CS533C Project Proposal

Visualization Tool for Flow Cytometry Project

Evgeny Maksakov
Department of Computer Science
University of British Columbia
maksakov@cs.ubc.ca

The project will be carried out in collaboration with Terry Fox Laboratory (TFL), BC Cancer Agency. My research will focus on the design and implementation of a visual tool for flow cytometry data representation of real datasets. It will be based on data obtained from the Flow Cytometry Data Standards Project [0] and an informal user study of TFL members who use flow cytometry in their everyday work.

Flow Cytometry Data Standards Project:
Principal Investigator: Dr. Ryan Brinkman, Assistant Professor, Medical Genetics, UBC (rbrinkman@bccrc.ca)
Postdoctoral fellow: Dr. Josef Spidlen (jspidlen@bccrc.ca)

Domain, task, and dataset

The project domain is the multidimensional data visualization, data filtering, data clustering and HCI.

Flow cytometry (FCM) is a technology that simultaneously measures and then analyzes multiple physical characteristics of single particles, usually cells, as they flow in a fluid stream through a beam of light. The properties measured include a particle's relative size, relative granularity or internal complexity, and relative fluorescence intensity. These characteristics are determined using an optical-to-electronic coupling system that records how the cell or particle scatters incident laser light and emits fluorescence. [1]

The data, collected using the FCM, can be of an average type dimensionality, which means it could range from 4 and up to 20 dimensions. Dimensions are represented by forward and side light scattering results and fluorometric results. Most often researchers use 5-10 dimensions in their analysis. Number of events (particles going through the laser beam) inside a dataset can be as high as a million, though it is typically in the hundreds of thousands for stem cell research, such as carried out at the TFL.

Although there are several commercial applications for the analysis of flow cytometry data, including FlowJo [2] and FACSDiva [3], there is a desire to have their own, more effective visualization tools. The existing systems provide visualization for the data in the form of scatterplots, contour diagrams and histograms (Fig. 1). To visualize data this way, we need one scatterplot per each pair of dimensions we are interested in. The problem with this representation of the multidimensional data is that it is hard to see the whole dimensionality at the same time. Usually, researchers use "gates" (an area enclosed in the contour) inside the scatterplots to analyze what is happening with this particular set of events on other scatterplots (Fig. 2).

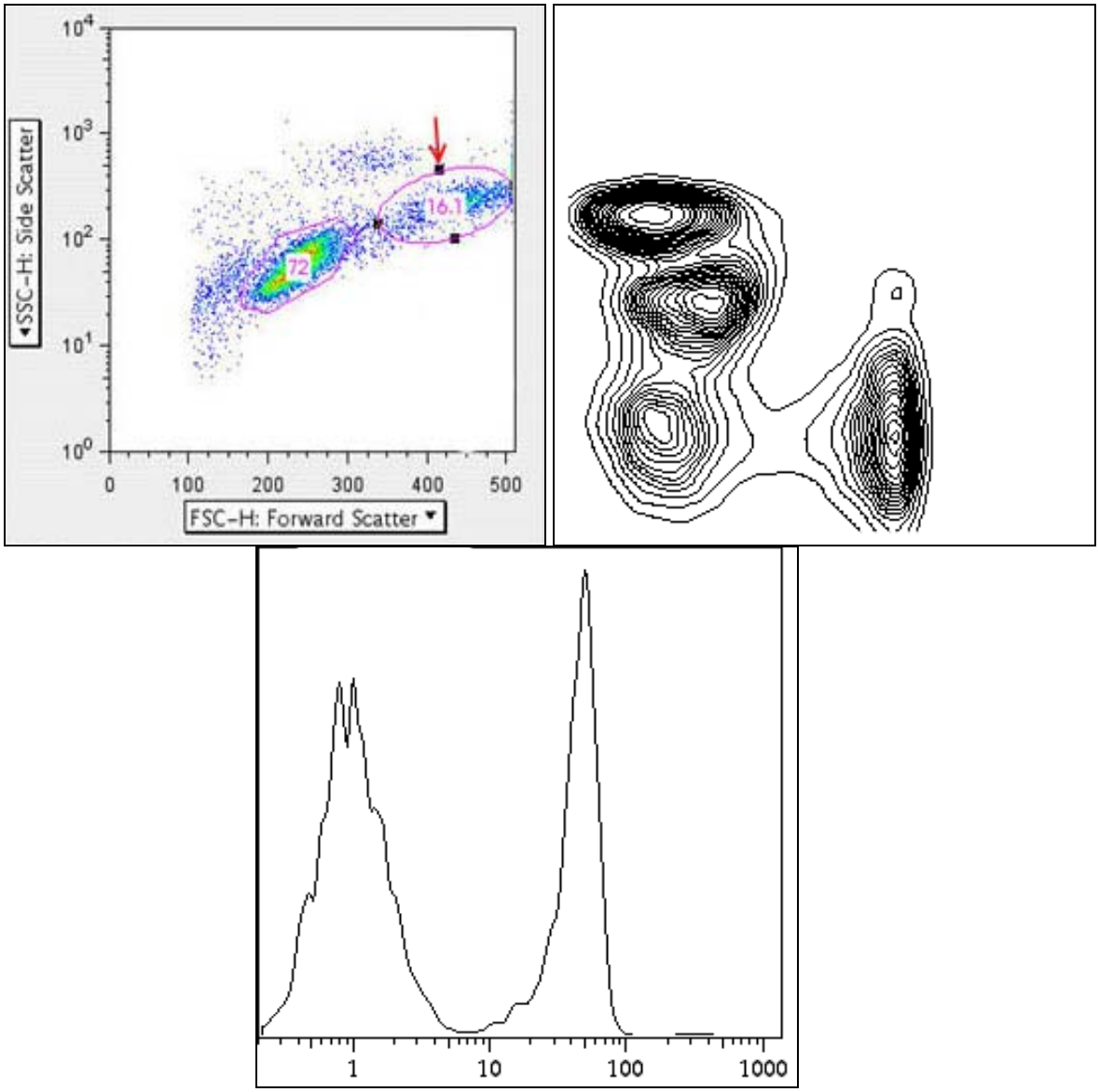

Fig. 1 FlowJo scatterplot, contour diagram, histogram (figure is taken from [2])

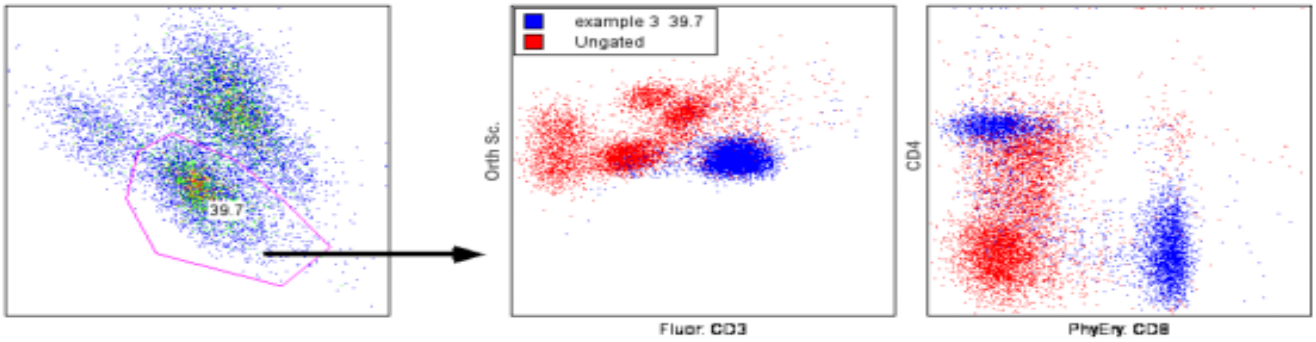

Fig. 2 Example of gating in FlowJo (figure is taken from [2])

My task is to create an initial visualization tool for the Flow Cytometry Data Standards Project and at the same time try to improve the visual representation of the datasets for easier analysis in comparison to the available commercial software.

Personal expertise

Before I decided to work on this task I have never heard of the flow cytometry and never visualized multidimensional data. Nevertheless, this project will provide me with an interesting opportunity of applying my theoretical knowledge of the task, which has an important practical impact.

Proposed solution

Scatterplots provides useful information about the data in two dimensions but scientists also want to see the whole picture. The solution to this problem is to find a way to show all dimensions at once and, at the same time, be able to understand the trends inside the dataset. The solution is the usage of parallel coordinates. Unlike the scatterplots, it connects values throughout dimensions.

Usage of the parallel coordinates for flow cytometry was recently attempted by Marc Streit *at al*. in 2006 [4]. But in this work, the emphasis is made on the representation of the event density in 3D, which does not give any additional vision of the data dimensions in comparison to the 2D version. It also does not show if clusters have common sections and, sometimes, it will be impossible to determine which cluster is where.

We plan to undertake a different approach (Fig. 3). Instead of showing just the event density, it is possible to cluster data throughout dimensions and show individual clusters even if they have common sections and intersections. Such approach for data visualization is described in the work by Ying-Huey Fua *et al*. [5], which will be taken as a component of this solution. Scatterplots will not be rejected either. They are very useful for certain tasks such as enclosing clusters in the polygon-shaped or elliptic gates. Histograms are also a useful tool for representation density along one of the dimensions. Contours are used for monochrome or color representation of the density; however, they do not provide additional information about the data and will be left for the further development.

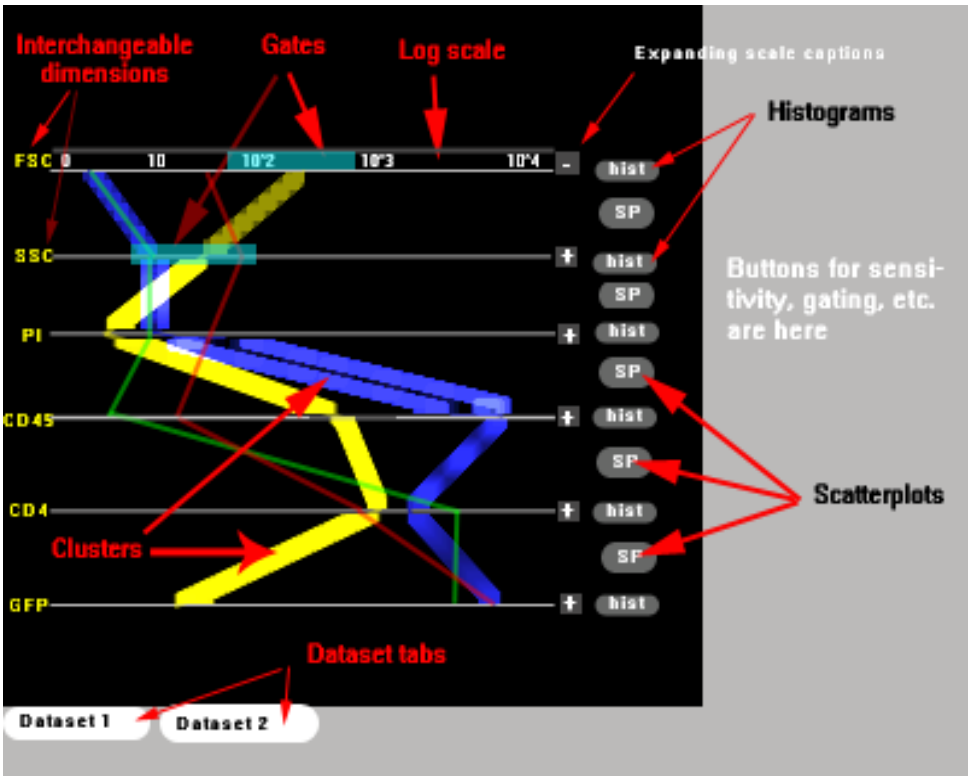

Fig. 3 Sketch of proposed visualization of flow cytometry dataset.

The parallel coordinate's interactive representation will be able to provide possibility to create gates and filter the data while reducing amount of manipulations necessary to make to organize scatterplots.

Hypothesis: The proposed system will allow the user in many cases to see the cluster that fits the user needs right away without the set of gates.

Scenario of use

Scenario, described further, is very simple but based on the information acquired by interviewing an actual user of the FlowJo system. Then actions of a user were mapped to the proposed solution.

John Flow just finished his experiment and wants to check his results using FCM analysis. He goes to the computer and starts the visual tool. Through the menu he loads from the file his negative control (cells that do not have fluorescence) dataset. Its representation successfully appears on his screen. He looks at the data between first two axes. These axes represent side-scattering/SSC (granularity) and forward-scattering/FSC (size). After a short glance, he creates gate #1 (selected region for further analysis) by selecting one interval on each axis. The gate separates now the distinct cellular cluster from the cellular debris. Then John looks on the data between second and third axes: FSC (size) and PI (dye to distinguish live cells) and by the single mouse move continues the gate to the PI dimension. Now, dead cells are separated from living ones by the gate. Next he gates negative control cell cluster on the fluorescence intensity axis and it will be the basis for comparison to the positive control result.

Then John one by one opens his several positive control datasets in the visualization tool. They appear in separate tabs. Since he already set the gate for the negative control dataset he does not need to repeat this operation anymore. Gates are already preset for each positive dataset. As soon as he finishes with loading John looks at the percentage of the fluorescently marked cells in the gate, relative to the whole set. An exact percentage is provided by the system. This is the result he wants but for his upcoming paper he also opens scatterplots and histogram views to save a few pictures of his resulting clusters.

An advantage of this process is that its beginning is exactly the same for all the analysis cases with, probably, few exceptions. It makes the choice of initial axes easy. For more complex analysis, steps following the described will vary depending of the task…

Implementation tools

Java 2D, a library to read flow cytometry data format.

Milestones

1. **October 27:** Gathering preliminary information of the flow cytometry and its datasets, informal interviewing of the technology users. (Done)
2. **November 1:** Submit project proposal.
3. **November 6:** Implement non-interactive parallel coordinates, scatterplots and histograms without clustering, initial GUI.
4. **November 20:** Implement clustering and visualize it.

5. **November 27:** Added interactivity for interface (interchangeable axes, gates, highlighting, etc.), some less important features might be omitted if there will be no time to implement them.
6. **December 7:** Implementation finished, user study.
7. **December 14:** Final report/presentation.

References

[0] http://www.flowcyt.org
[1] Introduction to Flow Cytometry: A Learning Guide by BD Biosciences, http://www.cancer.umn.edu/exfiles/research/fcintro.pdf, 2000.
[2] FlowJo, http://www.flowjo.com
[3] FACSDiva, http://www.bdbiosciences.com/features/products/display_product.php?keyID=93
[4] Marc Streit , Rupert C. Ecker , Katja Österreicher , Georg E. Steiner, Horst Bischof , Christine Bangert , Tamara Kopp, Radu Rogojanu, 3D parallel coordinate systems - A new data visualization method in the context of microscopy-based multicolor tissue cytometry, Molecular Cell Biology, Cytometry Part A, Volume 69A, Issue 7, Pages 601-611 (2006)
[5] Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner, Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets, IEEE Visualization (1999)