Efficient and Rapid Encoding of Mass Spectral Data for Interactive Visualization of 2D and 3D Hyperspectral Images

Alex Henderson[1,2], Stephen E. Reichenbach[3], Qingping Tao[4], John C. Vickerman[1,2]

[1]Surface Analysis Research Centre, University of Manchester, Manchester, UK [2]SurfaceSpectra Ltd, Manchester, UK
[3]Computer Science & Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska, USA [4]GC Image LLC, Lincoln, Nebraska, USA

Background

With the advent of SIMS instrumentation that can provide 2D and 3D image data with high spatial resolution and high mass resolution at the same time, there is a problem with the volume of data. Here we present a data encoding system proving both efficient compression and quick access times.

SIMS Data Characteristics

Spectra from image pixels have a high number of zero intensity values, >90% in many cases.
Probability distribution of intensity values decreases with intensity.
Therefore, many of the non-zero intensity values are 1 and few intensity values exceed 255.
Mass spectral peaks are wider than the ToF channels, so non-zero intensity values occur in clusters.

Data Encoding

Each pixel's spectrum is encoded individually. Store ***differential*** between **non-zero** time channels.
Non-zero intensities and their ToF differentials coded separately.
Zero intensities and their ToFs are not encoded explicitly.
Two-bit length codes indicate number of bytes required to store differential.

- For ToF differential:
  - `00` for differential zero (no bytes required)
  - `01` for differential 1 – 255 (one byte for the value)
  - `10` for differential 256 – 65535 (two bytes for the value)
  - `11` for differential greater than 65535 (four bytes for the value)
- For intensities, the value is decreased by one and the same length codes are used.

So, only two bits are required for ToF differentials equal to zero and only two bytes are required for intensity values equal to one. For fast decoding, coding system uses integer byte lengths.

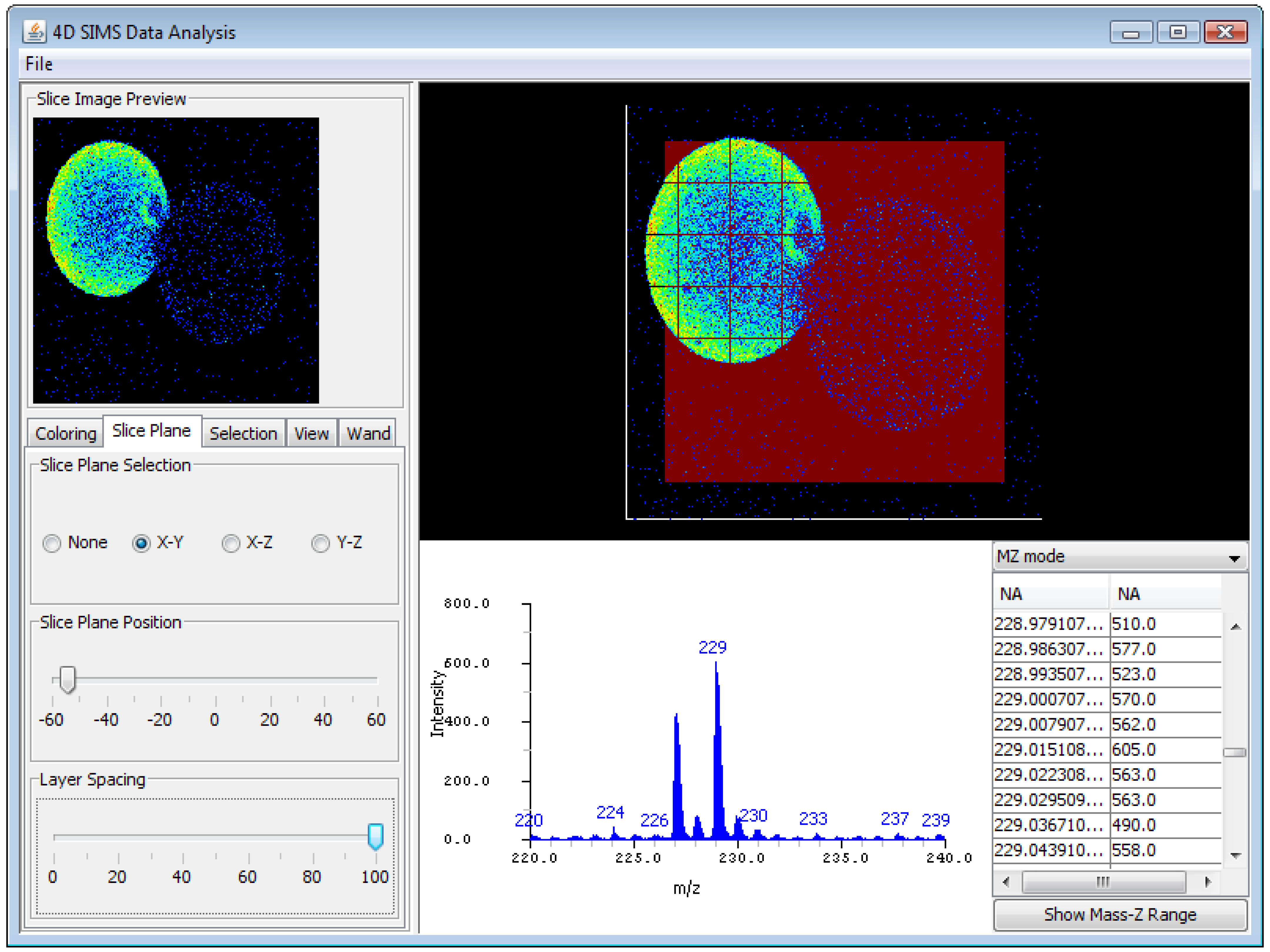
Data Decoding

Each pixel encoding is indexed for fast retrieval.
The two-bit length codes can be decoded quickly.
- If the ToF differential is zero, then no additional decoding is required.
- Similarly, if the intensity value is one, then no additional decoding is required.

If additional decoding is required, only byte copies are necessary.

Coding Example

Non-zero ToF channels: 7, 50, 89, 188, 189, 198, 199, 200, …
ToF differentials (zeroes between non-zeroes): 6, 42, 38, 98, 0, 8, 0, 0, …
Two-bit length codes for ToF differentials: `01, 01, 01, 01, 00, 01, 00, 00`, … (binary)
Additional bytes for ToF differentials: `06, 2A, 26, 62, 08`, … (hexadecimal)



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)
Example 1	3750	88	28	76	1621	1846	56	1743	2546	54	26	1	50
Example 2	5520	134	40	98	2386	2739	74	2760	4172	73	38	1	68
Example 3	5486	126	40	72	2380	2735	55	2501	3669	53	37	1	48
Example 4	5440	123	41	52	2414	2788	41	2347	3345	38	37	1	33

Examples

SIMS coding compared with GZIP (in java.util.zip), arithmetic coding with adaptive unigram model (in com.colloqual.arithcode[1]), and Prediction with Partial Matching (PPM(3), in com.colloquial.arithcode[2]).

[1] B. Carpenter, www.colloquial.com, 2003. [2] J.G. Cleary and I.H. Witten, *IEEE Trans. Computers*, 32(4):396-402, 1984.

GC IMAGE Software for Multidimensional Chromatography

http://manchester.ac.uk/sarc/