Whale Sharks, Boolean Set Operations, and Direct Manipulation

Ramik Sadana*
Georgia Institute of Technology

Alistair Dove
Georgia Aquarium

John Stasko
Georgia Institute of Technology

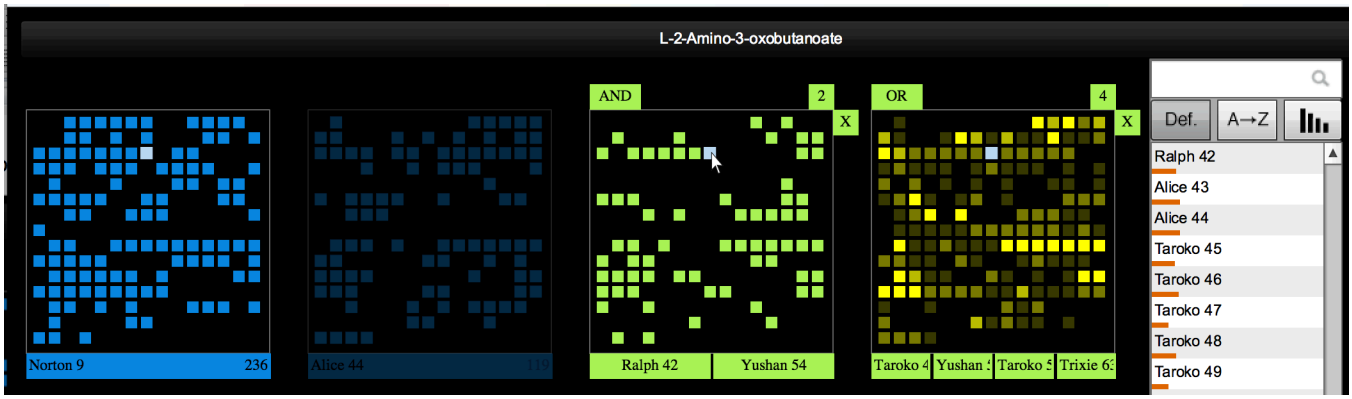

Figure 1. PixelLayer interface showing four blood samples and the compound L-2-Amino-3-oxobutanoate selected.

ABSTRACT

Whale sharks are the largest form of fish and are on the list of vulnerable species. We have developed a visualization technique to help marine biologists explore blood samples taken from whale sharks and the different bio-chemical compounds the samples contain. The visualization technique models each sample as a set that may or may not contain the individual compounds. Interactions on the samples show commonalities and differences as well as patterns in compound presence. Biologists can use the visualization to compare samples across days or weeks, find anomalies and trends resulting from diet, and thus gain better insights into the health of the fish.

Keywords: Design Studies, Molecular Visualization

1 INTRODUCTION

Whale sharks are the largest species of fish, growing to a length of up to 13 meters. They are on the list of vulnerable species and, as a result, marine biologists closely monitor their population and wellbeing. A recent study [1] used NMR and mass spectrometric methods ("metabolomics") to analyze whale shark blood as part of this health monitoring effort. It generated large tables of chemical data that were difficult to analyze and understand.

We have been able to procure these data for whale sharks that were or are currently resident at the Georgia Aquarium. Each blood sample consists of all the biochemical compounds that were detected as present in the sample. Every sample contains between 150 and 300 compounds, with a total of about 1100 distinct compounds throughout all of the data. Thus, the data consist of long lists of chemical names for each sample.

Biologists want to use these data to gauge the health of the sharks by analyzing and comparing the samples across days or weeks, or by comparing healthy and unhealthy individuals. Their intention is to find anomalies or trends in the chemical composition and relate that to the diet and health status of the sharks. Doing this based on chemical names rather than quantitative measures, such as concentration, presents a unique analysis challenge. In many cases, biologists approach this data without specific questions but seeking to explore and discover insights. Additionally, they want to detect trends in the blood samples and identify characteristics that may predict health issues.

One can model the samples within the dataset as a multi-variable binary-state data. That is, every sample is defined by the presence or absence of each compound throughout the entire data set. In this paper, we describe an interactive visualization technique we created to help biologists understand the data.

*email:{ramik@gatech.edu, adove@georgiaaquarium.org, stasko@gatech.edu}

2 SYSTEM DESIGN

Our data included samples for ten whale sharks, with an average of about 5 samples per shark, each taken on different days. The first step in our process was to combine all the compounds listed in the samples into a single list. This generated a list of 1100 distinct compounds, but only about 200 of those compounds were present in more than 10% of the blood samples.

If we consider an individual sample as a set with the individual compounds being the elements of the set that are present or not, operations such as set union (OR) and intersection (AND) are helpful to understand the data. One approach to visualize set data is the Euler diagram [2], but our data has too many set elements for this approach to be helpful.

In addition to detecting the presence of particular compounds in the samples, it is important to clearly observe the absence of specific compounds as well. In order to address these issues and assist marine biologists to explore the data, we developed a visualization technique that we call PixelLayer.

In this approach, we represent a sample as a square that consists of a matrix of smaller squares that we call pixels. Each pixel represents a compound from the master set of compounds. We use a 15x15 grid of pixels to represent the 225 most frequent compounds. Each unique compound takes the same (x,y) position within all samples and can have one of two states – present or absent. The present state is represented by blue color and absence is represented by black. Our visualization begins with an empty stage (exploration area for samples) and a list menu on the right containing the samples. Users can open the list and click on a "+" button to add a sample to the stage (Figure 1). Once the sample is on the stage, hovering the mouse cursor over any pixel highlights the pixel and displays the name of the compound at the top of the stage. Any number of samples can be added to the stage; each appears sequentially from the top left. Hovering the mouse cursor

Figure 2. Initial design using a fisheye approach with compounds in rows and samples in columns.

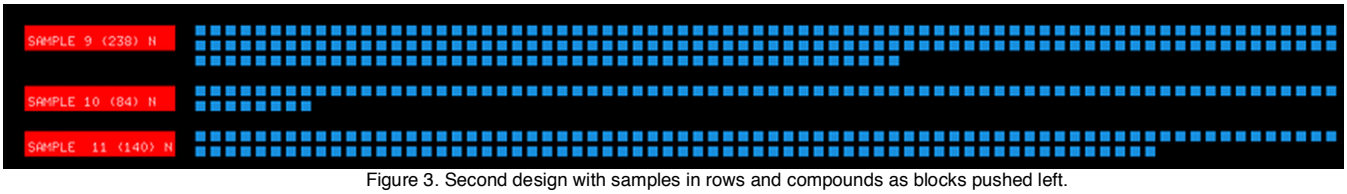

Figure 3. Second design with samples in rows and compounds as blocks pushed left.

on a pixel (compound) in any sample highlights the pixel in all other samples on the stage that also contain it. If a sample does not contain the compound, the entire sample fades out (second sample in Figure 1), thus helping the user to more easily observe samples containing the compound.

Samples are labeled at the bottom with their name as well as the number of compounds they contain. Samples can be interactively dragged around the stage using the mouse and act like "layers", as found in software such as Adobe Photoshop. When any portion of a sample is dragged on top of another sample, the two samples combine to show a "multisample" that depicts only those pixels common to the two samples in yellow (e.g., the third sample in Figure 1). In essence, this overlap operation acts like an AND operation between the compounds of the two samples and we indicate this at the top left of a multisample. Hovering over a pixel inside a multisample highlights the pixel in other samples in the same way as described earlier.

More samples can be added to a multisample in order to see common elements across a larger number of samples. The names of all included samples appear at the bottom, and the sample count of the number of samples it contains appears at the top right. Any of the labels can be dragged out to remove that sample from the multisample. The multisample can also be broken into all the constituents by clicking on the 'X' button at the right top.

In addition to showing the common elements across samples via the AND state, each multisample also provides an OR state in which the pixels take a color in a gradient scale of black to yellow (far right sample in Figure 1). The color for a pixel is based on the number of samples within the multisample that contain the compound it represents, with black being zero and bright yellow being the total number of samples in the multisample. The AND and OR states can be toggled by clicking on the button at the top left of the multisample.

The interactive, direct manipulation of the samples and compounds provides the user with a "hands on", flexible environment for investigating different combinations of samples. The user can explore similarities and differences across samples, as well as patterns and peculiarities across multiple samples simply by dragging samples around the stage.

In the list of samples to the right, each sample indicates the number of compounds it contains via a bar at the bottom of the label. The samples can be ordered alphabetically, by the date/time they were taken, or by number of compounds in the sample.

3 Discussion

We iterated through two different designs before finally settling on the current version. Our first design used a fisheye lens (Figure 2). Each row corresponded to a different compound and the columns contained the different samples. An element (x,y) was blue if the compound X was contained in sample Y and black otherwise. Hovering on an element expanded the row and colored it red, showing the name of the compound and the sample. To perform comparison of compounds across samples, one could drag a row up or down and place it next to the row of the compound to compare against. The compounds could be sorted based on frequency across samples. A search bar supported finding a particular compound.

One of the shortcomings of this design, however, was that due to screen space limitation, only limited samples could show up horizontally. As a result, a user needed to scroll, which was not efficient. Another shortcoming was that the interface did not give any information about either the total number of compounds or the number of samples in the data.

To address these concerns, in our second iteration, we switched the position of compounds and samples and collapsed the single row to multiple rows to prevent scroll (Figure 3). The rows, in this case, contained only the compounds that were present in a sample. Hovering over a compound highlighted all other samples that also contained it. This visualization design clearly showed the total number of compounds present in a sample. However, it did not support easy sample comparison due to compounds being in different positions in different samples and, thus, being hard to locate. Also, initial feedback revealed that auto loading of samples at startup was not preferred and instead, samples could be loaded on demand. This led us to our current PixelLayer interface.

In initial user testing with researchers from the Georgia Aquarium, we learned that they wanted the system to support importing multiple instances of one sample. This was needed to compare one sample to two different samples at the same time. As a result, we replaced a checkbox in the sample list with a plus button. Pressing the plus button would import a new instance of the sample, and one could thus import as many instances as required. For removing – instead of unchecking the checkbox, a user could drag and drop any sample from the canvas onto a recycle bin that appeared when the dragging began.

Informal user testing also identified the desire to more simply see the overall similarity of two samples. One suggestion was to draw a link between samples with its visual properties indicating the similarity. Finally, we received suggestions about allowing complex set functions on multiple samples. For instance, for four samples on the stage, users wanted to be able to see (S_1 AND S_2) OR (S_3 AND S_4). The current version of the system only permits using the same function across the expression – i.e. (S_1 AND S_2) AND (S_3 AND S_4) or (S_1 OR S_2) OR (S_3 OR S_4).

References

[1] A.D.M. Dove, J. Leisen, M. Zhou, J.J. Byrne, K. Lim-Hing, et al. Biomarkers of Whale Shark Health: A Metabolomic Approach. *PLoS ONE*, 7(11): e49379, 2012. doi:10.1371/journal.pone.0049379.

[2] N. H. Riche, T. Dwyer. Untangling Euler Diagrams. In *IEEE Trans. On Visualization and Computer Graphics*, 16(6), pages 1090-1099, Nov. 2010.