Published in final edited form as:
Microsc Microanal. 2013 October ; 19(5): . doi:10.1017/S1431927613012737.

Rapid and Accurate Analysis of an X-Ray Fluorescence Microscopy Data Set through Gaussian Mixture-Based Soft Clustering Methods

Jesse Ward[1], **Rebecca Marvin**[2], **Thomas O'Halloran**[2,3], **Chris Jacobsen**[1,4], and **Stefan Vogt**[1,*]

[1]X-Ray Science Division, Advanced Photon Source, Argonne National Laboratory, Argonne, IL 60439, USA

[2]Department of Chemistry and Chemistry of Life Processes, Northwestern University, Evanston, IL 60208, USA

[3]Interdepartmental Biological Sciences, Northwestern University, Evanston, IL 60208, USA

[4]Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA

Abstract

X-ray fluorescence (XRF) microscopy is an important tool for studying trace metals in biology, enabling simultaneous detection of multiple elements of interest and allowing quantification of metals in organelles without the need for subcellular fractionation. Currently, analysis of XRF images is often done using manually defined regions of interest (ROIs). However, since advances in synchrotron instrumentation have enabled the collection of very large data sets encompassing hundreds of cells, manual approaches are becoming increasingly impractical. We describe here the use of soft clustering to identify cell ROIs based on elemental contents, using data collected over a sample of the malaria parasite *Plasmodium falciparum* as a test case. Soft clustering was able to successfully classify regions in infected erythrocytes as "parasite,""food vacuole,""host," or "background." In contrast, hard clustering using the *k*-means algorithm was found to have difficulty in distinguishing cells from background. While initial tests showed convergence on two or three distinct solutions in 60% of the cells studied, subsequent modifications to the clustering routine improved results to yield 100% consistency in image segmentation. Data extracted using soft cluster ROIs were found to be as accurate as data extracted using manually defined ROIs, and analysis time was considerably improved.

Keywords

X-ray fluorescence microscopy; cluster analysis; soft clustering; hard clustering; Gaussian mixture models; *k*-means clustering; expectation maximization; image segmentation; bioinorganic chemistry; *Plasmodium falciparum*

Introduction

Metal ions, nanoparticles, and metal–protein complexes are enormously important in biology and medicine. While they comprise only a small fraction of living organisms by weight (Kaim & Schwederski, 1994), transition metal ions such as iron, copper, and zinc

*Corresponding author. svogt@aps.anl.gov.

carry out diverse activities that are essential to life as we know it, including catalysis, electron transfer, structure stabilization, and oxygen transport. Heavy metals such as mercury and cadmium are important environmental contaminants and are toxic to organisms. Finally, certain metals such as platinum and gadolinium are not normally present in biological tissues but have useful medical applications in cancer treatment and imaging. Whether one studies mechanisms of metal homeostasis, heavy metal toxicity, or processing of metal-based drugs or contrast agents, knowledge of how the element of interest is spatially distributed in cells and tissues is an important clue to its mode of action.

While a variety of techniques exist for studying metals in biology, X-ray fluorescence (XRF) microscopy offers several advantages in that it can produce simultaneous, quantitative, and element-specific images of each element in the sample, at high resolution and sensitivity (Paunesku et al., 2006). The resolution of XRF microscopy depends on the optics, but is typically on the order of 30–200 nm. Since typical cells are 5–40 μm across with organelles on the order of 100–1,000 nm, XRF microscopy is well suited for answering questions of spatial distribution of metals in cells. Synchrotron radiation produces high-intensity X-rays that allow detection of a variety of elements of interest at parts per million or lower detection limits, which provides sufficient sensitivity to answer questions on both total metal levels and spatial distribution. The penetration depth of X-rays is greater than that of electron or proton probes, which allows analysis on whole-cell samples without the need for thin sectioning, permitting analysis of cells closer to their native state.

Trace elements are not distributed homogeneously through the cell; rather, chemical composition varies from organelle to organelle. Both plants and yeasts contain vacuoles used to store toxic heavy metals (Benavides et al., 2005). In eukaryotic cells, the nucleus contains a large portion of the cell's total zinc content, where zinc serves as a component of DNA-regulating zinc finger proteins (Vallee et al., 1991). Cellular iron is found primarily in iron/sulfur clusters within the mitochondria (Chitambar, 2005). The Golgi apparatus is a primary store for free calcium within the cell (Dolman & Tepikin, 2006). In plants, manganese is found primarily within chloroplasts, where it is present at the active site of the oxygen-evolving complex (Dismukes, 1986). Manganese is also commonly found in mitochondria, where it binds one form of superoxide dismutase (Keller et al., 1998). These metals are required for the organelles to carry out their essential functions; therefore, organelles can be characterized in part by their trace metal composition. When analyzing X-ray microscopy data, features in the elemental maps can be used to identify organelles. This allows subcellular analysis of element composition without the need for subcellular fractionation.

Currently, subcellular analysis of elemental maps is commonly carried out “by hand”: the researcher manually draws a region of interest (ROI) around a particular feature in the X-ray images, and the average per-pixel contents of an element (in $\mu g/cm^2$) within the ROI multiplied by the ROI area yields the total element content within that region (Vogt, 2003). Automating this analysis would offer several advantages. First, manual analysis can be tedious and time consuming, especially for large data sets. Improvements in source intensity as well as X-ray optics, faster detectors, and software developments that allow fly scanning capabilities have already enabled collection of very large XRF microscopy data sets: scans of millimeter length scale, but at submicron resolution, encompassing hundreds of cells of a variety of different types. As upgrades to existing technology continue to be made, manual segmentation of increasingly large images becomes untenable. Second, automating analysis has the advantage of removing ambiguity in ROI definitions. Two different researchers may disagree on the significance or spatial extent of a feature in an X-ray image, thus introducing errors when elemental analysis is performed; algorithm-driven analysis is expected to produce more consistent results.

We describe here the use of soft clustering methods for automatic feature identification and image segmentation. Conventional or "hard" clustering methods assign data points completely to one of several groups based on some measure of similarity between data points (Everitt, 1980), and have successfully been used in soft X-ray spectromicroscopy (Lerotic et al., 2004, 2005). Soft clustering relaxes group membership criteria by assigning data points to multiple groups simultaneously, with variable membership weights or probabilities based on similarity. This is appropriate for organelles in cells with nonuniform metal concentrations, and is required for two-dimensional (2D) projections of cells where there may be overlap of different organelles in a particular direction. Fluorescence tomography (de Jonge et al., 2010) can help resolve this latter ambiguity, but at present this is time consuming and has only been used in a few example studies.

The data set used in this study is a series of X-ray maps of human red blood cells infected with the parasite *Plasmodium falciparum*, the causative agent of malaria. *P. falciparum* is an attractive model organism for testing clustering methods, since it is already known that infection of an erythrocyte by *P. falciparum* causes shifts in cellular iron and zinc distributions that are necessary for parasite survival (Ginsburg et al., 1986; Marvin et al., 2012). Internalized parasites feed on the host cytoplasm, using hemoglobin as its primary source of amino acids (Francis et al., 1997). This process releases heme, which is toxic to the parasite because of its ability to generate superoxide anions (Muller, 2004). The parasite detoxifies heme by crystallizing it into a nontoxic hemozoin crystal within its food vacuole (Francis et al., 1997). The parasite also makes a series of modifications to the host membrane that allow it to import excess zinc, though the reason for this zinc accumulation is unknown (Ginsburg et al., 1986; Marvin et al., 2012). The accumulation of heme iron within the food vacuole and the accumulation of zinc within the parasite cause inhomogeneities in the elemental maps that can be taken advantage of to automatically identify cellular subregions (Marvin et al., 2012).

Materials and Methods

Sample Preparation

The 3D7 strain of *P. falciparum* was cultured in 10% human serum in RPMI-1640 (Trager & Jensen, 1976). Cultures were synchronized with 5% sorbitol, transferred to fresh culture media to 1.5% parasitemia, and grown at 37°C for 40 h. At 40 h, cultures were pelleted and washed in PBS, then re-suspended in PBS at 5% hematocrit. A 2 μL volume of suspended cells was spotted onto 200 nm thick silicon nitride windows obtained from Structure Probe Inc. (West Chester, PA, USA). After allowing the cells to settle, slides were blotted and immediately frozen by submerging in liquid nitrogen for 20 s. Samples were dehydrated by submerging in liquid acetone just above its freezing point for 20 s, followed by air drying.

XRF Imaging

XRF imaging was performed at Beamline 2-ID-E at the Advanced Photon Source at Argonne, IL, USA. A Leica optical microscope was used to locate appropriate cells for imaging (Leica Microsystems, Wetzlar, Germany). Undulator-produced X-rays at 10.4-keV incident energy were focused to a spot size of 0.5 μm horizontal by 0.3 μm vertical using Fresnel zone plate optics (Xradia Inc., Pleasanton, CA, USA). Samples were raster scanned through the X-ray beam using 0.5 μm horizontal and 0.3 μm vertical steps. Fluorescence spectra were collected for 1 s/pixel using a single-element Ge detector (Canberra Industries, Meriden, CT, USA). Data were collected over a series of $n = 27$ total cells.

Quantitative analysis of elemental contents was performed on a per-pixel basis using MAPS software (Vogt, 2003). Normalized fluorescence intensities for each element at each pixel

were converted to elemental contents in micrograms per square centimeter by fitting spectra to those obtained from the NIST thin-film standards NBS-1832 and NBS-1833 (National Institute of Standards and Technology, Gaithersburg, MD, USA).

Cluster Analysis

The general procedure for performing hard or soft clustering of an X-ray data set is outlined in Figure 1. XRF microscopy data sets consist of a series of N element-specific images. If i is the number of rows in each image, and j the number of columns, then $P = i \times j$ is the total number of pixels in the image. Let each elemental image in the data set be indexed by the number $n = 1,\ldots, N$, and let each pixel be indexed by the number $p = 1,\ldots, P$. In that case, the data set can be represented as a $P \times N$ matrix, where each row represents the elemental composition of pixel p, and each column represents a reshaped image of element n. Another way to represent this data set is as a scatter plot with N dimensions and P points, where each point characterizes the elemental composition of a particular pixel. Two scatter plot points located in close proximity have similar elemental compositions and may be expected to represent fluorescence signal derived from the same organelles within the cell. Alternatively, two points located distant from each other would have distinct relative elemental compositions and would be expected to represent signal derived from separate organelle types. Cluster analysis is used to group together pixels with similar elemental compositions into G groups based on some distance measure.

Soft clustering is accomplished by fitting the N-dimensional scatter plot to a series of G Gaussians using the expectation maximization algorithm (Moon, 1996). The output is a $P \times G$ matrix, where each matrix element represents the intensity of normalized Gaussian $g = 1, \ldots, G$ at pixel p. The relative "ownership" of each pixel by each cluster can be calculated by normalizing across each row p of the $P \times G$ matrix, and each ROI calculated from the soft clustering routine can be plotted by reshaping each column g into an $i \times j$ ROI image matrix. Entries of each of the $i \times j$ ROI image matrices can take on any value between 0 and 1, depending on the relative ownership of a particular pixel p by a particular ROI g.

For the purpose of this study, images were divided into four ROIs based on the iron and zinc distributions. The average 2D elemental contents in micrograms per square centimeter was calculated for each element in each ROI by multiplying element matrix n by ROI matrix g entrywise, summing all the entries in the resulting matrix, and normalizing to the sum of entries in ROI matrix g. The background ROI was defined as the ROI with the lowest average per-pixel iron content, and subsequent inspection of the calculated ROIs showed that this was a reasonable assumption. The average iron and zinc per-pixel contents over the background ROI were subtracted from the average per-pixel contents in the other ROIs in the set, so that automatic background subtraction could be achieved. Total elemental contents in each ROI were calculated by multiplying the background-subtracted 2D elemental contents (in $\mu g/cm^2$) by the effective ROI area, defined as the sum of all the entries in ROI matrix g multiplied by the area of a single pixel (determined by the raster scan parameters). In some cases, an elemental image would contain not only the cell of interest, but the edge of an adjacent cell. In these cases, quantification was carried out only over the cell of interest by manually defining a large ROI that encompasses both the cell of interest and the background, while excluding the edge of the neighboring cell.

Initially, starting parameters (means, standard deviations, and covariances) for the Gaussians used in the scatter plot fit were chosen at random. Since there is no guarantee that multiple fits for the same data set will converge on the same result if the initial values are chosen at random, consistency was checked by running each cell through the soft clustering algorithm at least ten times and checking for the existence and relative frequencies of multiple solutions. Later on, consistency was improved by initializing the fits with more appropriate

starting parameters. This was achieved by creating a "master" scatter plot using all the pixels of all the images in the data set (all 27 cells), performing a fit of G Gaussians to the master scatter plot, and using the results of this fit as the initial starting parameters for fits for individual cells. Consistency was evaluated for the fits to both the master scatter plot and the individual cells as before.

For comparison, hard clustering was performed on the data scatter plot using the *k*-means clustering algorithm (Everitt, 1980). To cluster the data into G groups, G points in the data scatter plot are chosen by random, and the Euclidean distances between all the points in the data scatter plot and each of the G cluster centers is calculated. Ownership of a particular pixel p by cluster g is assigned by whichever distance is the smallest. Once clusters are determined, the cluster centers are recalculated based on the current cluster assignment, and the process is repeated iteratively until convergence is reached. The output is a $P \times G$ matrix, where each matrix element can take on a value of either 0 or 1 depending on whether or not pixel p is closest to cluster center g. Each column g can be reshaped into an $i \times j$ matrix and plotted to visualize the ROI.

Finally, clustering results were compared with the results of manual X-ray image segmentation. An ROI was drawn around each cell, and the iron distribution was used to identify an ROI for the food vacuole. In addition, an ROI was drawn around a large portion of the background for each image. As before, background subtraction was achieved by subtracting the average elemental per-pixel contents over the background ROI from the average elemental per-pixel contents in the cell and food vacuole ROIs. Total elemental contents within each ROI were calculated by multiplying the background-subtracted per-pixel elemental contents (in $\mu g/cm^2$) by the ROI area.

Results

Representative light micrographs and elemental distributions for several malaria-infected erythrocytes are shown in Figure 2. The dark spot in each of the light micrographs is the hemozoin crystal, which overlaps with the bright spot in the Fe distribution. The hemozoin crystal is formed in the food vacuole as heme is released from digested hemoglobin. In addition, one can observe aggregation of P and Zn in their respective maps, close to but not co-localized with the bright spot in the Fe distribution. The high phosphorous content suggests the presence of DNA, which is not present in the uninfected erythrocyte host, but is actively synthesized by the parasite as it replicates. The uninfected host has a baseline level of zinc, but this increases by over 400% in the course of the infection cycle, and the zinc imported by the parasite serves as a marker for parasite position (Marvin et al., 2012).

Each of the 27 cells in the data set were divided into four clusters using both soft and hard clustering methods, using the Zn and Fe distributions as input. Typical output for both soft and hard clustering is shown in Figures 3a and 3b, respectively. The four soft clusters obtained can be loosely interpreted as corresponding to a "food vacuole" region, a "parasite" region (i.e., those portions of the parasite growth compartment that are outside of the food vacuole), a "host" region, and a "background" region. Hard clustering of the raw data through *k*-means is usually observed to divide the food vacuole into two regions based on iron content and is able to define the parasite region through the zinc content, but does not typically distinguish between "host" and "background" regions (Fig. 3b). The inclusion of other biologically relevant elements, such as phosphorous, into the scatter plot for fitting failed to improve the quality of image segmentation (data not shown). Hard clustering results can be slightly improved in most cases (~74%, or 20 out of 27) by normalizing the elemental data to either their standard deviations or their ranges (the difference between maximum and minimum values), in which case one obtains four regions comparable to what

one obtains through soft clustering methods (Fig. 3d). However, the hard clustering algorithm still has difficulty distinguishing cell from background, and in 26% of the cases (seven out of 27), no improvements in image segmentation were observed. In contrast to the hard clustering case, normalizing the elemental data to standard deviation or range did not significantly improve the soft clustering output in any of the cells studied (Fig. 3c). Subsequent analyses were carried out using the soft clustering ROIs alone.

When initial values for the Gaussians in the scatter plot fits are chosen at random, the fit was observed to converge on two or three solutions in 60% of cases (16 out of 27), although in 14 of the 16 cases, the algorithm tended to converge on one particular result much more frequently (≥70% of the time). Figure 4a shows an example of this inconsistency between trials. In order to improve fit consistency, a “master” scatter plot was defined using all of the per-pixel data in the entire data set (all 27 cells), a four-Gaussian fit was performed on this master scatter plot, and the fit results were used to define ROIs in the individual cells without further fitting. While this approach vastly improved consistency (<5% variation between fit results when fitting to the master scatter plot), the resulting Gaussian parameters were not appropriate in all cases for defining a meaningful division of the data. For example, in Figure 4b, one can distinguish “food vacuole” and “parasite” ROIs; however, the “host” ROI includes significant contribution from the background. In order to provide enough flexibility to accommodate differences in individual images while improving fit consistency, fits were made to the individual cell scatter plots using the results of the fit to the master scatter plot to define the initial fit parameters. When this is done, the ROIs again define “food vacuole,” “parasite,” “host,” and “background” regions (Fig. 4c), while consistency is much improved—in this case 100% consistency (ten out of ten trials) for all 27 cells in the data set. Initializing the scatter plot fits this way had an additional advantage in that the cluster assignments were consistent from cell to cell (e.g., “cluster 1” could be designated the “food vacuole” cluster for all cells in the data set), whereas with random seeding, cluster assignments have to be made after the fact (e.g., “cluster 1” for the first cell might correspond to the “food vacuole”, while “cluster 1” for the second cell may represent “parasite”).

To check the accuracy of metal quantification through soft ROI fitting versus manual ROI definition, two figures of merit were chosen: total zinc per cell (to check the ability of soft ROIs to distinguish cell from background), and the percentage of total iron within the hemozoin crystal (to check the ability of soft ROIs to distinguish prominent organelles from the rest of the cell). These values were calculated for each cell using both methods, and a scatter plot showing the results of manual ROI definition versus soft ROI calculation is shown in Figure 5a for total zinc and Figure 5b for percentage of hemozoin iron. In both cases, the scatter plots approximately define a straight line with a slope close to 1, indicating that quantification through soft ROI calculation produces similar results to quantification using manually drawn ROIs in all cases.

Discussion

As advances in synchrotron beamline instrumentation continue to be made, the amount of data one can collect during a single experiment will continue to increase, as will the time and complexity of data analysis. To keep up with this demand, it is necessary to develop new software tools for rapid, accurate analysis. Presented above is a method for dividing large XRF microscopy data sets into biologically relevant subregions based on their elemental contents. Since the elemental contents vary from organelle to organelle and since the trace element levels in a given organelle are essential to that organelle's function, defining organelle positions through the elemental contents of each pixel is a reasonable

approach. In the example presented here, the images could be divided into regions corresponding to food vacuole, parasite, host, and background.

Initially, we found that the fitting algorithm would converge on multiple solutions for many of the cells under study, when the starting values were chosen at random. This is a common problem for numerical fits with randomly initialized values. While the algorithm may converge on a solution, one cannot be certain whether the algorithm converged on a global minimum or a local minimum without repeating the fit multiple times. Here, the problem was addressed by defining a master scatter plot built from all the images in the data set. If an organelle's elemental content is related to its function, then one can imagine that each organelle has an ideal range of probable values for each trace metal. Although the exact contents may vary from cell to cell, organelles of a given type could be expected to conform to a certain range of metal content, with a certain average and standard deviation over the data set. The scatter plot for an individual cell would be like a random sampling of this variation for the different organelle types. By defining the scatter plots for multiple cells simultaneously, one can obtain a distribution of metal levels that more closely matches the "true" distribution of metal levels within a certain organelle type. The master scatter plot can be considered a larger sample of the "true" distribution, and when one performs a fit to this scatter plot, it is more likely that the results of this fit will be closer to the "true" solution. When fits to the individual cell scatter plots are performed using the results of the fit to the master scatter plot as the initial values, the initial values are closer to the global minimum of the fit, and more consistent results are obtained.

Soft clustering of the X-ray data was hypothesized to produce a more accurate division of the data compared with "hard" clusters, where each pixel is assigned 100% to a given cluster. This is because the signal from a given pixel may have contributions from multiple biologically relevant subregions because of overlap of these subregions in the 2D image. A soft clustering approach is a promising way to disentangle the contributions from different subregions to a given pixel. Comparing the quantification over soft clusters to quantification using manually drawn ROIs (Fig. 5), one can observe that the two values are similar for all cells in the data set. Thus, it is not clear from this data set that soft clustering is more or less accurate than manual ROIs. However, the time savings of performing soft clustering as opposed to manual ROI definition are substantial.

In the work presented here, soft clusters were defined through Gaussian mixture models. However, it should be noted that this is not the only way to achieve a soft clustering of scatter plot data. An alternative method, fuzzy *c*-means clustering (FCM), is also commonly used to assign fractional membership (Bezdek, 1981). The FCM algorithm has features in common with *k*-means clustering. Cluster centers are randomly initialized and data points are assigned to clusters based on some distance measure. The cluster centers are recalculated based on the data point assignment, and the data point assignment and cluster center recalculation steps are alternated until convergence is achieved. FCM differs from *k*-means clustering, however, in that each data point is assigned a fractional membership for each cluster center, defining a $P \times G$ membership matrix (P pixels and G clusters) which gets updated at each step. The calculation of the membership matrix and the cluster centers depend strongly on a "fuzzy weighting exponent" parameter. The fuzzy weighting exponent can take on any value between 1 and ∞, and controls the degree of "fuzziness" of the clusters. This parameter has to be chosen by the researcher before performing FCM, and the interpretation of the results can change depending on its value (Stork & Keenan, 2010). In practice, the optimal value of the fuzzy weighting exponent depends on the data set, and several different values have to be tested. In our view, one of the advantages of using a Gaussian mixture model to define soft clusters rather than FCM is the lack of *ad hoc* parameters, facilitating unsupervised data analysis.

One of the limitations of the method presented here is that in cases where a second cell is captured in the XRF image along with the cell of interest (see, e.g., the upper left-hand side of the images in Fig. 4), the cell has to be manually edited out before quantification for accurate analysis. While this did not significantly slow down total analysis time for the cells in this study, since regions with well-separated cells were chosen before collection of XRF images, this method would be inappropriate for the analysis of denser cell cultures or wider-area images. One of the challenges in the future will be to combine this method with a method for automatically identifying multiple cell ROIs simultaneously.

We have not yet found a clear method for automating the determination of the appropriate number of clusters. In the study presented here, four clusters were chosen because the calculated ROIs appeared to be reasonable based on comparison to the X-ray maps and what was known about the parasite biology. It would be desirable to have some metric to suggest to the researcher when the appropriate number of clusters has been chosen in the more general case where less is known about the sample of interest. In some cases, it may be advantageous to combine clustering methods with principal components analysis (PCA). PCA can be used to divide a set of hyperspectral images into a ranked set of eigenimages that account for decreasing amounts of variance in the data set (Keenan & Kotula, 2004; Stork & Keenan, 2010). The eigenimages can be divided into two groups: a set of "principal component" images that encode the "real" chemical variations in the sample, and a set of images that are essentially because of experimental noise. The original data can be reconstructed with little loss in information by excluding the noise eigenimages, and in theory the number of principal components identified can serve as a useful starting point when deciding on the appropriate number of clusters to use. In practice, it can be difficult to decide where to draw the line between principal components and noise eigenimages. The number of principal components to include is commonly determined by plotting the component rank versus the component eigenvalue (Keenan & Kotula, 2004). The component eigenvalues will initially decrease sharply as a function of rank, eventually leveling off to a constant value, and the position of the "elbow" in this plot will determine the number of components. In data sets with nonuniform noise, the position of this transition will blur. Such is the case with XRF microscopy, where variance is proportional to signal intensity because of Poisson statistics. In this case, it will be necessary to pre-treat the data to correct for Poisson noise, and when this is done, the division between "real" and noise eigenimages becomes less ambiguous (Keenan & Kotula, 2004). On the other hand, even if a clear division between "real" and noise eigenimages can be achieved, the number of chemically relevant components the researcher may be interested in can be greater than the number of components identified by PCA. PCA will produce the most parsimonious division of the data set—the minimum number of factors needed to account for all the variation. However, it is possible for one region in a series of images to possess unique physical and chemical properties, while being compositionally linearly related to adjacent regions. In one such case, Stork and Keenan performed energy-dispersive X-ray spectroscopy on a solder bump sample composed of Cu, Sn, and a Cu–Sn intermetallic region. The Cu–Sn intermetallic region is brittle compared with pure Cu or Sn, and therefore of special interest to materials scientists. Principal components analyses of this data set yielded two significant eigenimages which correspond to the Cu and Sn distributions; however, when soft clustering was performed, the Cu–Sn intermetallic phase could be independently identified (Stork & Keenan, 2010). Thus, when performing PCA to estimate the number of clusters one should use, the result obtained should probably be considered a lower bound to the number of chemically relevant components. The problem of deciding on the appropriate number of clusters is still an active area of research, and ultimately, the researcher must still exercise judgment in deciding when an appropriate division has been reached.

Summary

We have developed an approach to XRF microscopy data analysis that automatically segments images into biologically relevant subregions based on their elemental contents. The method is as accurate as manual image segmentation, but is significantly faster to perform. The potential drawback of the algorithm converging on multiple solutions was avoided by using the results of a fit using the entire data set to seed the initial values for the fits to individual cells. This approach will allow the speed of data analysis to keep pace with the increased throughput of modern XRF microscopy experiments.

Acknowledgments

This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences program under contract DE-AC02-06CH11357 and by the National Institutes of Health grant GM038784-22S1.

References

Benavides MP, Gallego SM, Tomaro ML. Cadmium toxicity in plants. Braz J Plant Physiol. 2005; 17:21–34.

Bezdek, JC. Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers; Norwell, MA: 1981.

Chitambar CR. Cellular iron metabolism: Mitochondria in the spotlight. Blood. 2005; 105(5):1844–1845. [PubMed: 15747401]

de Jonge MD, Holzner C, Baines SB, Twining BS, Ignatyev K, Diaz J, Howard DL, Legnini D, Miceli A, McNulty I, Jacobsen CJ, Vogt S. Quantitative 3D elemental microtomography of *Cyclotella meneghiniana* at 400-nm resolution. Proc Nat Acad Sci. 2010; 107(36):15676–15680. [PubMed: 20720164]

Dismukes GC. The metal centers of the photosynthetic oxygen-evolving complex. Photochem Photobiol. 1986; 43(1):99–115.

Dolman NJ, Tepikin AV. Calcium gradients and the Golgi. Cell Calcium. 2006; 40(5–6):505–512. [PubMed: 17023044]

Everitt, B. Cluster Analysis. Wiley; New York: 1980.

Francis SE, Sullivan DJ, Goldberg ADE. Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*. Annu Rev Microbiol. 1997; 51(1):97–123. [PubMed: 9343345]

Ginsburg H, Gorodetsky R, Krugliak M. The status of zinc in malaria (*Plasmodium falciparum*) infected human red blood cells: Stage dependent accumulation, compartmentation and effect of dipicolinate. Biochim Biophys Acta. 1986; 886(3):337–344. [PubMed: 3518809]

Kaim, W.; Schwederski, B. Bioinorganic Chemistry: Inorganic Elements in the Chemistry of Life: An Introduction and Guide. John Wiley & Sons; San Francisco, CA: 1994.

Keenan MR, Kotula PG. Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images. Surf Interface Anal. 2004; 36(3):203–212.

Keller JN, Kindy MS, Holtsberg FW, St. Clair DK, Yen H-C, Germeyer A, Steiner SM, Bruce-Keller AJ, Hutchins JB, Mattson MP. Mitochondrial manganese superoxide dismutase prevents neural apoptosis and reduces ischemic brain injury: Suppression of peroxynitrite production, lipid peroxidation, and mitochondrial dysfunction. J Neurosci. 1998; 18(2):687–697. [PubMed: 9425011]

Lerotic M, Jacobsen C, Gillow JB, Francis AJ, Wirick S, Vogt S, Maser J. Cluster analysis in soft X-ray spectromicroscopy: Finding the patterns in complex specimens. J Electron Spectros Relat Phenomena. 2005; 144–147:1137–1143.

Lerotic M, Jacobsen C, Schäfer T, Vogt S. Cluster analysis of soft X-ray spectromicroscopy data. Ultramicroscopy. 2004; 100(1–2):35–57. [PubMed: 15219691]

Marvin RG, Wolford JL, Kidd MJ, Murphy S, Ward J, Que EL, Mayer ML, Penner-Hahn JE, Haldar K, O'Halloran TV. Fluxes in “free” and total zinc are essential for progression of intraerythrocytic stages of *Plasmodium falciparum*. Chem Biol. 2012; 19(6):731–741. [PubMed: 22726687]

Moon TK. The expectation-maximization algorithm. IEEE Signal Process Mag. 1996; 13(6):47–60.

Muller S. Redox and antioxidant systems of the malaria parasite *Plasmodium falciparum*. Mol Microbiol. 2004; 53(5):1291–1305. [PubMed: 15387810]

Paunesku T, Vogt S, Maser J, Lai B, Woloschak G. X-ray fluorescence microprobe imaging in biology and medicine. J Cell Biochem. 2006; 99(6):1489–1502. [PubMed: 17006954]

Stork CL, Keenan MR. Advantages of clustering in the phase classification of hyperspectral materials images. Microsc Microanal. 2010; 16(6):810–820. [PubMed: 20964877]

Trager W, Jensen J. Human malaria parasites in continuous culture. Science. 1976; 193(4254):673–675. [PubMed: 781840]

Vallee BL, Coleman JE, Auld DS. Zinc fingers, zinc clusters, and zinc twists in DNA-binding protein domains. Proc Natl Acad Sci USA. 1991; 88(3):999–1003. [PubMed: 1846973]

Vogt S. MAPS: A set of software tools for analysis and visualization of 3D X-ray fluorescence data sets. J Physiq (Proc). 2003; 104(2):635–638.

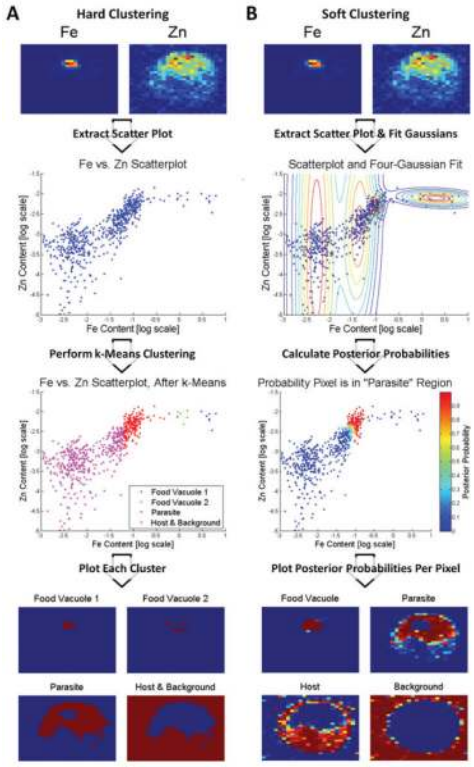

Figure 1.

Work flow for performing hard or soft clustering of X-ray fluorescence imaging data. Each column shows the raw Fe and Zn maps for a particular cell in the data set (top panel), a scatter plot of Fe versus Zn contents in each pixel (second panel from top), the same scatter plot after sorting the pixels through hard or soft clustering methods (third panel from top), and the region of interest (ROIs) calculated from clustering results (bottom panel). **A:** Outline of hard clustering. The iron and zinc maps are used as inputs, and a two-dimensional scatter plot is defined, consisting of 744 individual points, one per pixel. Data segmentation is performed using *k*-means clustering over four groups. ROI images are defined using the pixel indices: if a given cluster g contains a pixel with index p, the corresponding pixel in that cluster's ROI image has a value of 1; otherwise, it is 0. **B:** Outline of soft clustering. The scatter plot of indexed pixels is defined as in the hard clustering example, but instead of using *k*-means clustering to segment the data into four groups, a four-Gaussian model is defined, and the best fit of the model to the scatter plot data is determined using the expectation maximization algorithm. The posterior probability that each pixel p is a member of each group g is calculated. This number varies between 0 and 1, inclusive. A scatter plot showing the posterior probabilities that each of the pixels in the scatter plot belong to the "parasite" ROI is shown in this example; for the sake of clarity, the scatter plots for the other ROIs have been omitted. The soft ROIs are imaged by plotting the posterior probabilities that each pixel p belongs to a given group g.

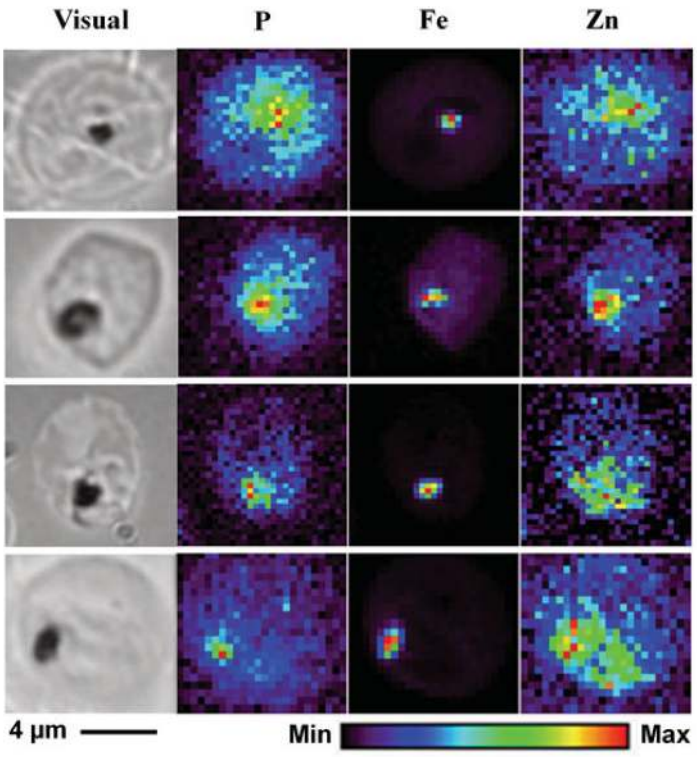

Figure 2.
Representative visible micrographs and phosphorous, iron, and zinc distributions for various infected erythrocytes.

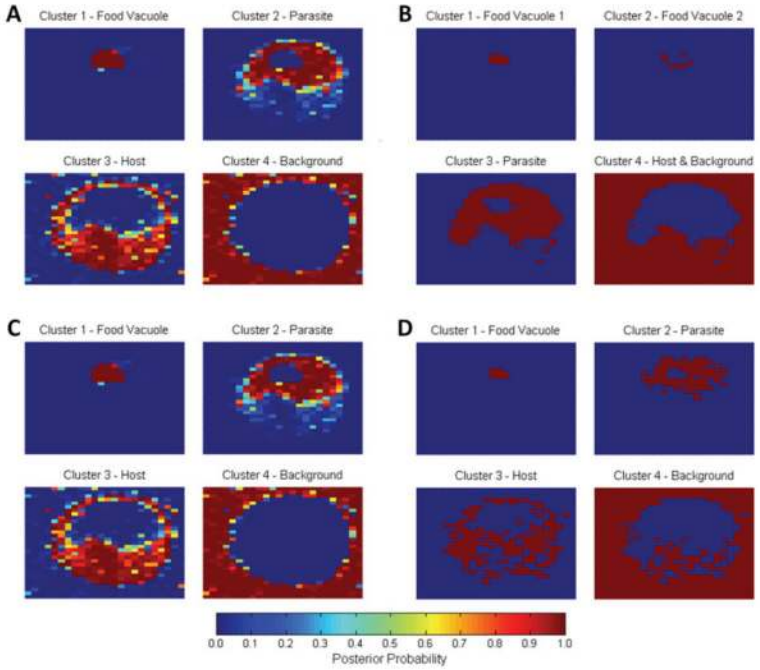

Figure 3.

Results of various soft clustering and hard clustering methods performed on a representative infected erythrocyte, using the iron and zinc distributions as inputs. The colorbar (at bottom) indicates the probability that a particular pixel belongs to a particular region of interest (ROI). **A:** Results of soft clustering performed on raw, un-normalized data, using a four-Gaussian model. The plotted ROIs can be interpreted as "food vacuole," "parasite," "host," and "background" regions. **B:** Results of hard clustering performed on raw, un-normalized data, using *k*-means clustering with four cluster centers. Unlike the soft clustered data, *k*-means clustering divides the food vacuole into two regions, and combines the host and background into a single region. **C:** Results of soft clustering performed on data normalized to standard deviation, using a four-Gaussian model. The results are qualitatively equivalent to the ROIs defined in (**A**). **D:** Results of hard clustering performed on data normalized to standard deviation, using *k*-means clustering with four cluster centers. In this case, the ROIs can be roughly interpreted as "food vacuole,""parasite,""host," and "background," as in (**A**), although there is still significant overlap between the "host" and "background" ROIs.

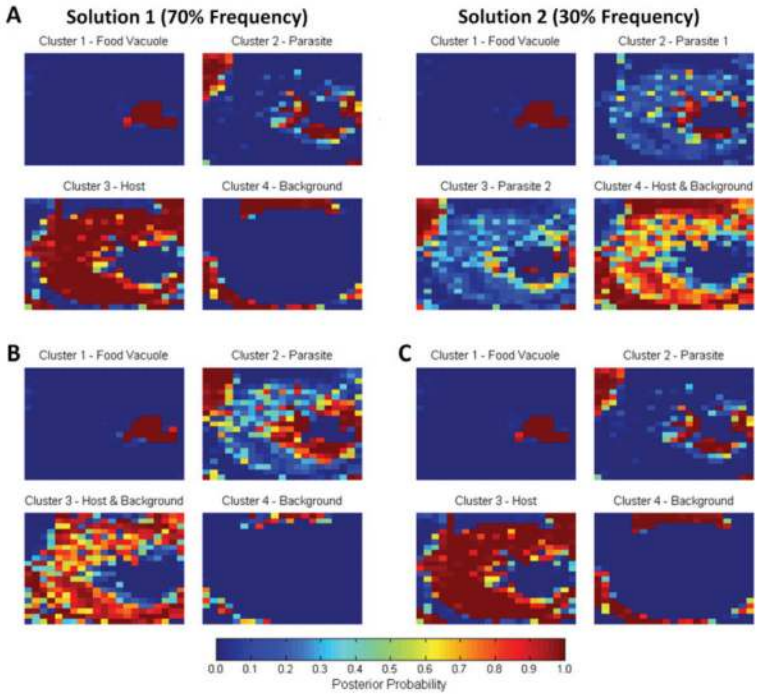

Figure 4.

Soft clustering results for an infected erythrocyte. The iron and zinc distributions were used as inputs and fit to a four-Gaussian model, with various choices for starting parameters. The colorbar indicates the probability a particular pixel belongs to a particular region of interest (ROI), as in Figure 3. **A:** Soft clustering with random starting parameters. In 70% of the trials, the image is successfully divided into "food vacuole," "parasite," "host," and "background" regions. However, in 30% of cases "host" and "background" were poorly distinguished, and the "parasite" region was divided into two ROIs. **B:** Soft clustering using the results of the fit to the "master" scatter plot (consisting of the Fe and Zn data from all the pixels in all 27 cells in the data set), with no additional fitting to the scatter plot for the individual image. In this case, the cell is not distinguished from the background, although "parasite" and "food vacuole" ROIs can be observed. **C:** Soft clustering results when performing a fit to the individual image, using the results of the fit to the master scatter plot as the initial values. In this case, ROIs that are straightforward to interpret physically are obtained 100% of the time.

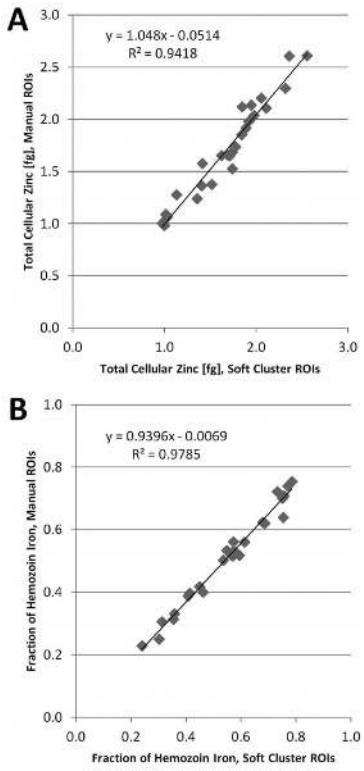

Figure 5.

A: For each cell in the X-ray fluorescence microscopy data set, total cellular zinc was calculated using data extracted from X-ray images by regions of interest (ROIs) found through soft clustering methods, and compared with data extracted using hand-drawn ROIs. Each data point represents a single cell. The data are highly correlated, with a slope close to 1, indicating that soft clustering methods are able to extract total zinc in this data set at least as accurately as hand-drawn ROIs are able to. **B:** Scatter plot comparing the percentage of total iron present in the hemozoin crystal, for data extracted using soft cluster ROIs versus data extracted using manually defined ROIs. The data are once again highly correlated, indicating that soft clustering methods can calculate iron within cellular subregions as accurately as hand-drawn ROIs.