Interactive Visualization for Understanding and Analysing Medical Data

Samar Al-Hajj*, MSc
School of Interactive Arts and Technology
Simon Fraser University

Richard Arias**, PhD
SCIENCE Lab
Simon Fraser University

Brian Fisher***, PhD
SCIENCE Lab
Simon Fraser University

Abstract

Massive amounts of biomedical data generated by the latest high throughput technologies are challenging to analyze. Visual Analytics (VA) tools and techniques are intended to amplify medical researchers' cognitive and perceptual capabilities and enable them to understand complex biomedical data. In this study, we explore how visualization tools can facilitate the exploratory analysis of this data. In order to assess and evaluate the effectiveness and usefulness of using visualization tools to enhance medical analysts' data exploration, we analyzed the use of Tableau and iPCA by biomedical researchers to explore immunological data. Our findings reveal that VA tools are efficient and powerful tools that can be integrated into healthcare systems to help health researchers get insights and generate knowledge from their complex medical data.

Keywords: iPCA, Tableau Software, Interactive Visualization.

1 Introduction

The latest high-throughput biomedical technologies used in flow cytometry produce massive amounts of medical data. The magnitude and complexity of these data are overwhelming to immunological researchers including immunologists and biologists. Analysing and extracting useful information from these data impose a great challenge on the medical research community. It is our argument that efficient and effective visualization tools can facilitate the exploration and analysis of complex biomedical data. Interactive visualizations provide biomedical researchers and analysts with efficient tools and techniques to amplify their cognitive skills and enhance their initial understanding of the data during the exploratory analysis process.

Visual Analytics (VA) is defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [5]. These interactive visual interfaces rely on advanced visualizations of data and interactive techniques to accelerate the data analysis process, derive insights, acquire knowledge and optimize decision-making [6]. The implementation of interactive visualization tools was introduced in various medical disciplines to amplify analysts' cognitive capabilities and address the challenge of extracting useful information from massive datasets.

In this study, we present a case study of immunologists and biologists analyzing massive and multi-dimensional datasets using two visualization tools: iPCA (interactive Principal Component Analysis) and Tableau Software. Furthermore, we demonstrate how the integration of real-time visualization tools can help biomedical researchers uncover hidden trends in complex data and expose data patterns that are not noticeable otherwise, and ultimately facilitate the exploratory data analysis process. Finally, we show how immunologists exploited these visualization results to generate valuable qualitative information and drive new research questions.

* samara@sfu.ca
**ariasher@sfu.ca
***bfisher@sfu.ca

2 Task, Materials and Data

2.1 Task and Data

In order to assess the accuracy and effectiveness of VA tools for medical data analysis, we used analysis immunological data as a case study.

Sub-Saharan Africa has the largest HIV-infected population in the world [2]. The vast majority of infants born to HIV positive mothers are not infected themselves. However, those **H**IV **E**xposed but **U**ninfected (**HEU**) infants are at a high risk of mortality during their first year of life; they suffer severe immune system deficiencies and an abnormal susceptibility to infections and diseases [7]. The causes of this mortality and morbidity are unclear and are currently the subject of a biomedical research carried out by immunologists and biologists at the Child and Family Research Institute (CFRI) in Vancouver, BC. The main analytical goal of this research project is to understand the immune responses of HEU infants and link these responses to causes of high mortality and morbidity.

The HEU dataset included laboratory data generated by Flow Cytometry Luminex high-throughput technologies at CFRI. Blood and tissue samples from HEU infants, HIV positive infants, and unexposed infants (EU) were stimulated with several infectious agents and fed into the Flow Cytometry device to measure infants' immunological responses by focusing on cytokine levels. The datasets were multidimensional, heterogeneous and complex. They included the flow cytometry data on cytokine responses to infectious agents, as well as the infants' demographics, feeding methods, and vaccine reactions data.

2.2 The Analytic Setting: Paired Analytics

To focus on the accuracy and effectiveness of VA tools in supporting the analysis of these multidimensional biomedical data, rather than wasting immunologists' time in tool training, we decided to follow a pair analytics protocol for collaborative visual analysis [8].

In a pair analytics protocol, a visual analytic Tool Expert (TE) is paired with a Subject Matter Expert (SME) to conduct a collaborative visual analytic session organized around a well-defined task, a dataset, and a visual analytic tool [8]. Since the TE lacked biomedical expertise to conduct a meaningful analysis of the HEU data, and the SME lacked tool expertise to operate the visual analytic tool proficiently, their collaboration was required to make the most of the visual analytic sessions. In our case, the SMEs were biologists and immunologists. The TE was the main author of this paper. The pair analysis was structured to help immunologists exploit the VA tool and increase the speed, efficiency, and accuracy of the exploratory data analysis process [8]. Both experts worked together and exchanged expertise to understand the HEU data and assess the relevance of using two different visualization tools (iPCA and Tableau) for exploring biomedical HEU data.

3 VISUALIZATIONS AND INSIGHTS

Intuitive and interactive data visualizations facilitated the exploratory analysis of HEU data and enabled immunologists and biomedical researchers to analyse and interact with HEU data at various levels of abstractions to identify trends, patterns and formulate hypotheses.

To study HEU infants' immune system reactions to infectious agents, we explored HEU and EU infants' cytokine reactions using two interactive visualization tools: iPCA and Tableau Software.

3.1 Interactive Principal Component Analysis (iPCA)

Interactive Principal Component Analysis (iPCA) is an interactive visual analysis tool developed by the Charlotte Visualization Centre. iPCA uses the Principal Component Analysis (PCA) technique to reduce high dimensional datasets and convert data into new meaningful representations in order to facilitate users analytical reasoning and expedite the data exploratory analysis process [3]. Since HEU datasets were multidimensional, we plotted the HEU data in iPCA to visualize the reduction of variables representing infants' cytokine reactions into principal components, and to analyze the distribution and contribution of variables to the principal components.

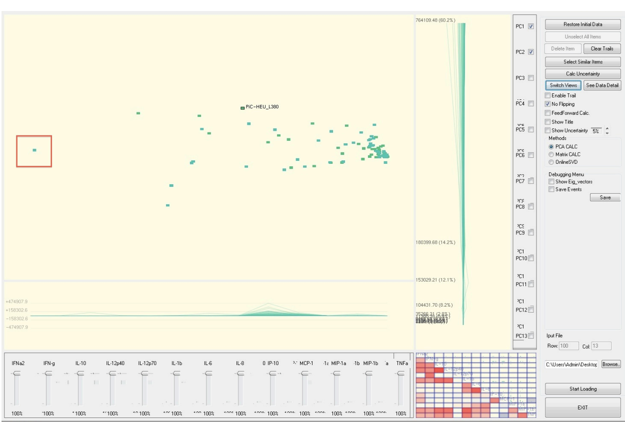

Figure 1. Visualization of Immunological data in iPCA Views: The Projection View, the Eigenvector View, the Data View and the Correlation View.

Fig. 1 shows cytokine reactions to one treatment (pIC). Each colour represents one group (i.e. HUE or EU), and each dot represents the values for one patient in that group. iPCA visually reveals the relationship between data variables, highlights outliers and provides immunologists with a comprehensive overview of existing correlations among pairs of variables. In Fig. 1, for example, outliers are quickly detected on the left side (highlighted here by the red box). It is also evident that the first principal component accounts for most of the variability in this dataset (60.2%). The slides associated with each variable (i.e. cytokines) allowed the TE to quickly show SME each variable' unique contribute to the principal components. In this particular visualization, most variables did not contribute significantly to the constitution of the first two principal components. One exception was the variable representing the cytokine IP10. Figure 2, shows the state of the visualization after the TE interacts with the "IP-10" slide dropping its contribution to the principal components to zero (highlighted by the red box). It was visually evident the dramatic reorganization of the values on the scatter plot representing the first two principal components. The first principal component, for example dropped from accounting for 60.2% to accounting for 42.3% of the variability in the dataset, while the second principal component increased from 14.2 % to 26.8%.

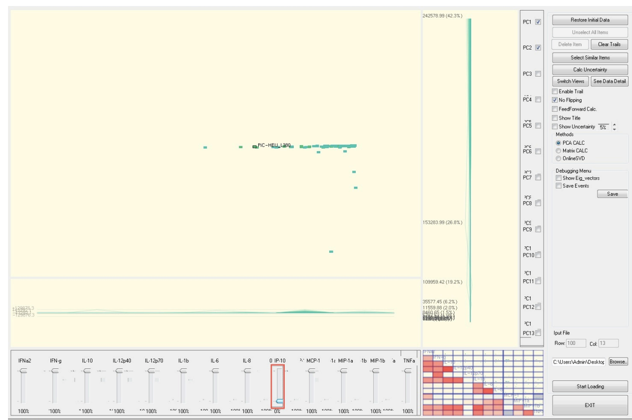

Figure 2. Interaction with individual variables to visualize their impact on the constitution of principal components

According to this analysis, the cytokine IP10 seemed to be a good candidate for further statistical analysis. To verify the analysis outcome, iPCA offered a matrix of correlations of pairs of variables. Every variable in the matrix is plotted against other existing variables to determine correlation coefficients (See Fig. 3). The correlation matrix proved very useful to quickly confirm the independence between IP10 and all the other cytokines in terms of responding to pIC treatment, which is visualized by the absence of dark red colours (i.e. indicators of high correlation) on the row corresponding to IP10 (highlighted by the red box).

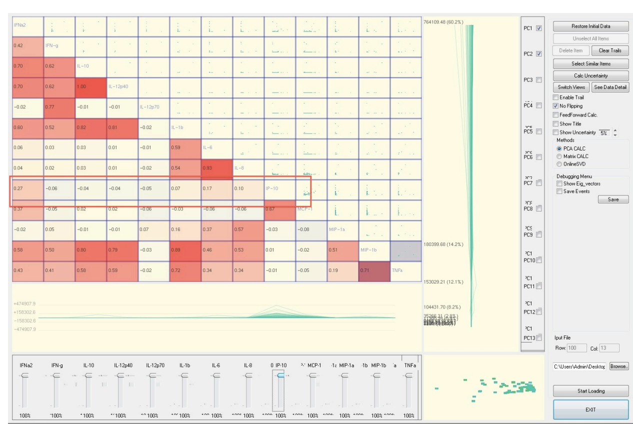

Figure 3. Correlation matrix

In summary, immunologists were able to quickly explore the HEU data and interactively identify which variables were more and less relevant for further statistical analyses. iPCA enabled immunologists to interact with the HEU data in real-time. Each time they changed a data item in one view; the change was automatically reflected in other views, giving immunologists the ability to understand data patterns and characteristics. By interacting with data, immunologists understood the influence and intuitively perceived the weight of separate variables on the constitution of the principal components. iPCA also allowed immunologists to visually detect and investigate outliers (Fig. 1) and their corresponding data items by eliminating an outlier from the data and observing its effect on the overall data visualization.

3.2 Tableau Software

Another visualization tool was used to visualize the HEU datasets: Tableau Software. Tableau is a commercial tool used for data exploration; it uses interactive visual dashboards to represent data and facilitate the exploratory data analysis process [1]. In order to compare the HEU, HIV and EU infants' cytokine reactions to infectious agents, we plotted and visually compared infants' IP10 cytokine reactions to each one of the 6 types of stimulations: CpG, pIC, R848, LPS, PG, and PAM, as well as the unstimulated control: Unstim.

The outcome of the graph, as shown in Fig. 4, depicts the infants' average cytokine reactions. Tableau represents cytokine reactions with different colors and saturations to reveal trends and show patterns in data. These patterns reflect variation across infants' groups, indicating that cytokine reactions are cohort-specific and vary between HIV unexposed and exposed infants. Tableau enables immunologists to drill down the HEU datasets and get further individual detailed information. Each bar of the graph represents all types of cytokine reactions per infant. The value of each cytokine reaction dictates the height of the bar. The shape of the bars represents a powerful visualization that provides immunologists with a comprehensive picture of the difference among HEU, HIV and EU cytokine reactions.

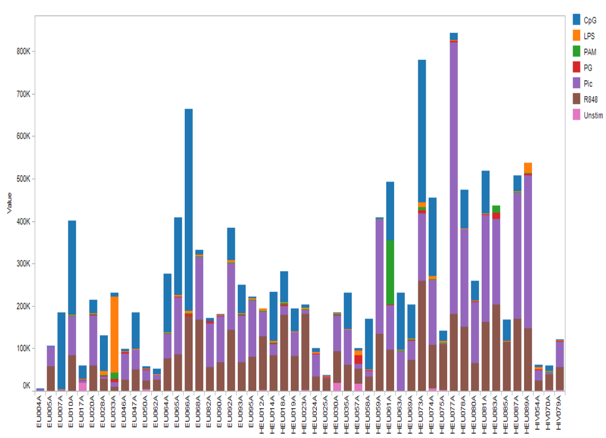

Figure 4. EU, HEU, and HIV infants' cytokine reactions to stimulations. [9]

The interactivity aspect of the HEU Tableau visualization supports immunologists' visual data exploration; it enables immunologists to hover the mouse over any particular patient and get on-demand detailed accurate statistical information about the infant's cytokine reactions to each of the stimulations. The observed variations in cytokine reactions proved the immunologists' hypothesis stating that HIV exposed but uninfected infants experience less immune defense against infectious agents compared to unexposed infants.

Immunologists were able to observe how HEU infants reacted differently to the majority of the stimulations, indicating that HEU infants' immune system differ from unexposed infants' immune system in terms of reacting to invading infectious agents and susceptibility to disease.

4 Discussion

Visual Analytics tools and techniques amplify biomedical analysts' cognitive and perceptual skills in order to observe and comprehend complex medical data, derive scientific insights and acquire knowledge to accelerate health discoveries [8]. Through real-time interactive visualization, Visual Analytics empowers biomedical analysts with the ability to reason and make sense of data under investigation.

Immunologists expressed their design and features preferences when interacting with the iPCA and Tableau visualization tools, which could be pivotal to inform the redesign of current tools to better fit the exploratory data analysis process. On one hand, immunologists pointed out flaws when interacting with Tableau. Immunologists were mainly concerned with the lack of the correlation option in Tableau, a feature that is perceived to be beneficial to the data exploration process. On the other hand, during the pair analysis session, the SMEs reported difficulties when interacting with iPCA. Something expected since iPCA is an experimental VA tool. Firstly, iPCA needs filtering; selecting and deleting groups as form of filtering is a cumbersome process. Secondly, SMEs couldn't directly access raw data from iPCA as the tool does not offer this functionality. The TE had to open a spreadsheet with the raw values on a second screen in order to have simultaneous access to the raw data. Thirdly, iPCA automatically assigned colours to groups and did not offer options for colour customizing to make differences between groups more visually salient. Fourthly, iPCA did not offer quantitative information about the exact contribution of each variable to each principal component. This information needed to be deduced by interacting with every single slide. A table with values may prove to be a better, faster, and more precise way to reach similar conclusions. Finally, iPCA did not provide features to export data. Since principal component analysis is an intermediary process in the statistical analysis process of multidimensional data, iPCA should enable users to export data to a statistical package to determine whether there is any statistical difference among the groups.

iPCA and Tableau encounter few pitfalls that constraint and limit their applications to our current HEU data. However, our preliminary findings indicate that VA tools support biomedical data exploration and knowledge dissemination. iPCA and Tableau visualization examples validate the relevance of using efficient VA tools and techniques for healthcare applications. iPCA and Tableau visualizations reveal important features about the HEU dataset and illustrate the useful application of Visual Analytics for data exploratory analysis. Furthermore, VA promoted collaboration and dissemination of information among health professionals, which is vital for the decision-making process [4]. Tableau visualization software enabled biomedical researchers to disseminate, share and communicate analysis results with a variety of audience through the creation of dashboards. Produced dashboards can be published to communicate information, interactively explore results and disseminate knowledge to colleagues as well as patients to ease dialogue with them. This is an efficient way to share knowledge and promote collaborative analytical reasoning.

Based on the visualization of the HEU data, VA provided powerful interactive visualizations needed to assist immunologists and medical researchers in data exploration as well as to generate hypotheses and test these hypotheses. VA enabled immunologists to engage and interact with the high dimensional HEU data, discover details and relationships among data variables, recognize relevant patterns, identify data clusters and outliers, and ultimately advance their research. Immunologists' experience motivates other health professionals and promotes the use of VA tools and techniques to explore complex data and to integrate powerful and effective visualization software in clinical practices.

5 Conclusion

High throughput flow cytometry technology provides immunologists with complex and multifaceted data. Exploring and examining massive and unstructured medical data exceed the ability of health professionals to synthesize meaningful information. Interactive and dynamic graphical presentations of data empower immunologists with a better perception of the HIV disease progression and a good understanding of the HEU infants' immunodeficiency. Visual Analytics uses interactive and intuitive visualizations to help medical researchers determine hypotheses, formulate research questions and conduct exploratory data analysis efficiently. Effective visualization of the HEU data represents a fundamental step in the data analysis process that can guide relevant medical discoveries and gain insights into valuable medical information. Understanding complex HEU data and drawing valid conclusions enable immunologists to identify the health determinants of HEU infants and eventually make decisive public health interventions to reduce HEU infants' sufferings and bring changes to the lives of over 300,000 HEU infants born annually [7].

We identified emerging challenges with iPCA and Tableau that could provide opportunities to improve the current version of the tools or design new tools that accommodate the needs of biomedical researchers and analysts. Further research into the potential implementation of visualization software for medical applications will determine how these visualizations can significantly affect the way analyst look at their data and guide effective integration of VA techniques and tools in various health care systems to help medical researchers generate knowledge and gain insights.

6 Acknowledgement

The authors would like to thank immunologists and biomedical researchers at the Child and Family Research Centre (CFRI) as well as our colleagues at SFU SCIENCE Lab for their collaboration and support.

References

[1] Tableau Software, 2011. http://www.tableausoftware.com/

[2] United Nations Programs on HIV/AIDS (UNAIDS). Report on the global HIV/AIDS Epidemic. Global report. 2010

[3] D. H Jeong et al., "iPCA: An Interactive System for PCA-based Visual Analytics," in *Computer Graphics Forum*, vol. 28, 2009, 767–77 http://www.vrissue.com/portfolio/iPCA/iPCA.php

[4] D. Keim et al., "Visual analytics: Scope and challenges," *Visual Data Mining* (2008): 76–90.

[5] J. J Thomas and K. A Cook, "Illuminating the path: The research and development agenda for visual analytics," *IEEE Computer Society* (2005).

[6] P. N Johnson-Laird, *How we reason* (Oxford University Press, USA, 2006).

[7] Peter Wall Institute for Advanced Studies. "HIV-Exposed but Uninfected (HEU) Infants: Exploration of the Causes of Enhanced Morbidity and Mortality". The University of British Columbia. 2009.

[8] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher, "Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics," in System Sciences (HICSS), 2011 44th Hawaii International Conference on, 2011, p. 1-10.

[9] S. Al-Hajj, E. Fortuno III, and B. Fisher, "Data Visualization of Immunological Competence of HIV Exposed but Uninfected (HEU) infants", *IEEE Visual Analytics for Science and Technology (VAST)*, 2011.