 520.

DOI: 10.1002/rcm