FlowCytoVis: Visualization Tool for Flow Cytometry Data Standards Project

CS533C Course Project Final Report

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

Note: This project was carried out in collaboration with the Terry Fox Laboratory (TFL), BC Cancer Agency. My research was focused on the design and implementation of a visual tool for flow cytometry data representation of real datasets. It was 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)

Abstract

The research in the Terry Fox Laboratory (TFL), BC Cancer Agency, Vancouver, BC involves the use of flow cytometry (FCM) technology. Current methods of visualization of these specific data include scatterplots, histograms and contour diagrams, which have their disadvantages in multidimensional data analysis. The work presented in this paper introduces a new visualization tool for flow cytometry multidimensional datasets, FlowCytoVis. This tool uses parallel coordinates and data clustering for the visual representation of the FCM data besides traditional scatterplots and histograms. FlowCytoVis prototype has shown that the parallel coordinates can be an effective visualization technique for the FCM and can be valuable addition to the currently used scatterplot techniques.

1. Introduction

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 (size) and side (granularity) light scattering values, PI (propidium iodide) fluorescent intensity (viability), GFP (green fluorescent protein) fluorescent intensity (gene expression) and 16 different fluorescent intensities that represent spectrum of light wavelengths that represent colors from infrared to blues of fluorochromes (cell markers). Most often researchers use 5-10 dimensions in their analysis. Number of events (particles going through the laser beam) in a dataset can be as high as a million, though it is typically in the tens to 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 individual, 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).

There are also generic data visualization tools, such as GGobi [8], which provide variety of visualization techniques, including parallel coordinates, but the problem here is to choose the right technique and to use it right way, which is not always trivial because people who use flow cytometry for work usually are not familiar with the InfoVis techniques advantages and disadvantages or even techniques themselves.

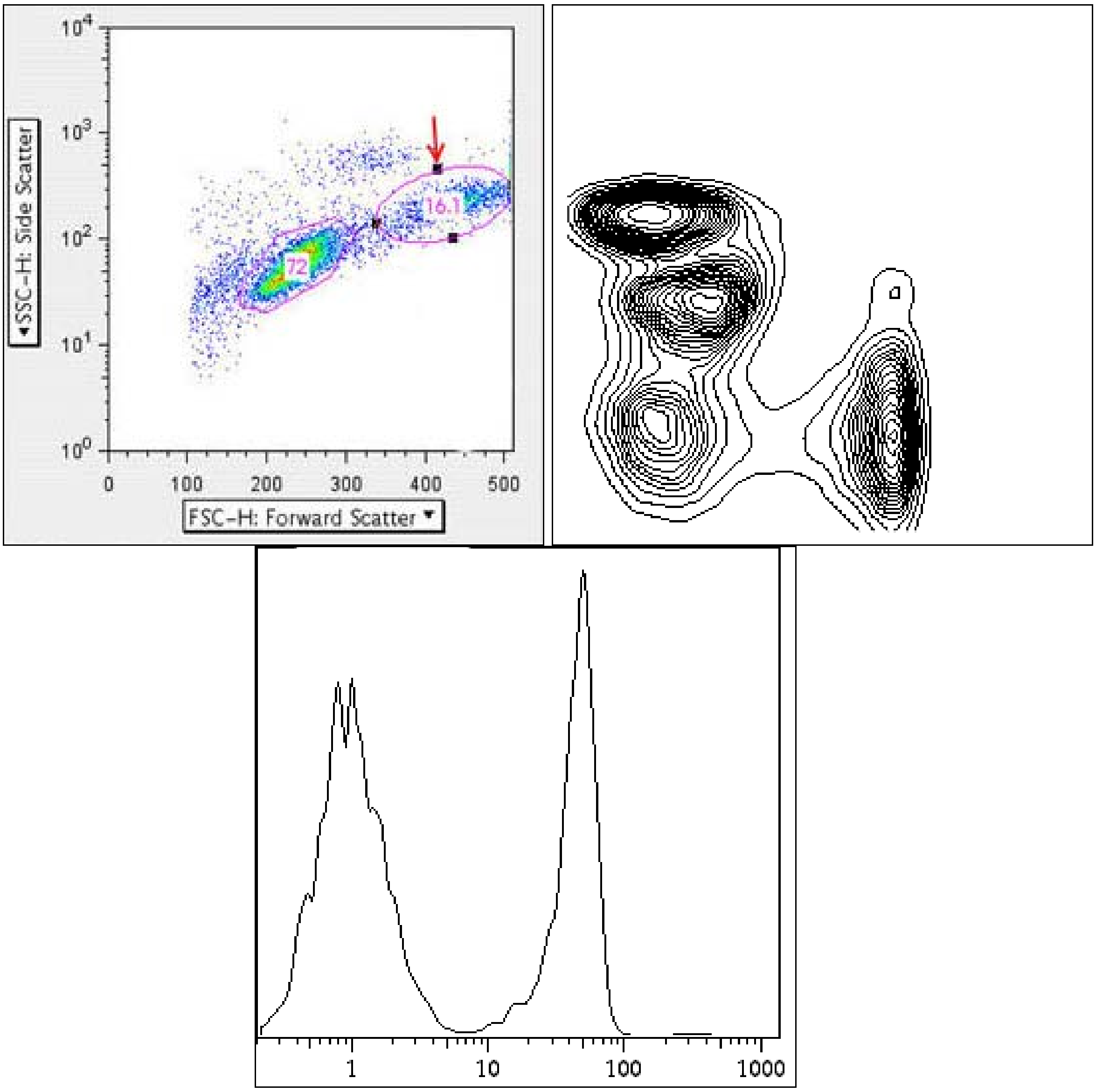


Figure 1. FlowJo scatterplot, contour diagram and histogram (figure is taken from [2])

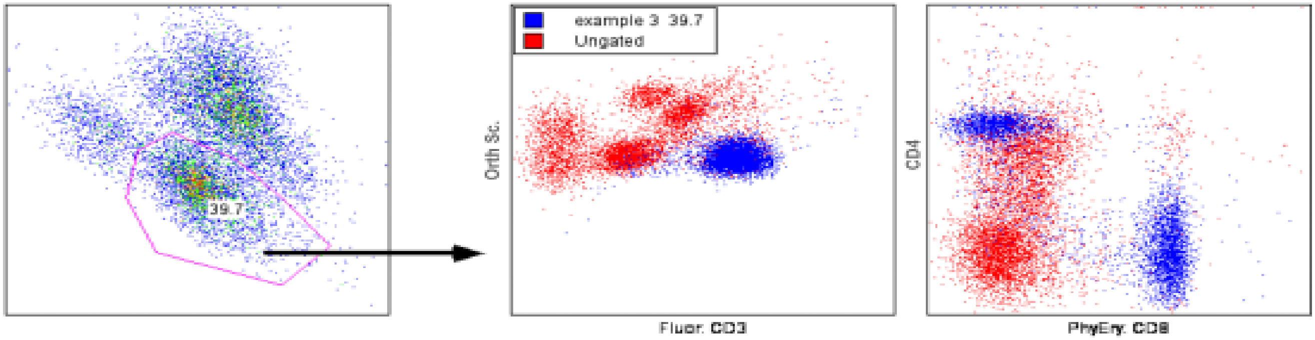


Figure 2. Example of gating in FlowJo (figure is taken from [2])

The direction of this project was the visual representation improvement of the datasets for easier analysis in comparison to the available commercial software besides creation of a visualization tool for the Flow Cytometry Data Standards Project.

2. Related work

Parallel coordinates technique exists since 1980s, flow cytometry was introduced in 1960s. There are numerous visualization systems that support parallel coordinates but before 2006 there were no indications that anybody tried to specifically apply parallel coordinates technique for the flow cytometry data visualization.

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, introducing the height axis to represent what previously has been shown by the coloration. Such approach does not provide any additional information of the data dimensions in comparison to the 2D view, except for increasing its computational cost and introducing additional interaction (rotation) to be able to see all the data as well as in 2D. It also does not solve the problem of separating clusters if they have common sections and, occasionally, it is impossible to determine which cluster is where. Thus, such 3D representation cannot serve as a good example of the appropriate use of the parallel coordinates for the aimed visualization.

Introducing the 3rd spatial dimension does not guarantee an advantage over the 2D visualization. There are methods that allow exploiting 2D representation more efficiently. Fua *at al.* [5] suggested using hierarchical parallel coordinates involving hierarchical clustering for large datasets and proximity-based coloring and variable-width opacity bands for providing better understanding of the data tendencies. Their solution was based on the publicly available XmdvTool [7] that provides several visualization features including scatterplots and hierarchical parallel coordinates.

3. FlowCytoVis

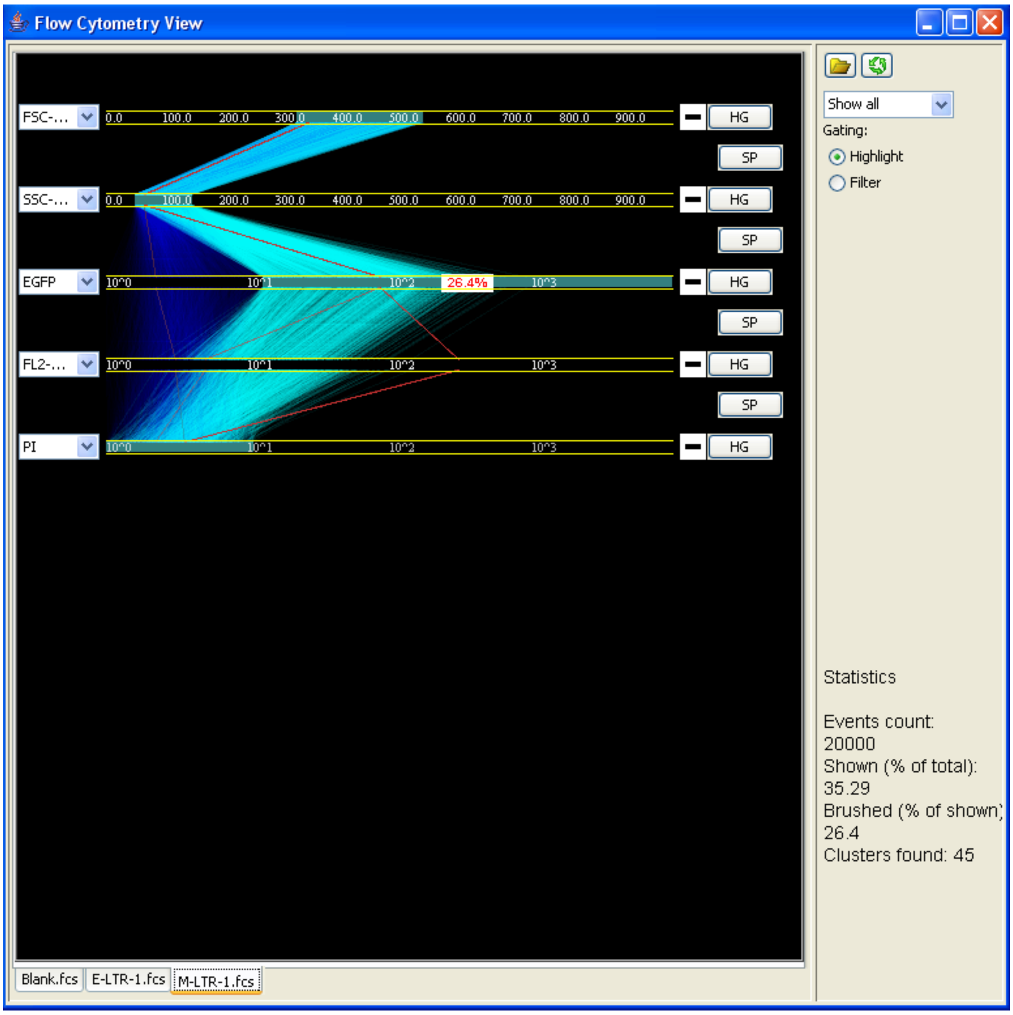
3.1. Implementation details

Java2D and Java Swing were taken as tools for the implementation of the FlowCytoVis. Java provides possibility to use it across platforms and also is the programming language the author of this report is most skillful with. FlowCytoVis uses the CFCS library of the Flow Cytometry Data Standards Project, also written in Java, to read FCM datasets from files in FCS format. Everything else was implemented by the author of this report.

3.2. FlowCytoVis design

Scatterplots provide useful information about the data in two dimensions but scientists also want to see the whole picture. This creates a demand for the visualization design to show all dimensions at once and, at the same time, be able to understand the trends inside the dataset.

Another inconvenience of the scatterplots in the available commercial software for FCM is that they employ many windows, which then require some substantial manipulations to open. And at the end there are multiple 2-dimensional slices of the data, which are hard to process by human brain without using additional filtering. The solution proposed here is the usage of parallel coordinates together with clustering. Unlike the scatterplots, it connects values throughout all dimensions.



Show all

Gating:

● Highlight

○ Filter

Statistics

Events count:
20000
Shown (% of total):
35.29
Brushed (% of shown)
26.4
Clusters found: 45

Figure 3. Data analysis using FlowCytoVis

This presented work has taken a different approach (Fig. 3) in comparison to the 3D parallel coordinates work, described in [5]. It improves every aspect of parallel coordinate visualizations in 2D which authors of [5] rejected as not-good-enough solution or did not mention at all. The density of the lines in FlowCytoVis is shown by using alpha blending instead of coloration. The new system provides interactive highlighting and filtering. Interchangeable axes allow organizing the data any way the user might want to. Clustering data throughout dimensions is aimed to improve understanding of data trends. A user can highlight individual clusters even if they have common sections and/or intersections.

Scatterplots are not rejected either. They are very useful for certain tasks such as enclosing clusters in the polygon-shaped or elliptic gates, which would be impossible to do manually having only parallel coordinates feature. However, since the dealing with scatterplots is very well known, this work did not make an emphasis on the copying of the full functionality of the FlowJo system.

The same approach was applied to the histograms. Histograms are also a useful tool for representation of event density along each dimension. In addition, the histograms were used as a base for the clustering method. Contour diagrams that exist in the FlowJo software are used for monochrome or color representation of the density; however, they do not provide additional information about the data and were left for the future development.

The parallel coordinate’s interactive representation is able to provide possibility to create gates and filter the data while reducing the amount of manipulations necessary to organize them as in the case of scatterplots.

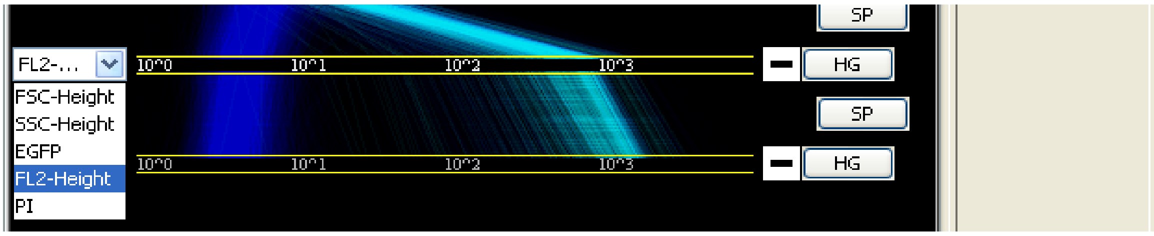


Figure 4. Axes selection in FlowCytoVis

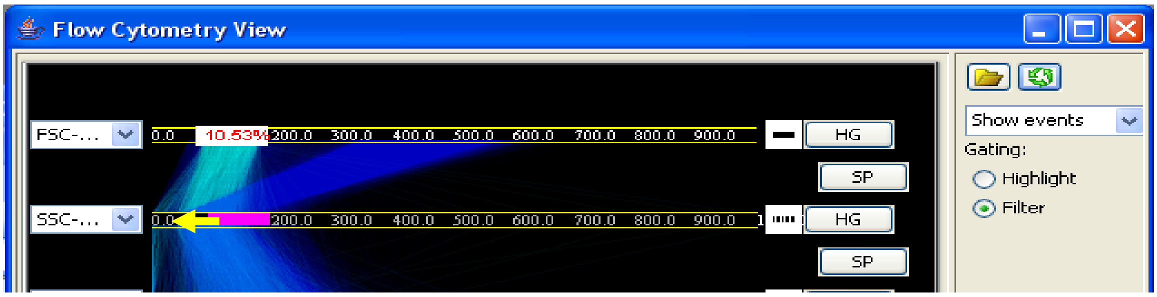


Figure 5. Applying a filter with a mouse dragging with pressed left button.
The arrow shows the direction of selection.

An advantage of the parallel coordinates use in FCM is the analysis specific first steps: sorting out cell fragments and dead cells. They are exactly the same for all the cases with, probably, few exceptions. It makes the choice of initial axes ordering easy. For more complex analyses, steps following the initial ones vary depending of the task.

3.3. FlowCytoVis interface functionality

The interface of FlowCytoVis allows an easy access to all the features available for the user. They include opening flow cytometry files, axes ordering, switching between visualization modes: data only, clusters only, and all at once; two ways of filtering: highlighting and exclusion; forced re-rendering, hiding/showing axes captions, histograms and scatterplots views, and switching between datasets.

To open fcs file the user needs to press the “open” button at the top-right of the screen. Very often fcs files have different extensions, which makes them hard to extract manually from the common mass of files. That is why the FlowCytoVis checks the headers of each file in the directory to determine if they belong to the FCM standard and provides the list of the appropriate files.

Axes interaction is provided by the set of the combo boxes on the left side of the window (Fig. 4). After the selection the user should press the green “render” button (At top right on Fig 5.) to construct a new picture. It is possible to select axes of the same name on the screen. Plus and minus buttons to the right of the axes lines allow to show and hide the scale captions for each axis individually. However, this operation requires re-rendering of the picture that can take a substantial amount of time, depends on dataset. Mainly, it was implemented for the case of “clusters only” visualization mode when there are many axes present at the same time. But it is not prohibited to use it in the other modes too.

The user can switch between three modes of visualization using the combo box on the right panel and clicking the rendering button (Fig 5.). These self-explanatory modes are “show all”, “show events” and “show clusters”.

Two filter types are represented on the interface by two radio buttons on the right panel. Traditionally, filters for flow cytometry analysis are called “gates” and in this project the author decided to use the same terminology. Highlighting changes the filtered data color and shows the percentage of the highlighted event of the whole visible amount (Fig. 5). Filtering selects the interval on which the events and clusters that pass through become the only visible data.

To set up a filter the user must press the left mouse button over the axis and drag it along the axis (Fig. 5). When the button is released the filter applies. To remove the filter the user can use right click on the filter he/she does not need.

There is a possibility to look at the histograms that represent event densities of dimensions their buttons positioned at. Scatterplots buttons pop up windows with the scatterplot based on the axes in between which the scatterplot button is located. Because the emphasis of the project was towards the parallel coordinates solution histograms and scatterplots were left without axes captions.

It is possible in FlowCytoVis to load several datasets and switch between them using the tabs at the bottom of the screen. It is important to know that all the filters applied to the current dataset are automatically applied to the just loaded dataset too. It was made due to the general FCM analysis practices.

In the right bottom corner the user can see a short statistical data about the current view. It provides information about the total number of events, the percentage of the events that appear on the screen at the moment and the percentage of highlighted events relative to all visible events.

3.4 Rendering

The FlowCytoVis uses dynamic alpha blending to show the density of the data and static alpha component to make filter bars transparent. The alpha component for events visualization is based on the number of events in the dataset. More events mean less salient the individual lines. This is not the case with cluster centroids because this way some clusters become almost invisible and impossible to see, which would be bad for the visualization purpose.

To prevent multiple re-renderings during the Java update events the system generates an image of the visualization and re-render it only if cannot be avoided. All the visualization is scalable and readjusts according to the window size.

4. Clustering

4.1 Background

The choice of proper clustering is an important decision. Bad clustering can make the data analysis more challenging. Thus, after looking through several available clustering options designed specifically for flow cytometry, the feature-guided algorithm by Zeng *at al* [8] was deemed to be the most promising since it reported 100% determination of predetermined cluster combinations and had acceptable computational cost. However, several changes to this algorithm have been made. The result of this was an easier implementation without theoretical loss of quality of the results, though accompanied by a slightly higher computational cost. Also, the modified version of the algorithm does not discard the outliers, which may be important for some research goals, when the events of interest are outliers themselves.

The idea of the algorithm is to analyze histograms of each dimension and use peak maximums as centroids of clusters. It is also assumed that data has a Gaussian distribution and, before running this analysis, it is necessary to smooth all histograms (Fig. 6) by averaging surrounding values. Such adjustment removes the noise from the curve and makes clustering meaningful. Zeng *at al* [8] suggest making the number of clusters equal to the maximum number of peaks per one dimension among all dimensions.

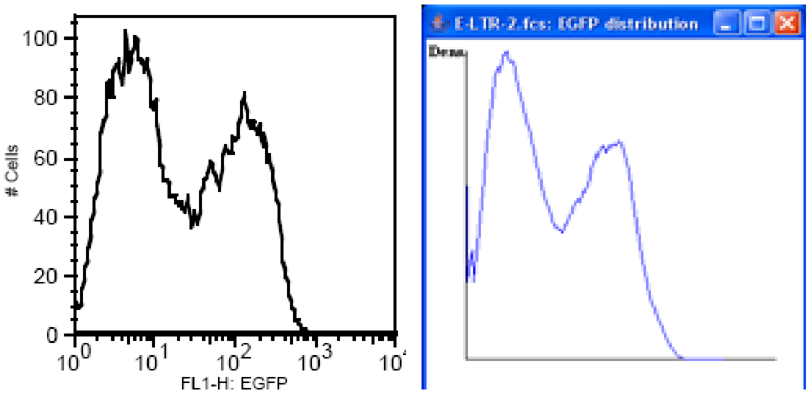


Figure 6. On the left: original histogram (FlowJo). On the right: the same smoothed histogram (FlowCytoVis).

The presented modified algorithm borrows this idea. However, in the FlowCytoVis the number of clusters is determined by the total number of data trends available in the dataset, including minuscule ones.

4.2. Modified feature-guided algorithm

Below is the step-by-step description of the algorithm used for the clustering in the FlowCytoVis. Smoothing histograms has a double purpose. It makes the histograms less noisy in visual appearance and is useful for clustering. The level of averaging can be made adjustable later but it is currently 5 points both to the right and to the left of the currently processed point.

1. Make each peak in each dimension a cluster base. The width of the cluster is determined by the tracing of the closest minimums to the left and to the right of the peak and taking the maximum width. The centroid of the cluster in this dimension is the coordinate corresponding to the peak maximum.

2. For the first dimension, distribute the events between the clusters determined by the peaks in this dimension. The event belongs to a cluster to which centroid it is the closest in this dimension.

3. Trace the direction of the events in the peak-based clusters of the next dimension.

3a. If all of the events from one of the clusters, organized by checking previous dimensions, go into another peak-based cluster of the next dimension, then merge these two clusters into one.

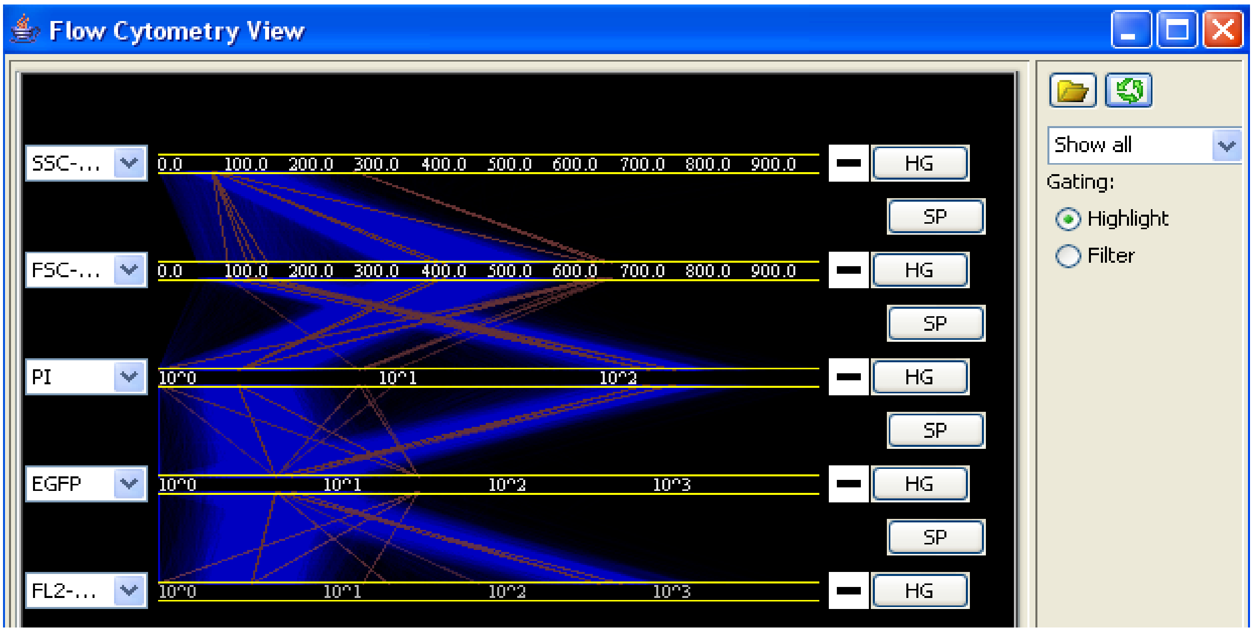
3b. Otherwise, organize as many clusters as the number of connected pairs between previously made clusters and the current peak-based clusters in the dimension being processed. Then, redistribute all relevant events between them according to proximity to centroids.

4. Go to step 3 until all dimensions have been processed.

Since the number of dimensions is limited to 20 maximum, it is acceptable to assume that the computational complexity of this algorithm is $C^2O(n)$, where $4 < C < 20$ represents a number of dimensions. Also, given that the number of dimensions greater than 10 is rarely used, this algorithm is appropriately fast for most of the needs the software might be used for.

5. Scenario of use

John Flow has finished his experiment and wants to check his results using FCM analysis. He goes to the computer and starts the visual tool. Through the clicking “open” button he loads from the file his negative control (cells that do not have fluorescence) dataset (Fig. 7). Its representation appears on his screen.



Show all

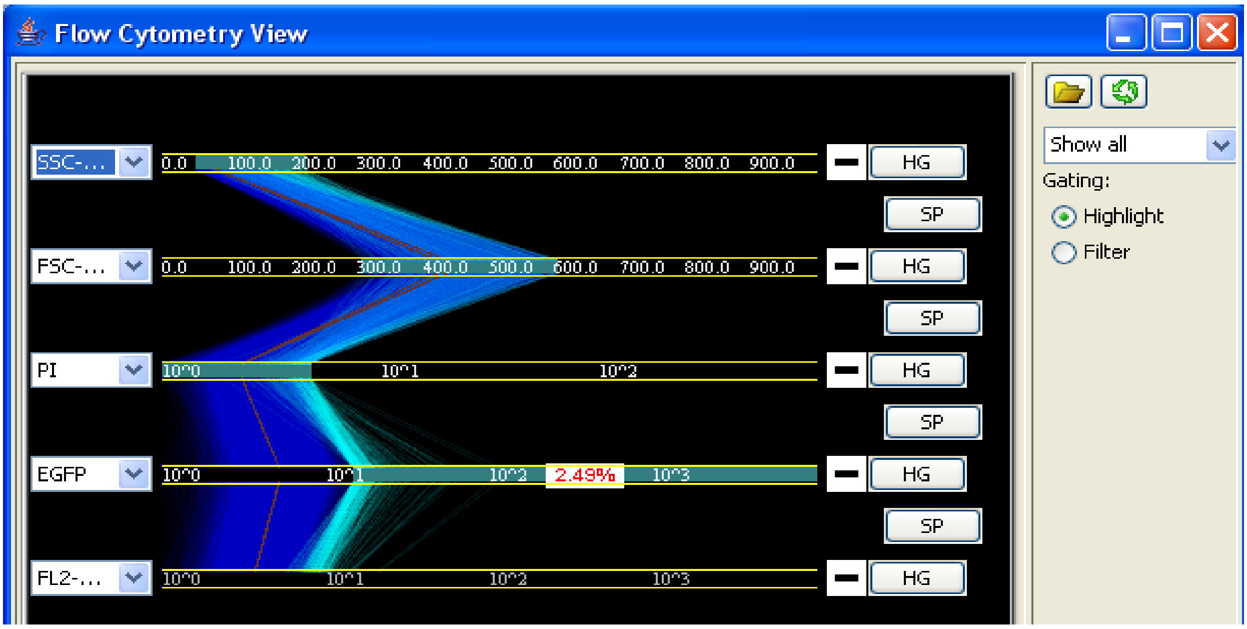
Gating:

● Highlight

○ Filter

Figure 7. Negative control dataset. Dark red lines represent clusters.

He looks at the data between the first two axes. These axes represent side-scattering/SSC (granularity) and forward-scattering/FSC (size). After a short glance, he creates the “gate” by selecting one interval on each axis with two mouse dragging movements. This filter separates now the distinct cellular cluster from the cellular debris, which has been now filtered out. Then John looks on the PI (dye intensity to distinguish living cells) axis and by the single mouse move continues the gate into the PI dimension. Now, dead cells are separated from living ones by the gate. Next, he highlights everything that is not negative control cell cluster on the GFP (green fluorescent protein) intensity axis and it will be the basis for comparison to the sample that includes positive result (Fig. 8).



Show all

Gating:

● Highlight

○ Filter

Figure 8. Negative control example. 2.49% will be an expected false positive rate.

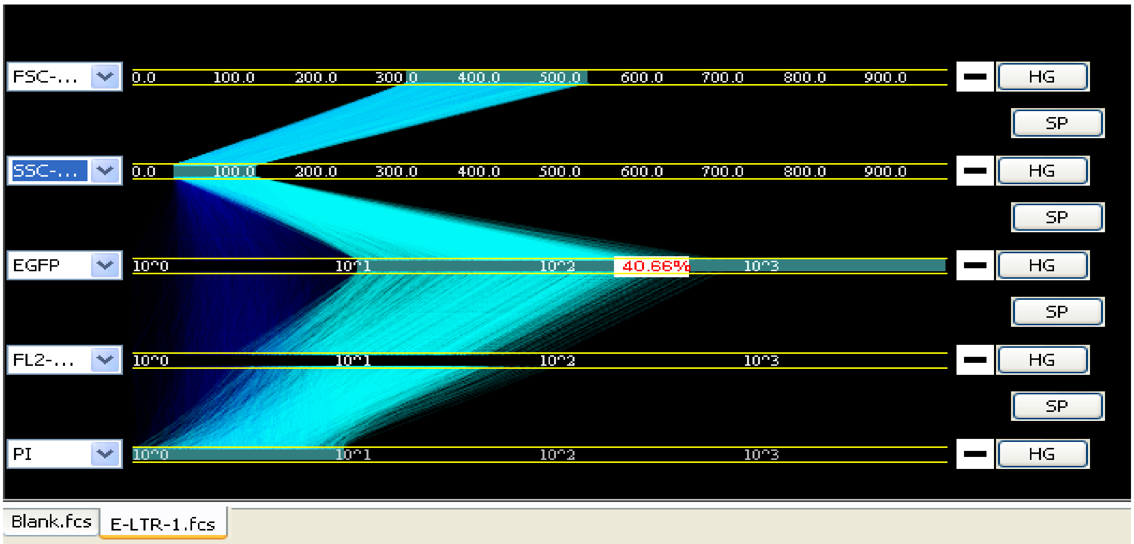


Figure 9. Positive results (40.66%) automatically applied to the negative control filters.

Then John one by one opens his several sample datasets (Fig. 9) in the FlowCytoVis. They appear in separate tabs. Since he has already set filters for the negative control dataset he does not need to repeat this operation anymore. Gates are already preset for each next dataset. As soon as data have finished loading, John looks at the percentage of the fluorescently marked cells in the gate, relative to the cells that do not show presence of the marker in the same sample. This is the result he wants to know but for his upcoming paper he also opens scatterplots and histogram views to save a few pictures of his results in different representations.

6. Conclusion and future work

In this work, the FlowCytoVis system is introduced. This is a new tool for visualization of the FCM data based on parallel coordinates technique with a clustering feature. The FlowCytoVis provides the interactive interface for convenient and intuitive data analysis.

It is important to note that the FlowCytoVis parallel coordinates solution has not been positioned to completely substitute scatterplots and histograms but rather be a powerful addition to them that can provide alternative possibilities of the data analysis.

The implementation revealed that the disadvantage of Java implementation of the parallel coordinates is a rendering speed for large datasets. Scatterplots take the lead here because the number of points needed to be drawn is significantly smaller. Clustering, in fact, helps in this case but clusters also need to be initially generated, also requiring substantial amount of time right after a dataset has been loaded. Thus, one of the future developments of this project is solving the rendering speed problem.

Clustering algorithm presented in this report provides decent results in determining clusters. However, it requires some improvement. Here goes the list of what still needs to be done.

1. Partially filtered cluster centroid does not appear on the screen if the centroid itself has not been caught by the filter. This creates some confusion in understanding if clusters have been found correctly.
2. The author is aware of a bug that sometimes provides the shift of the found cluster centroid along the axis.
3. Another existing issue that with provided algorithm it is possible to get the cluster with 0 events inside and it will be shown on the screen, fortunately, this issue is easy to fix.

Flow Cytometry Data Standards Project has also several clustering algorithms at the disposal that are worth integration into the FlowCytoVis system. Due to the very tight time limitation of the project they have not been considered but might be incorporated in the future.

The main apprehension before the start of the project was that parallel coordinates might not be intuitive enough for the users to use them without having to struggle to understand what they represent. Some FCM users, for example, stated that they associate dots on the scatterplots with actual cells and it could be hard to imagine cells as lines.

But parallel coordinates proved to be a relevant solution for such visualization and was accepted with enthusiasm by the researchers in the TFL. At the end, the FlowCytoVis prototype was introduced to some of the flow cytometry technology users. One of them shared the opinion that it is easy to realize the mapping of the scatterplot to parallel coordinates especially after few comparisons of the two types of representations and immediately suggested several features that exist in the FlowJo they would like to try with the FlowCytoVis. Another user of the FCM suggested using different highlighting colors to be able to see few more data trends for comparison.

On several datasets that were tested it was possible to see the expected trend before setting up all of the filters that usually required for the data analysis on the FlowJo system. For example, in one of the datasets, provided for testing selecting only the biggest cell size cluster showed that all of the cells of this size have roughly the same granularity, they do not glow (low GFP) and that they almost all alive (< 10 on PI axis). With the FlowJo the same result is impossible to se right away because it requires the manual selection of the axes to show each of them as a scatterplot.

During the informal discussion members of the Flow Cytometry Data Standards Project suggested that this tool might provide additional insights into the data besides the general FCM analysis and that they decided to investigate such possibility. Also they confirmed that suggested approach has a perspective to be widely used. This is an area of the future work.

The FlowCytoVis system can be improved in many ways and provides new opportunities for investigation how this visualization can help users to have better experience with the FCM data analysis. Thus, current project seems to have a future and this can be considered as a success.

References

[0] http://www.flowcyt.org, official website of the Flow Cytometry Data Standards Project
[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)
[6] Qing T. Zeng, Juan Pablo Pratt, Jane Pak, Dino Ravnic, Harold Huss, Steven J. Mentzer, Feature-guided clustering of multi-dimensional flow cytometry datasets, Journal of Biomedical Informatics, *In Press*, doi:10.1016/j.jbi.2006.06.005. (2006)
[7] Matthew O. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. Proc. of Visualization ’94, p. 326-33, (1994)
[8] http://www.ggobi.org, official web site of the GGobi data visualization system