A Visual Analysis System for Metabolomics Data

Philip Livengood*, Ross Maciejewski*, Wei Chen†, David S. Ebert*
* Purdue University Visualization and Analytics Center
† State Key Lab of CAD&CG, Zhejiang University

ABSTRACT

When analyzing metabolomics data, cancer care researchers are searching for differences between known healthy samples and unhealthy samples. By analyzing and understanding these differences, researchers hope to identify cancer biomarkers. In this work we present a novel system that enables interactive comparative visualization and analysis of metabolomics data obtained by two-dimensional gas chromatography-mass spectrometry (GCxGC-MS). Our system allows the user to produce, and interactively explore, visualizations of multiple GCxGC-MS data sets, thereby allowing a user to discover differences and features in real time. Our system provides statistical support in the form of mean and standard deviation calculations to aid users in identifying meaningful differences between sample groups. We combine these with multiform, linked visualizations in order to provide researchers with a powerful new tool for GCxGC-MS exploration and bio-marker discovery.

1 INTRODUCTION

In recent years, GCxGC-MS has become an invaluable laboratory analysis tool. However, this procedure produces large (gigabytes of data per sample), four dimensional datasets (retention time one, retention time two, mass and intensity). Such data is cumbersome, and researchers must spend time formatting and processing the data in order to remove acquisition artifacts, and quantify and identify chemical compounds [9]. Furthermore, while statistical analysis has played an important role in this work (because of the need to reduce the thousands of acquired spectral features to a more manageable size), the large data size, inherent biological variability and measurement noise makes the identification of bio-markers through purely statistical processes extremely difficult and time consuming.

As such, our work focuses on the development of a visual analysis suite for exploring differences across samples and groups of samples. In this work, we have created a system that utilizes a series of multiform linked visualizations to enable interactive exploration, filtering and comparison of multiple samples simultaneously. This allows users to quickly locate feature similarities and differences across samples and drill down into the mass spectrum data for detailed analysis. These techniques allow researchers to form and explore hypotheses about sample alignment, acquisition artifacts, and (most importantly) cancer bio-marker sites. For example, Figure 1 (top) illustrates a cancerous and non-cancerous data sample visualized with total ion count (TIC) images. In this example, differences between the two samples are not obvious, and, once differences are found, it can be difficult to determine whether or not those differences are meaningful. By using our tools, researchers are able to quickly identify and explore sample differences as seen in Figure 1 (bottom).

Our work is being developed in collaboration with analytical chemists and biology researchers, and is designed to provide an interactive comparative visual analysis environment for GCxGC-MS data exploration. While this work is similar to previous work [1, 2, 7] in that we provide linked views that support the interactive visual exploration of mass spectrometry data, our system provides several features not currently available in other GCxGC-MS analysis systems:

1. a comparative visualization window that allows multiple samples (and multiple views of individual samples) to be displayed simultaneously,
2. data exploration tools for exploring mass spectra and filtering and comparing TIC images in real-time,
3. grouping of samples and calculation of group means for comparison and difference calculation,
4. the application of mean and standard deviation TICs to the color-mapping of difference measures,
5. a dynamic color scale adaptation tool for discovering differences in low-intensity peaks.

As automated analysis tools take days to run, our system serves as a front end in order to make the calculation tractable by specifying a region for evaluation. With our tool, users may visualize multiple samples of GCxGC-MS data, interactively search for sample differences, probe areas of interest to bring up mass spectrum plots, and compare regions where the mass spectra deviate the most. These features provide valuable data analysis for bio-marker discovery, and make an excellent complement to existing workflows.
2 GCxGC-MS

In cancer care engineering, researchers collect GCxGC-MS data in order to search for biomarker differences between cancerous and non-cancerous samples. In GCxGC-MS, a sample to be analyzed is first mixed with a carrier gas that transports it through the machine. The device has two columns through which a sample passes before being analyzed by a mass spectrometer. Different components of the sample move through the columns at different rates, resulting in two levels of separation according to how long it takes each of the components to move through each of the columns. The mass spectrometer gives an additional level of separation as the mixture is ionized when it exits the second column. This process results in a four-dimensional dataset with two time axes (retention time one (RT1) from the first column and retention time two (RT2) from the second column), mass, and intensity.

Up to this point we have referred to mass as one of the dimensions of data produced by the mass spectrometer. More correctly, we should be referring to this as the mass-to-charge ratio, designated m/z. A mass spectrometer breaks molecules up into ionized fragments. It is then able to determine the mass-to-charge ratio by using the fact that two particles with the same m/z will move in the same path in a vacuum when subjected to the same electrical and magnetic fields. A particular chemical compound will have a very specific mass spectral signature. The spectral signature of unknown compounds can be compared to a library of known compounds to help identify the compound or its composition. This type of identification can be difficult because different chemical compounds may have the same retention times, and thus their spectral peaks may have significant overlap, a problem known as co-elution.

GCxGC-MS is marked by its large peak capacity, an order of magnitude increase in chemical separation ability [1], and its improved speed [3]. However, the data obtained exhibits several complexities. Samples must be mixed with a carrier gas as part of the acquisition process. This carrier gas shows up in the resulting data set, and can obscure sample peaks of interest. Inconsistencies in sample amounts will lead to differences in peak intensities. Peak retention times and shapes exhibit slight differences that are uncontrollable, yet are unrelated to actual chemical differences in the samples [1]. Data samples exhibit background noise that can vary from sample to sample, and make accurate peak quantification difficult. Finally, despite the improved separation ability over one-dimensional GC-MS, GCxGC-MS still exhibits some peak overlap due to coelution.

Despite these difficulties, GCxGC-MS is used in a wide array of applications, such as quality control, chemical identification, and biomarker detection. As a result, much research has been done to determine effective ways to analyze the large amounts of data produced and overcome the previously mentioned pitfalls. Alignment algorithms [14], background removal algorithms [8], and compound identification [10] are common areas of study. Other researchers have worked on identifying potentially significant compounds, such as the orthogonal partial least-square (OPLS) approach used by Wiklund et al. [13].

3 RELATED WORK

The thrust of this work is the development and deployment of an interactive visual analysis tool for analyzing data captured through the GCxGC-MS process. Kincaid and Dejgaard [6] developed a system to explore protein complexes in tandem mass spectrometry data through modifications to a scatter plot visualization. Jourdan et al. [5] also utilized scatterplot matrices and parallel coordinate plots as a means of exploring metabolomics data. Work by de Corral and Pfister [2] presented a system for the three-dimensional visualization of Liquid Chromatography - Mass Spectrometry (LCMS) data. The goal of this system was to provide LCMS practitioners with an overview of their data set and enable them to identify features that can vary with mass and time, and this research primarily focused on rendering speeds of large scale LCMS data. Work by Linsen et al. [7] also focused on LCMS data looking at characteristic patterns of isotopes. Their visual exploration tools allowed users to click on labeled peaks to explore mass spectrum data similar to our proposed system. However, the key difference lies in the ability of the user to compare multiple samples simultaneously through difference mappings and linked views.

Perhaps the closest is the work done by Hollingsworth et al. [1], in which they utilize image processing based techniques based on the total ion count image. Prior to visualization and comparison each image must undergo background removal and peak detection. Once this pre-processing is complete, a difference image is computed. This difference can be visualized using tabular data, a 2D image, or 3D height field visualization in either grayscale or color. Their main contribution is the calculation of a fuzzy difference that compares each pixel value in one image with a small neighborhood of pixels in the other, rather than doing a pixel-by-pixel comparison. This technique helps to reduce the incidental differences in peak shape and retention time that may still exist even after alignment. While these techniques represent a good first step, they are ill-suited for biomarker detection as they rely solely on images produced from the total ion count data, meaning that peaks of interest could easily become obscured. Since only the magnitude of the difference is considered, the results obtained can be misleading. Large, yet unimportant differences may be emphasized while small, yet meaningful differences may not be noticeable. Additionally, there is only limited user interaction available for exploring peaks and mass spectra.

4 COMPARATIVE VISUAL ANALYTICS SYSTEM

Our work provides a comparative visual analytics system for GCxGC-MS data. The primary goal is to increase data exploration and analysis in order to help researchers determine meaningful differences between groups of samples. Finding such differences is the first step in bio-marker identification. We provide both new and traditional visualization techniques and couple them with linked, interactive exploration. By incorporating mean and standard deviation calculations we are able to visualize the significance of differences in peak intensity rather than merely the magnitude of those differences.

In this section, we first discuss our interactive visualization methods. We describe our scheme for total ion count visualization, mass spectrum visualization and the various applied color mapping functionalities. We then discuss the linked views and data exploration tools including the area selection tool, mass filtering and TIC lens.
4.1 Visualization Methods

Our system allows researchers to view multiple samples, as well as display multiple views of individual samples. Each view can be configured with a different visualization that provides researchers with a different perspective on the data. The visualization methods include 2D total ion count visualization, mass spectrum visualization, and comparative visualization.

4.1.1 Total Ion Count (TIC) Visualization

The total ion count is a reduction of the data set from four dimensions down to three. For each retention time coordinate, the intensities of the entire mass spectrum at that point are summed together to obtain the total intensity, or total ion count. Once the TIC data has been computed, it can be visualized either in 2D, or as a 3D height rendering.

2D TIC: The two-dimensional total ion count image is one of the most common visualization techniques for GCxGC-MS data. Ion count values are mapped to a color using the specified transfer function and rendered to the screen as a flat, two-dimensional image, as in Figure 1.

3D TIC: A total ion count height rendering is nearly identical to the two-dimensional total ion count image. For the height rendering, the intensity is used as the z-coordinate in a polygonal mesh, and can be scaled linearly or logarithmically to fit within a reasonable dimension. The x and y coordinates are evenly spaced points corresponding to retention time 1 and retention time 2. When used in conjunction with a normal color mapping, this does not actually convey any more information than a two-dimensional TIC image (Figure 2, left). However, this is still a useful technique as data can often be portrayed more effectively by mapping the data values to multiple display parameters, in this case color and height. Not only do the two parameters serve to reinforce each other, but one may overcome deficiencies in the other.

Additionally, as a new application, the height field can be used with alternative color mapping schemes, similar to work done by Linsen et al. with LC-MS data [7]. In this case, the height of the peaks is a useful method for communicating peak intensity compared with the color mapped attribute. Alternative color mapping schemes are discussed in more detail in Section 4.1.2.

As an alternative to color mapping, we also provide a ‘high contrast’ rendering option for the height field. For this technique we enable OpenGL lighting and create a single light source positioned along the positive z-axis with ambient, diffuse, and specular components. We apply diffuse, specular, and shininess material properties to the polygons. Vertex normals are calculated at each vertex in the mesh, corresponding to each RT1/RT2 coordinate. The end result is a high-gloss, metallic looking rendering with high contrast.

With this technique, even small peaks are highlighted and readily noticeable, as seen in Figure 2 (right). Background noise is also highly visible in this view, as it produces a large number of small peaks and valleys.

4.1.2 TIC Color Maps

The total ion count visualizations both support color mapping based on intensity, difference, and standard deviations away from a mean. Each of these methods can be configured to use either a continuous or discrete color scheme. For the continuous color scheme, the system uses a set of three curves that allow independent control of the hue, saturation, and brightness.

For the discrete color scheme, the system presents the user with a histogram that displays bin colors and data distribution. The user can modify the number of bins, data range for the bins, and bin colors interactively. The color for each bin can be specified by the user, or the system can automatically generate a color mapping. In each case, values are initially mapped to a logarithmic scale where a larger color range is used to represent small intensities, and the scale of large peaks is greatly reduced.

We provide three different types of visualization modes for the TIC color maps: intensity, difference and standard deviation.

Intensity: The intensity mode is a simple mapping of the peak intensity to color. This technique is useful for providing a high-level view of the data. It reveals the location and relative intensity of peaks, and can be useful in helping a user identify any samples that may contain data collection errors. An example of the intensity mode mapping is shown in Figure 3 (left).

Difference: The difference mode calculates the difference between two samples. The result is then displayed by simply mapping the difference to a color. The system uses a separate set of HSV curves for positive and negative differences. By default, the hue for negative differences is set to pure green, and the hue for positive differences is set to pure red. An example of the difference mode is shown in Figure 3 (middle). Here, the user can quickly find areas of high positive or negative differences between two samples.

Standard Deviation: The standard deviation mode also calculates the difference between samples and renders an image based on that difference. However, a color mapping based solely on the magnitude of the difference in intensities may not be what is most interesting. Even using the mean of two sets of samples, the difference in intensity between two large peaks could be relatively high in magnitude compared to two smaller peaks, but this does not necessarily mean that difference is meaningful. By analyzing the standard deviation within user specified groups of samples, differences...
can be visualized in more certain terms. We create a standard devi-
ation color mapping from a sample group by first calculating a mean TIC for all the samples within that group. Note that the samples are chosen by the user such that they have been pre-normalized as input to the system.

Figure 4: The point selection tool is used for comparing TIC values, as well as interactively exploring mass spectra.

Next a standard deviation TIC is calculated as:

\[ \sigma_A = \sqrt{\frac{1}{n_A} \sum_{i=1}^{n_A} (x_i - \mu_A)^2} \]  

(1)

Here, \( n_A \) is the number of samples in group A, \( \mu_A \) is the mean TIC of group A, and the \( x_i \) are the TICs of the \( i \)th sample. Once we have computed the standard deviation TIC, it is stored to use for color mapping. This color mapping can then be applied to a sample visualization. Generally, it would be applied to a sample that is part of another group. In order to determine the color at a particular retention time coordinate, the system calculates the corresponding z-value for each point \( b \) in the new sample, as shown in Equation 2. The z-value is simply how many standard deviations different a value is than the calculated mean. We then use that difference to determine the appropriate color.

\[ z = \frac{\mu_A - b}{\sigma_A} \]  

(2)

This can help a user to determine whether an observed difference is truly meaningful. Additionally, this technique may effectively reveal areas of difference in smaller peaks that are significant in terms of standard deviations, but were not previously noticed simply because the peaks themselves are smaller. An example is shown in Figure 3 (right), note the green streaks that are not seen in Figure 3 (middle). This helps the user explore regions in the image that are statistically different in a sample when compared to a group of samples.

4.1.3 Mass Spectrum Visualization

A mass spectrum view is simply a plot of intensity on the y-axis vs. the mass-to-charge ratio on the x-axis, as seen in Figure 4 (bottom). The user is given the option to plot this spectrum as a bar graph, or as a connected line.

4.2 Linked Views and Data Exploration

Linked views are a common technique for comparative visualization. This type of interaction aids a user in quickly, and intuitively exploring data. Within the context of our system, we provide a user with multiple tools he may use in order to interact with the views, including scale, pan, rotate, point selection, area selection, mass filter, and tic lens. Such linked views are not scalable as the number of samples to be compared grows; however, by using the difference and standard deviation views, users can explore samples within groups for comparison, thereby reducing the need for numerous side-by-side comparison.

4.2.1 Point Selection

The point selection tool is used on 2D TIC visualizations. The tool generates point selection messages that can then be handled by other 2D TIC visualization, and by mass spectrum visualizations. As the user interactively moves the cursor over a 2D TIC visualization, a cursor is displayed at the same location on all 2D TIC visualizations. The total ion count value at that location is displayed in the lower left corner of the window, and mass spectrum visualization are updated with the corresponding mass spectrum at that location, as shown in Figure 4. All of this happens interactively, providing real-time TIC and mass spectra exploration and comparison for the data sets.

4.2.2 Area Selection

Rather than selecting a single point, a user may choose to select an area in a TIC image. This tool allows a user to draw a rectangular region on the TIC image. The aggregated TIC value is then displayed in the lower left, and mass spectrum visualization are updated with the total sum of intensity values for each mass in the entire region. In combination with linked views, this can be used to select a region containing a specific peak. The total integrated value of the region can be compared across samples, as well as individual mass values for any spectra that are displayed. The selected area can be made large enough so that the peak will be completely within the region across all samples, even if there is a slight variation in retention time across some of the samples. This is illustrated in Figures 8 and 9.
4.2.3 Mass Filter

When a user is looking for bio-markers and meaningful differences between samples, the user will often identify a few mass values that make up some interesting compound. The mass filter tool allows users to select a unique mass or set of masses from the mass spectrum display. Total ion count data for active TIC visualizations is then re-acquired using only the selected masses. The color mapping is then rescaled to match the new range of values across all of the TIC displays, and the TIC displays are refreshed. This allows a user to quickly determine differences among samples for a selected set of mass values. While other applications provide mass filtering via dialog options, none currently support real-time updating of the mass filter in an interactive manner. Mass values can quickly be added to or removed from the filter as the data sets are being interactively explored for differences in a unique set of mass values. Mass filtering is demonstrated in Figure 5.

4.2.4 TIC Lens

In any data visualization, color mapping may lead to a loss of visual features. This is due both to the limited color depth of the display, as well as to the limited perceptual ability of human vision. As a result of these properties, it can be difficult to compare peaks across separate TIC visualizations, and small peaks may not be noticeable at all. The TIC lens implements a dynamic colorscale adaption technique that was described by Elmqvist et. al [4]. In order to more closely examine a particular region of the dataset, the TIC lens tool allows the user to optimize the TIC color mapping for a particular region of interest. When the user selects a region using the TIC lens tool, the color mapping is rescaled to match the range of values that lie within the lens. This is accomplished via linear transformation from the old intensity range to the new intensity range. This technique is similar to Magic Lens filters [11]. However, while those techniques employed a local lens, our TIC lens tool is a global lens that rescales the color mapping over the entire image. This is demonstrated in Figure 6 (right). Our TIC lens tool has two modes of operation that are activated by the left or right mouse button. The first mode causes the color mapping for each visualization to be rescaled using the range of values within the lens of the image it was activated on. This allows absolute intensity values to be compared across samples. The second mode will cause each visualization to rescale based on its individual range of values.
within the lens region. This allows relative intensity values to be compared across samples. Using this tool with the 2D TIC visualization, the intensities of peaks can be more accurately compared, and small peaks that may have been previously hidden can be seen and compared.

5 Results

To evaluate the benefits of our system, we have worked with several biological and chemical researchers on several datasets. In order to better illustrate the use of this system, we describe below one session of our working side-by-side with a GCxGC-MS researcher to analyze one of his data sets. This particular data set consisted of canine serum samples. There were five samples each from healthy canines and canines with cancer. No preprocessing was performed on any of these samples.

He created ‘Cancer’ and ‘Healthy’ groups in order to classify the samples. After averages for the two groups were calculated, he then calculated the difference of the two averages. In order to get a quick overview of the data, the researcher first displayed 2D TIC visualizations of the means and difference side-by-side. In the difference image we observed several red peaks that indicated higher intensities in the cancer mean sample. However, there was a bright green peak about one-third of the way up on the difference image toward the left side, as seen in Figure 6 (top middle) labeled as Arrow 1. As this seemed interesting, he then used the region select tool in order to get a feeling for the relative peak sizes. He created two more visualizations in order to simultaneously show the mass spectra of these peaks for the mean samples. The first thing we noticed was that the mean of the healthy samples had a much higher background level than the mean of the cancer samples. Unfortunately, this makes comparison using the TIC values problematic. However, individual mass values can still be compared from the mass spectra views, and using this we see that the healthy mean sample has about twice the intensity of the cancer mean sample for that peak. He next wanted to verify that this difference was consistent. All ten samples were visualized simultaneously using 2D TIC views. This allowed him to verify that this peak was consistently bigger in the healthy samples. We also noticed a single sample among the cancer group that had significantly less background level than the other samples. This one outlier was responsible for most of the difference in background levels between the two means, and was therefore moved into a new group so that it would not be included in future comparisons between the calculated means.

In order to dig deeper into the differences between the samples, he next created standard deviation color mappings from the healthy and cancer samples. These color mappings indicated that several peaks showed higher than normal intensities in the cancer samples. After adjusting the color mapping to highlight only the largest of those differences, he chose to investigate one particularly noticeable difference near the center of the image, which is shown in Figure 6 (bottom, middle). After visualizing all of the samples simultaneously, use of the region select tool showed a consistent difference between the two groups. Again, differences in background levels reduce the accuracy of using the displayed TIC values as difference measures, but by visually inspecting the spectra we could see about a six-fold difference in intensities between the two groups, as seen in Figure 7.

He next began exploring this area by using the TIC lens tool to compare the two mean samples. He quickly noticed a peak in the cancer sample that barely showed up in the healthy samples, Figure 6 (middle) labeled as Arrow 2. Because this peak was small, it is barely noticeable using the standard color mapping. We zoomed into that area and immediately noticed two peaks in this area, Figure 6 (right) labeled as Arrow 3 and 4. The zoom and differences can be seen in Figure 6 (right). We again visualized all the samples simultaneously to look for consistency. The first peak (Arrow 3) was very prominent in 4 out of the 5 cancer samples. We saw no trace of it in 3 out of 5 healthy samples, and very slight presence in the other 2 healthy samples, as shown in Figure 8. The second peak (Arrow 4) was even more consistent, showing a significant presence in all 5 cancer samples, and little to no presence in all 5 healthy samples, as shown in Figure 9. By noting the retention time coordinates, he was then able to quickly look up and confirm these differences using the peak tables generated by commercial software. These consistent differences across samples mean that these peaks are potential biomarkers.

6 Discussion

Initial feedback from researchers working with GCxGC-MS data has been very enthusiastic. This system has provided them with their first opportunity to visualize multiple samples simultaneously. Enthusiasm has also been expressed about mean, standard deviation, and difference calculation for a set of samples. By visualizing these calculated data, the human eye can quickly identify differences. This system can be used to identify a particular peak or region of difference, and then the mass spectra can be explored to provide validation of differences, and hypotheses about the compounds involved. With this information, their existing tools can be used to obtain information about compound identification and intensity details much more quickly than was previously possible.

The researchers also frequently mentioned how pleased they were with the speed of the software. Other commercial systems will often take tens of seconds or minutes to even display a TIC image. Additionally, these software systems allow masses to be filtered, and individual spectra to be visualized, however, it is a slow and cumbersome process to change parameters and redisplay a new spectra or filter out different mass values. No other system currently used by this group was able to provide the fast, interactive filtering and mass spectra exploration of our system. As an example, the researchers can now quickly change the mass filter for a specific value, and slowly move through the entire range of mass values. Multiple samples can be visualized, and as the mass filter is updated the researchers can very quickly visually identify cases in which a unique mass has an unusual abundance in some samples. Currently, a version of this system is deployed for use on the Cancer Care Engineering Hub at Purdue University, http://ccehub.org.

7 Conclusions and Future Work

As our system was evaluated, we also received several suggestion for future work. Plans for future work involve producing output that can be used by other tools. For example, we could allow the user to select a single peak (or what appears to be a single peak) from a TIC image. This could be used to reconstruct a one-dimensional chromatogram for that region, and input that data into other existing tools that would then perform peak deconvolution (if necessary) and identification. Finally, many of these features could benefit from incorporating gradient based value mapping into the display. This was applied to GCxGC datasets in [12], and could be very effective when used with the types of comparative visualization techniques provided by this system.

Acknowledgements

The Cancer Care Engineering project is supported by the Department of Defense, Congressionally Directed Medical Research Program, Fort Detrick, MD (W81-XWH-08-1-0065) and the Regenstrief Cancer Foundation administered jointly through the Oncological Sciences Center at Purdue University and the Indiana University Simon Cancer Center. This work is supported by the U.S. Department of Homeland Security’s VACCINE Center under Award Number 2009-ST-061-C10001 and under the 973 program of China (2010CB732504), NSFC 60873123, and NSF of Zhejiang Province (N0. Y1080618). Thanks to Amber Janasch and Bruce Cooper for...
Figure 8: Examining the third difference. The TIC lens tool was used to rescale the color mapping in the TIC images, and the region select tool was used to highlight the peak and display the corresponding mass spectra.
their willingness to answer questions, provide input and feedback, and for data sets provided during development.

REFERENCES