SIMS Proceedings Papers

SURFACE and INTERFACE ANALYSIS

Received: 8 September 2009 Revised: 1 April 2010 Accepted: 4 April 2010 Published online in Wiley Interscience:

(www.interscience.wiley.com) DOI 10.1002/sia.3553

Interactive spatio-spectral analysis of three-dimensional mass-spectral (3DxMS) chemical images

Stephen E. Reichenbach,[a]* Xue Tian,[a] Robert Lindquist,[b] Qingping Tao,[b] Alex Henderson[c] and John C. Vickerman[c]

Emerging technologies for chemical imaging provide high-resolution three-dimensional (3D) surveys with high-precision mass spectrometry (MS), promising to open unprecedented vistas for understanding complex phenomena such as cellular metabolism. However, there are critical challenges in transforming the large, complex, multidimensional, multispectral data sets into useful chemical information for biological research and other applications. This paper describes new informatics for advanced interactive spatio-spectral analysis of three-dimensional mass-spectral (3DxMS) chemical images. The technical challenges for interactive informatics are rapid access to large datasets, visualization of 3D hyperspectral images, and pattern recognition for spatio-spectral mapping. This paper describes an effective compression method for time-of-flight secondary ion mass spectrometry (ToF-SIMS) data that provides rapid spatial-spectral access; a framework for 3DxMS visualization that supports multiple views with multiple layers of information; and a suite of pattern recognition tools for spatio-spectral drawing, clustering, and classification. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: hyperspectral image processing; three-dimensional image processing; secondary ion mass spectrometry (SIMS)

Introduction

Advanced informatics are required to support interactive spatio-spectral analysis of three-dimensional mass-spectral (3DxMS) chemical images generated by an emerging generation of time-of-flight secondary ion mass spectrometry (ToF-SIMS) instruments. These instruments provide high spatial resolution and fine mass-spectral precision of biological samples and promise to provide an informational basis for important scientific advances, but the volume and complexity of data pose significant challenges for interactive visualization and analysis.

A new generation of ToF-SIMS systems, exemplified by the J105 3D Chemical Imager developed by the University of Manchester[1,2] and the hybrid quadrupole orthogonal ToF-SIMS system developed by Penn State University,[3] utilize polyatomic primary ion beams with high duty-cycles to achieve faster analyses and higher spatial resolution without sacrificing mass resolution and with subsurface degradation that is small enough to allow depth profiling for 3DxMS imaging of single biological cells.

Despite the promise of ToF-SIMS for biosciences, a lack of information technologies to support advanced data analysis and 'push-button' methods for routine applications is a significant impediment to its adoption.[4,5] Extracting the rich chemical information offered by emerging chemical imaging technologies from large, complex data is a substantial challenge. Currently, there is an insufficient knowledge basis for fully automated processing of 3DxMS data from biological samples, so this research and development focuses on interactive and semiautomated operations.

Three significant challenges for interactive informatics with three-dimensional (3D) chemical imaging are:

- Rapid access to large datasets.
- Interactive visualization of complex multidimensional multi-spectral data.
- Pattern recognition for spatio-spectral mapping.

This paper describes an effective compression method for ToF-SIMS data that provides rapid spatial–spectral access; a framework for 3DxMS visualization that supports multiple views with multiple layers of information; and a suite of pattern recognition tools for spatio-spectral drawing, clustering, and classification.

Data Compression for Rapid Access

A practical issue for informatics software is that the data size exceeds the computer memory of typical desktop computers. For example, a ToF-SIMS image acquired with the J105 with 256×256 pixels at each of ten layers and spectra with 10^5 channels defines more than 65 billion data points.

ToF-SIMS data is *hyperspectral*, with intensities for tens of thousands of ToF intervals. Tretter, Memon, and Bouman[6] surveyed two principal approaches for lossless compression of hyperspectral images: predictive coding and reversible transforms,

* *Correspondence to: Professor Stephen E. Reichenbach, Computer Science & Engineering Department, University of Nebraska – Lincoln, Lincoln NE 68588-0115, USA. E-mail: reich@cse.unl.edu*

a *Computer Science & Engineering Department, University of Nebraska – Lincoln, Lincoln NE 68588-0115, USA*

b *GC Image, LLC, PO Box 57403, Lincoln NE 68505-7403, USA*

c *Surface Analysis Research Centre, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester M1 7DN, UK*

Table 1. Compression and decompression times and sizes [7]

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

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 interactivity, low computational complexity is more important than high compression rates, so predictive coding methods are better suited. Given the primary need for spatial visualization in ToF-SIMS analysis, rapid access should be provided to each data-point spectrum.

Reichenbach *et al*.[7] recently described a method that codes individual spectra, consistent with the predominant access mode for ToF-SIMS analysis, based on statistical and structural characteristics of ToF-SIMS spectra. Unlike hyperspectral data generated by most remote sensing satellites, for which popular hyperspectral compression methods were developed, ToF-SIMS spectra have many zero values and the probability distribution of the intensity values is skewed, decreasing rapidly with magnitude. Also, many of the nonzero values are in adjacent ToF channels, forming peaks in the mass spectra. These statistical characteristics can be exploited to give highly compressed data that can be accessed quickly.

The frequency of zero values and the adjacency of nonzero values suggest that run-length encoding (RLE) may be used to effectively code long runs of zeros. Commonly used sparse array representations of mass spectra (i.e. recording the mass and value for each nonzero 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 nonzero-valued channels is used (i.e. the difference between the index of the next channel with a nonzero value and one more than the index of the current nonzero channel), the result is a run-length code.

The compressed data must be decoded quickly for interactive visualization, so the approach represents the run lengths with 1, 2, and 4 byte integers 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 1 byte; '10' if the differential is in the range 256 – 65 535, with the differential coded in two bytes; and '11' if the differential is 65 536 or larger, with the differential coded in 4 bytes. So, 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.* The differential codes can be retrieved quickly using byte copies.

The nonzero intensity values, of which many are one and many others are small, can be compressed similarly. The integer byte-length scheme allows quick retrieval of the intensity for a specific channel, with decoding of only the ToF differentials and the intensity byte lengths to locate the byte(s) with the intensity value. The nonzero intensity values are reduced by one (which maps the ones to zeros) and then coded using the length-coding scheme described above. Decoding restores the nonzero values by adding one.

Table 1 compares compression rates and coding times for the SIMS method and three popular general coding methods: GZIP (in java.util.zip[8]), arithmetic coding with an adaptive unigram model (in com.colloquial.arithcode[9]), and arithmetic coding with Prediction by Partial Matching[10,11] (PPM(3) in com.colloquial.arithcode). The input data, which has been segmented into single layers for processing on desktop computers, is large. For example, in uncompressed form, with 4-byte integers for all values, the Grid Spot data requires 3.75 GB (for 16K pixels, 57K channels/pixel, 4 bytes/channel). Even in list format with time-and-intensity pairs, a commonly used representation for sparse spectra, the 20 071 213z0 data requires 307 MB (for 38M nonzero intensities with 4 bytes each for time and intensity). All of the methods compared in Table 1 substantially compress this data. The SIMS method achieves excellent compression (33 to 68 MB) and fast decoding of ToF-SIMS data for interactive visualization (less than 1 s).

Interactive 3DxMS Visualization

Visualization is the process of converting scientific data into visual information.[12–16] Conventional computer monitors offer two dimensions for displaying the four dimensions of 3DxMS data. Accordingly, a framework for visualization should support different views of the same data, each showing different dimensions of the data. In addition to the four dimensions of data, data processing defines additional spatial and spectral features for visualization, for example, the spatial region of a cell nucleus or the mass-spectral channels indicative of cholesterol. So, the visualization framework also should support mapping of multiple spatial and spectral features generated during data analyses.

A software multi-view multi-layer (MVML) framework for 3DxMS data visualization with a model-view-controller (MVC) architecture[14,17,18] supports various views of ToF-SIMS data with multiple layers of feature information. The model components maintain various data and metadata objects, such as data source, a collection of data access components and a 3D array of mass-spectral vectors; aspect function, a component for computing features; and geometric aspect, a region map in the data space. The controller components handle the interaction between users and

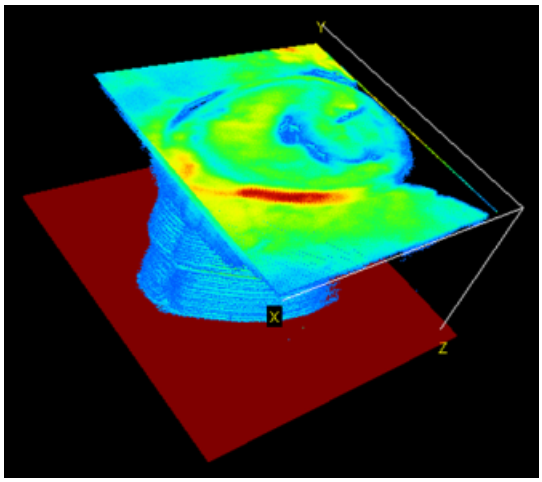

Figure 1. Thresholded values show data points inside a 3D data-cube.

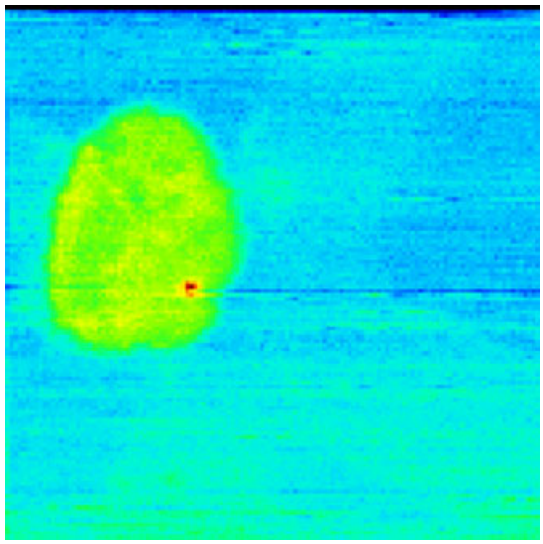

Figure 2. A two-dimensional slice from a three-dimensional data-cube.

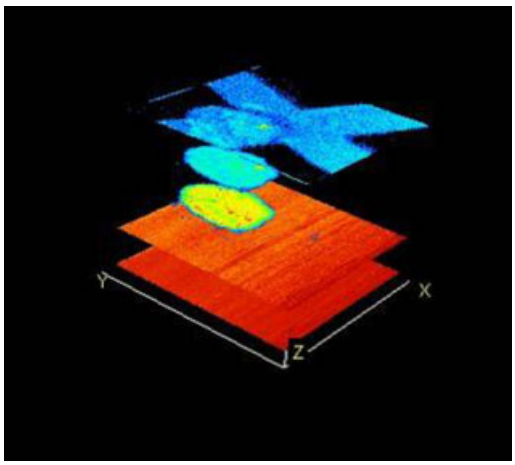

Figure 3. Mass to charge is presented as the third dimension for a slice.

the program, accepting events from users and dispatching events to appropriate receivers. The view components render various visualizations, including 3D spatial projection, two-dimensional (2D) spatial plus mass-spectral projection, 2D spatial slice, and one-dimensional (1D) mass-spectral graph. Each view has configuration parameters, e.g. color mapping, opacity, *etc.*

The 3D visualization in Fig. 1 shows a perspective projection of the spatial data-cube, with pseudocolor mapping of the mass spectrum at each point. A pseudocolor mapping function generates a color for a given spectrum, e.g. using a linear, logarithmic, or exponential function of the total intensity count (TIC) to index into a cold-hot color scale or mapping the selected intensity count (SIC) for user-defined m/Q interval(s) indicative of chemically important ions. Thresholds can be used to make some data points invisible, e.g. creating an isosurface that shows data points with the same (or similar) intensities. Recomputation of the value for each data point (when the m/Q selection changes) is relatively fast if the compressed data can be held in memory. Mouse-controlled rotation allows viewing the data from any 3D angle and zoom moves the data-cube closer or more distant. Radio buttons provide six standard orthogonal views (one of the six data-cube faces shown fully in front) and eight isometric views (one of the eight data-cube corners with the three intersecting faces shown fully at equal angles). A control also is provided for the aspect ratio.

The 2D visualization in Fig. 2 is convenient for viewing a planar slice through the interior of the data-cube. The slice plane can be positioned along any axis in the 3D view and repositioned with a slider. As the slicing plane is repositioned, the selected image slice is shown in real-time in the image viewer. This allows interactive 2D animation of the 3D data along any spatial dimension. Displaying a 2D slice as an image provides a convenient interface for precisely indicating data points or drawing regions, e.g. the circle indicating a point of interest. Other drawing interfaces are single-point, polygon, free-hand, and scribble. The software allows users to build composite regions using discard (new), addition (union), subtraction, and replace (discard followed by addition). Composite regions are maintained as geometric aspect functions that can be saved, loaded, edited, and visualized in the 2D and 3D views.

Each data point in the 2D view is a pseudocolorized mass spectrum, but the mass-spectral dimension also can be visualized as a third dimension, as in Fig. 3. Because the mass-spectral array for ToF-SIMS typically is 1–2 orders of magnitude larger than the screen resolution, resampling is required. Pseudocolorizing each resampled interval independently can highlight spatio-spectral structures (such as the grid in Fig. 3 that may be difficult to discern in other views (e.g. the grid in Fig. 2).

1D visualization is convenient for showing the mass spectrum of an indicated data point or the summed mass spectrum for a spatial region. The spectrum viewer displays a spectrum in graphical and tabular formats. The abscissa of the spectrum can be set to ToF, m/Q, or integer mass to charge (rounded to whole numbers). ToF-SIMS systems generate hyperspectral data – intensity arrays with tens of thousands of values – so neither display screen resolution nor visual acuity is sufficient to perceive all spectral intensities simultaneously. Therefore, the graphical view of the spectrum allows zooming to show sub-ranges of the spectrum and the tabular view supports scrolling. The tabular view can be sorted either by abscissa or intensity (the ordinal) with either increasing or decreasing values. The graphical view allows interactive delineation of a SIC range by mouse click-and-drag. The SIC can then be visualized in the 3D and 2D spatial views with a button click. Generating new SIC visualizations requires summing

intensities for the indicated range and may take a few seconds depending on the size of the data and SIC range. Specific SIC features (e.g. that are indicative of various chemically important ions) can be saved, loaded, and edited, as well as visualized.

Spatio-Spectral Pattern Recognition

Analyses of ToF-SIMS images may entail both spatial and spectral features. For example, it may be useful to delineate a spatial feature such as a cell or regions within a cell. Similarly, it may be useful to identify spectral features, such as a spectral signature indicating the presence of a drug. Pattern recognition can be used to find and/or delineate such features. This section describes a Magic Wand tool, spectral clustering, and spectral classification.

The Magic Wand tool selects data points based on spectral similarity and spatial proximity. First, the user selects a data point to provide both a reference spectrum and a seed for the selected spatial region. Then, the software iterates a region-growing process. In each cycle of the iteration, data points within a specified spatial distance of any selected data point (initially just the seed) are tested for similarity with the reference spectrum. Every data point within the specified distance that meets the similarity criterion is added to the selected region. This iterative process repeats each time the region grows, then stops when no more data points nearby any selected data points are similar enough. Two sliders interactively parameterize the Magic Wand: the jump parameter, which specifies the distance from a selected data point at which that the region-growing process can include additional data points, and the similarity threshold, which specifies the level of similarity required for new data points to be added to the selected region. Two types of similarity can be used: spectral similarity, computed as the cosine between mass-spectral vectors, and TIC similarity, the difference between TIC values divided by the difference between the largest and smallest TIC values in the data.

Figure 4 illustrates the use of Magic Wand on ToF-SIMS data from two 200-µm polystyrene beads coated with different peptide mimics with distinct mass-spectral features at m/Q 226 and 547 on a silicon substrate.[19] In each image, the region selected by the Magic Wand is shown as a mauve-colored overlay on the colorized TIC image. In Fig. 4(A), the seed point is in the background and the Magic Wand successfully selects much of the background. In Fig. 4(B) and 4(C), the seed point is respectively in the left and right bead and the Magic Wand selects much of the indicated bead.

The spectral clustering tool provides a uniform interface for various clustering algorithms, including *k*-means,[20] hierarchical clustering,[21] and spectral clustering.[22] Clustering algorithms perform unsupervised grouping of objects such that those in the same group are more similar with one another than with objects in other groups. First, the user selects a subset of data points to be clustered. Then, the user selects the clustering algorithm and provides its parameters, e.g. some algorithms require the user to specify the number of clusters. Then, the algorithm separates the data points into clusters based on their spectral similarity.

Figure 5 illustrates clustering of ToF-SIMS data from HeLa cells: (A) a TIC image; (B) *k*-means clustering with two clusters separates the cells and background; and (C) *k*-means clustering with three clusters apparently separates the cell edges and interiors.

Spectral classification is based on supervised training: two (or more) user-defined geometric aspect functions designate data points in distinct classes, then the classification algorithm

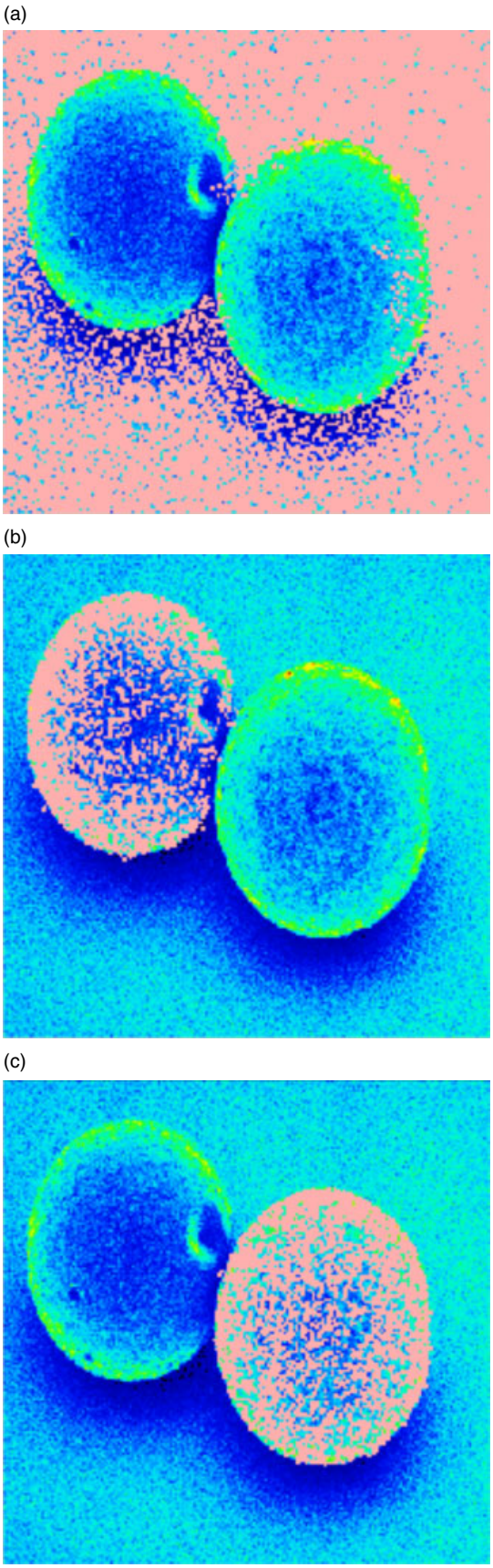

Figure 4. The Magic Wand can be used to select proximal and spectrally similar data points for an indicated seed point. (A) Seed in the background. (B) Seed in the left bead. (C) Seed in the right bead. (Data from Winograd and Braun.[19]).

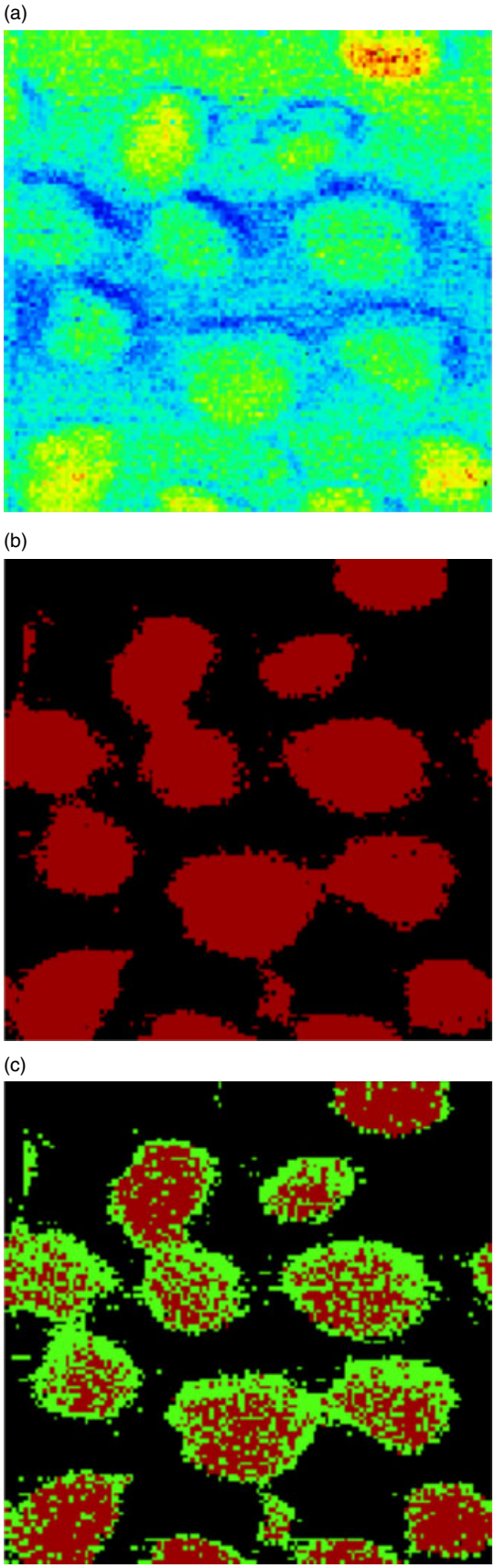

Figure 5. Clustering of ToF-SIMS data of HeLa cells. (Data from S. Rabbani and J. Fletcher, Surface Analysis Research Centre, University of Manchester.) (A) TIC. (B) Two clusters with *k*-means. (C) Three clusters with *k*-means.

Table 2. Classification results with four classification methods for two datasets

Classifier	Dataset 1 accuracy (%)	Dataset 2 accuracy (%)
Decision trees	90.00	93.50
SIMCA	80.00	83.00
PCA with DFA	91.00	92.50
MSN-PSSM	89.00	90.00

assigns class membership to other data points based on mass-spectral characteristics. Experiments with the bead data compared four classification algorithms: C4.5 decision trees,[23] soft independent modeling of class analogy (SIMCA),[24] principal component analysis (PCA)[25] with discriminant function analysis (DFA),[26] and the most similar neighbor with a probability-based spectrum similarity measure (MSN-PSSM).[27] Two data sets were constructed: the first with 100 data points from each bead and the second with 50 data points from each bead. With these data sets, leave-one-out cross-validation, which is commonly used in chemometrics, was used for testing. Table 2 shows the classification accuracy (the fraction correctly classified) for each classification algorithm. Decision trees and PCA with DFA performed best, followed by MSN-PSSM, then SIMCA, but only the lower performance of SIMCA is statistically significant. In experiments with SIMS data from bacterial samples related to urinary tract infections, MSN-PSSM significantly outperformed the other classification methods.[27]

Conclusion

This paper summarizes new software methods and tools for computer-based visualization and analyses of high-resolution, 3D, hyperspectral ToF-SIMS data. The informatics suite includes a coding scheme for efficient storage and fast access, interactive interfaces for visualizing and operating on 3D hyperspectral images, and spatio-spectral clustering and classification. The goal of the work to-date is proof of concept and the development of a prototype foundation for future work. Future work will include continued evaluation of the coding effectiveness as ToF-SIMS instruments evolve, interactive spatial operations for 3D drawing such as rotation and extrusion, improved clustering and supervised classification methods, and a general framework for spectral aspects functions such as PCA, ion ratios, *etc.*

Acknowledgements

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

References

[1] J. S. Fletcher, X. A. Conlan, E. A. Jones, G. Biddulph, N. P. Lockyer, J. C. Vickerman, *Anal. Chem.* **2006**, *78*, 1827.

[2] J. S. Fletcher, S. Rabbani, A. Henderson, P. Blenkinsopp, S. P. Thompson, N. P. Lockyer, J. C. Vickerman, *Anal. Chem.* **2008**, *80*, 9058.
[3] A. Carado, M. K. Passarelli, J. Kozole, J. E. Wingate, N. Winograd, A. V. Loboda, *Anal. Chem.* **2008**, *80*, 7921.
[4] R. Heeren, L. McDonnell, E. Amstalden, S. Luxembourg, A. Altelaar, S. Piersma, *Appl. Surf. Sci.* **2006**, *252*, 6827.
[5] A. V. Walker, *Anal. Chem.* **2008**, *80*, 8865.
[6] D. Tretter, N. Memon, C. A. Bouman, Multispectral image compression, in *Handbook of Image and Video Processing*, (Ed: A. Bovik), 539, Academic Press, San Diego, CA, **2000**, pp. 539–554.
[7] S. E. Reichenbach, A. Henderson, R. Lindquist, Q. Tao, *Rapid Commun. Mass Spectrom.* **2009**, *23*, 1229.
[8] Package java.util.zip. **2004**, http://java.sun.com/j2se/1.5.0/docs/api/java/util/zip/package-summary.html/.
[9] B. Carpenter. Arithmetic Coding. **2003**, http://www.colloquial.com/ArithmeticCoding/javadoc/index.html.
[10] J. Cleary, I. Witten, *IEEE Trans. Comput.* **1984**, *32*, 396.
[11] I. Witten, R. Neal, J. Cleary, *Commun. ACM* **1987**, *30*, 520.
[12] K. A. Frenkel, *Commun. Psychopharmacol.* **1988**, *31*, 110.
[13] K.W. Brodlie, L.A. Carpenter, R.A. Earnshaw, J.R. Gallop, R.J. Hubbold, A.M. Mumford, C.D. Osland, P. Quarendon (eds), in *Scientific Visualization: Techniques and Applications*, Springer-Verlag, New York, NY, **1992**.
[14] J. Bergin, K. Brodie, M. Patiño-Martínez, M. McNally, T. Naps, S. Rodger, J. Wilson, M. Goldweber, S. Khuri, J. Sami and R. Jiménez-Peris, *et al*., **1996**, *8*, 192.
[15] G. M. Nielson, H. Hagen, H. Muller, *Scientific Visualization: Overviews, Methodologies, Techniques*, IEEE Computer Society, Washington, DC, **1997**.
[16] E. R. Tufte, *The Visual Display of Quantitative Information* (2nd edn), Graphics Press, Cheshire, **2001**.
[17] J. Cooper, *The Design Pattern: Java Companion*, Addison-Wesley, Boston, MA, **1998**.
[18] D. Geary, *Graphic Java 2: Swing*, vol. 2, Sun Microsystems, Palo Alto CA, **1999**.
[19] N. Winograd, R. M. Braun, *Spectroscopy* **2001**, *16*, 14.
[20] J. B. MacQueen, Some methods of classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, **1967**, 281.
[21] S. Johnson, *Psychometrika* **1967**, *32*, 241.
[22] A. Y. Ng, M. I. Jordan, Y. Weiss On spectral clustering: analysis and an algorithm, in *Advances in Neural Information Processing Systems*, vol. 14, MIT Press, Cambridge, MA, **2002**, pp 849.
[23] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo CA, **1993**.
[24] S. Wold, M. Sjostrom, *Am. Chem. Soc. Symp. Ser.* **1977**, *52*, 243.
[25] I. T. Jolliffe, *Principal Component Analysis* (2nd edn), Springer, New York, **2002**.
[26] W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective* (revised edn), Oxford University Press, New York, **1988**.
[27] X. Tian, S. E. Reichenbach, Q. Tao, A. Henderson, Classification and cluster analysis of complex time-of-flight secondary ion mass spectrometry for biological samples, *International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics (BCBGC-09)* ISRST, Worthington, OH, **2009**, pp. 78–85.

