Distribution Agreement

Signature:

__________________________ __________________

Kyle Thayer Date

Self-Organizing Maps and Wind-Rose Charts in the Visualization and Analysis of Flow Cytometry Data

By

Kyle Thayer
Masters of Science

Computer Science

Dr. Vicki Hertzberg
Advisor

Dr. Li Xiong
Advisor

Dr. Aneesh Mehta
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Self-Organizing Maps and Wind-Rose Charts in the Visualization and Analysis of Flow Cytometry Data

By

Kyle Thayer
B.S., Colorado State University, 2006

Advisor: Vicki Hertzberg, Ph.D
Advisor: Li Xiong, Ph.D

An abstract of
A thesis submitted to the Faculty of the James T. Laney School
of Graduate Studies of Emory University in partial fulfillment
of the requirements for the degree of Master of Science in
Computer Science
2011

Abstract

Self-Organizing Maps and Wind-Rose Charts in the Visualization and Analysis of Flow Cytometry Data

By Kyle Thayer

Using Flow Cytometry (FCM) technology, multi-faceted measurements can be taken of many individual particles, and thus it is often used in tissue sample analysis. As the resulting data set can have over ten dimensions and millions of points, analysis can be complicated and visualizing the data requires significantly reducing the number of dimensions or condensing the volume of the data. In studies where FCM data has been taken from multiple patients at multiple times, comparing these data sets complicates the analysis. I developed a software package (IFC Soft) for analyzing and visualizing FCM data using Self-Organizing Maps (SOMs) and Wind-Rose Charts (WRCs). I demonstrate the use of SOMs and WRCs on FCM data taken from two different transplant studies. In each study, blood samples were taken from post-op transplant patients at regular time intervals and FCM data were produced from each blood sample. The data were first pre-processed by medical researchers using FlowJo software to remove cellular debris and miscellaneous cells and save the general cell types for investigation. I used SOMs to visualize and manually cluster these cells from the different blood samples and save the amount of each cell type in each sample. I visualized these results using WRCs and SOMs to look for common trends among different classes of patients (eg. responded well/poorly to the transplant). The SOMs proved to be useful for selecting clusters of cells, but were difficult to use directly for finding patterns between patients because they displayed more information than could be easily managed and because of the variability of the patients. WRCs made from the cluster summaries were found to be more useful for finding patterns, but still provided too much information to easily extract patterns. The SOMs of the summary data, on the other hand, were easy to use in finding patterns among patient groups. The WRCs and original SOM could then be examined to confirm the pattern and see greater detail in each sample.

Self-Organizing Maps and Wind-Rose Charts in the Visualization and Analysis Flow Cytometry Data

By

Kyle Thayer
B.S., Colorado State University, 2006

Advisor: Vicki Hertzberg, Ph.D
Advisor: Li Xiong, Ph.D

A thesis submitted to the Faculty of the James T. Laney School
of Graduate Studies of Emory University in partial fulfillment
of the requirements for the degree of Master of Science in
Computer Science
2011

Contents

List of Figures

Chapter 1

Introduction

1.1 Flow Cytometry

Flow cytometry (FCM) is a technology in which multiple properties are measured for many individual cells [Bashashati and Brinkman, 2009] and is commonly used in medical science to count cell types within tissue samples. By using fluorochromes to label cell surface markers then passing the cells through an electronic device in which they are profiled optically by a series of lasers, the cells can be measured and even be physically sorted into phenotypic subpopulations according to the combinations of fluorochromes that are present or absent [Eisenstein, 2006]. Tremendous amounts of data can be produced from tissue and blood samples. Millions of cells for a single sample can be measured in a single pass with the FCM device, and each of those cells can have many channels measured. For instance, BD Biosciences now offers an instrument with five lasers that can distinguish up to 18 different fluorochromes [BD Biosciences, 2009]. Hence 262,144 (2^{18}) different cell populations could be distinguished based solely on the presence or absence of different fluorochromes.

Working with the vast amounts of multi-dimensional data that FCM generates can be difficult. The classical method of analyzing FCM data is to examine one or two channels (also called parameters or markers) at a time using histograms, scatter plots and contour maps. Subsets of the cells are selected using gates with software packages such as FlowJo [TreeStar Inc., 2011]. Gates are regions of the plots that are used to select cell populations of interest. These cell populations are then typically shown in a different pair of dimensions and more gates are made. This process of selecting subpopulations through gating, then making new displays of the subpopulations is repeated until the most important subpopulations are found. Statistics about these populations are then saved and used in analysis.

Gating is often performed manually by visual inspection. This approach is time-consuming, requires significant knowledge and experience, lacks standardization and is subject to human error [Bashashati and Brinkman, 2009]. Beyond the disadvantages of being tedious and subject to operator error and intra- and inter-operator variability, only two dimensions are examined at a time and important multi-dimensional patterns may be missed. Recently, some clustering algorithms and tools have been developed for FCM data that cluster cells into cell groups using all dimensions (channels) at once [Bashashati and Brinkman, 2009, Scheuermann et al., 2009]. However, most of these tools or algorithms provide limited visualization capabilities or still require significant user intervention and hence limited utility.

The next challenge comes in comparing different FCM data samples, whether from the same subject at different times or between subjects and groups of subjects. Standardizing the FCM process and calibrating the devices is a challenge as well; samples run by different operators and at different times can vary significantly [Maecker et al., 2010]. Thus a gating that is valid for one sample, may not work on another. Programs like FlowJo allow the user to adjust gates for each sample to fit the calibration differences [TreeStar Inc., 2011]. Once the cell counts for the different subpopulations are taken, these values may be compared visually with bar charts, histograms or pie charts.

1.2 Contributions

Given the difficulty in processing FCM data and the limitations of two-dimensional analysis methods, new methods of visualization and analysis are needed.

Self-Organizing Maps (SOMs) can display large amounts of multi-dimensional data without ignoring or reducing any of the dimensions. Additionally, the SOM organizes the data in a way that is convenient for finding and selecting clusters of cells in FCM data. In this paper, SOMs are used to visualize cells from a number of blood samples, simultaneously showing the cell clusters and which of each type of cell the blood samples had. This allows detailed comparison of the different blood samples, but contains too much information to quickly find patterns. Instead clusters are selected based on the SOM and statistics were saved on how many cells of each type the blood samples had.

Wind-Rose Charts (WRCs) are used to display this summary data. Because of the common orientation of the dimensions (representing amount of cell types), comparisons of the same dimension across blood samples are simple. Additionally, since WRCs display their value as a length rather than an area, more accurate comparisons are possible than with pie charts.

SOMs are then used to display the same summary information as the WRCs. These graphs prove effective for discerning commonalities and differences between types of patients and changes over time. These patterns can then be verified and examined in more detail on the earlier WRCs and SOMs.

The SOMs additionally give a signature hit histogram for blood samples and summary statistics of different groups of patients. This signature can be used to compare new patient blood samples to the previously studied patients.

Chapter 2

Related Work

2.1 Wind-Rose Charts (WRCs)

Wind-Rose Charts are one of several closely related circular charts, all of which are essentially based on wrapping a bar chart around a circle. The data values are either represented by the area of the sectors or the length of the sector from the center of the circle. These graphs have a long history of use. In 1829 André-Michel Guerry made polar-area charts [Friendly, 2009a]. In 1843, Léon Lalan drew the first Wind-Rose diagram to show the amount of wind in different directions [Friendly, 2009a]. Florence Nightingale made her own polar-area chart (or "coxcomb diagram") in 1857 to show causes of death in the British Army [Friendly, 2009b]. Modern meteorological wind-rose charts (Fig. 2.1) show the frequency and strength of wind in the different compass directions.

These circular charts have also been known as "Circular Column Graphs," "Sector Graphs," [Harris, 1999], "Sector Charts" in Mathematica [Wolfram Research, Inc., 2010], and "Stars" with draw.segments enabled in R [R Development Core Team, 2008].

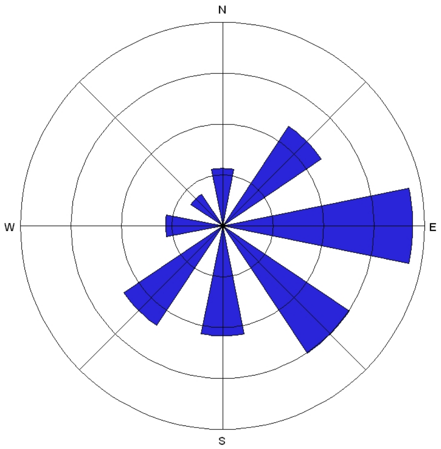

Figure 2.1: Wind-Rose Chart generated in SAS^{TM}software [SAS Institute Inc., 2011].

These circular charts have significant advantages over the traditional pie charts and bar charts. Unlike pie charts, which are difficult to compare due to the human eyes difficulty in discerning subtle differences in angles [Helsel and Hirsch, 2002], relative sizes are based on the length (or area, which is proportional to the square of the length) of the sectors. These also have an advantage over bar charts in the ease of comparing related plots since a consistent direction is also associated with the sector. See figure 2.2.

2.2 Self-organizing maps (SOMs)

SOMs [Kohonen et al., 2001] comprise a special type of neural network that provides a mapping of high dimensional data points, such as cells in FCM data, to a reduced output space consisting of a regular maps of nodes, usually a 2D array. The underlying algorithm preserves, as far as possible, the spatial relationship among points in the input space (topology conservation). In other words, data points close to each other in the input space will generally be

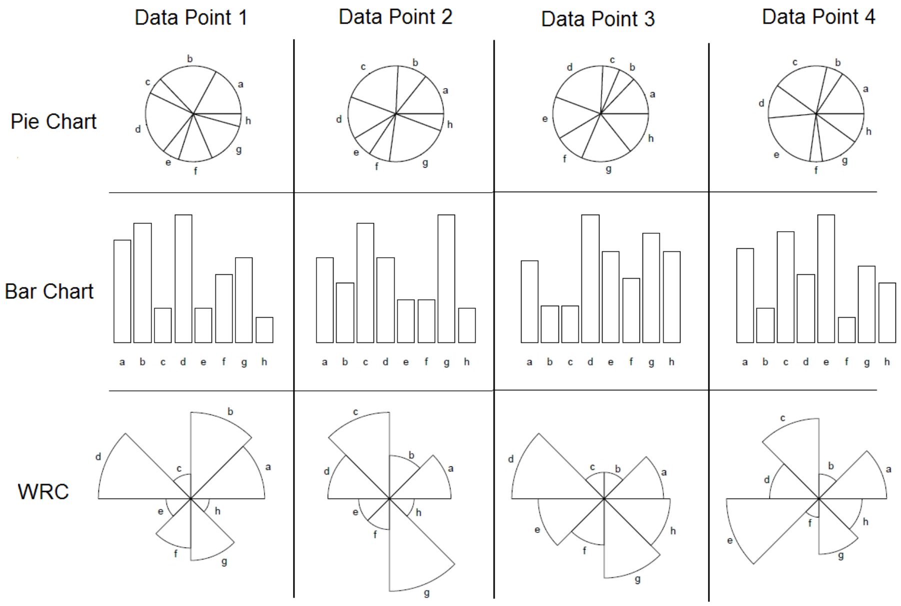

Figure 2.2: Comparison of Pie Charts, Bar Charts and WRCs. It is difficult to find and compare the value of "d" in each pie chart. While more accurate comparisons can be made in the bar chart, it is difficult to quickly find "d" in each. The WRCs allow for fast, accurate comparisons of all values.

mapped to nodes that are close to each other in the output map.

If there are more data points than SOM nodes, then some nodes will represent multiple data points. Thus, the SOM condenses the data volume into a smaller number nodes, allowing this simplification of the data to be effectively displayed.

Each node of the SOM has weights in each of the dimensions of the input data. The algorithm uses the calculation of Best Matching Units (BMUs) in its process of generating the map. For a given data point, the BMU is the node in the SOM that is the closest to the data point according to some metric. The most common metric, and the one used in IFC Soft, is that of Euclidean distance. BMUs are used both in the organizing process of the SOM and in visualizing and clustering the data based on the SOM once it is produced.

The SOM algorithm uses random selections of data points from the data set to produce its organization. This process is entirely automated and does not need or accept any user input, and therefore is not subject to human variability.

If the SOM is a 2D array and the data has over two dimensions, there are two straightforward ways of displaying the nodes once the organization process has finished. The first is to display the node array with a graph in each node, showing the node's weights (e.g. A bar graph as in Figure 2.3 below). The second is to, for each dimension, display the node array with colors based on their weight in that dimension (Figure 2.4).

2.2.1 The Incremental SOM Algorithm

The original algorithm for producing SOMs was the Incremental SOM Algorithm. In this algorithm, the organization is achieved through a large number of iterations, each with a single data point.

For each iteration, a data point is picked randomly from the input data set

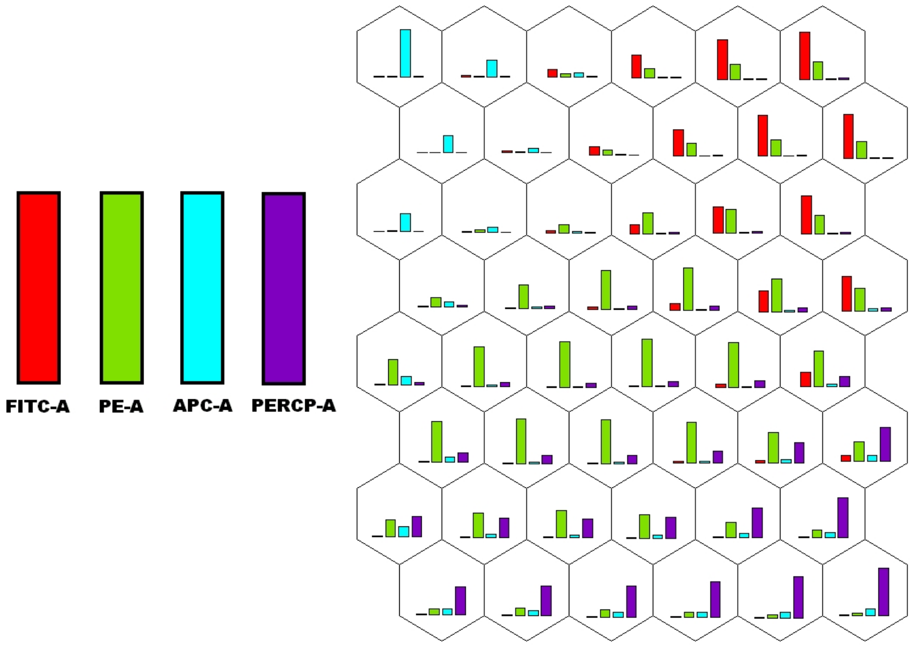

Figure 2.3: A bar graph representation of an SOM made from FCM calibration beads. Made in Matlab®[MATLAB, 2010] with SOMToolbox [anb Johan Himberg et al., 2005].

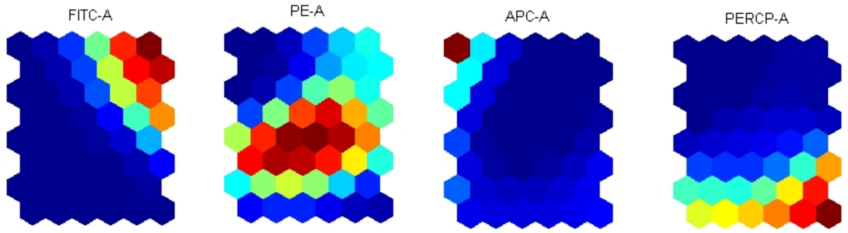

Figure 2.4: The same SOM as Figure 2.3, but the SOM is displayed one time for each dimension. Red is high in the given dimension, and blue is low. Made in Matlab [MATLAB, 2010] with SOMToolbox [anb Johan Himberg et al., 2005].

and its BMU is found. The weights of the BMU node and the nodes surrounding it are shifted towards the weights of the data point.

When the algorithm starts, the neighborhood that is influenced is very large and the amount that the nodes are shifted towards the data point value is large. As the algorithm continues, the neighborhood shrinks so that only a few nodes are changed, and they are only changed a very small amount. This continues until the map stabilizes.

2.2.2 The Batch SOM Algorithm

An alternative to the Incremental SOM Algorithm is the Batch SOM Algorithm. This algorithm needs a much smaller number of iterations than the incremental SOM does, and each iteration uses a large number of input data points.

For each iteration, the BMU is calculated for a large number of data points randomly picked from the data set. Each node then has a set of data points that belong to it. After all the points have been placed for the iteration, each node's weights are replaced with the average of all the points that belong to itself and the nodes in its neighborhood.

At the start of the algorithm, the neighborhood is large. This means that not many data points are needed since the average of all data points that landed in the neighborhood of the node are averaged together, and that could be hundreds of nodes. At the end of the algorithm, the neighborhood will be small and the number of points needed will be large so that a reasonable average can be taken of the points in a node's neighborhood. If the neighborhood is only the node itself, then this is equivalent to a step of k-means.

2.2.3 SOM Feature Maps: U-Matrix and Hit Histograms

The U-Matrix (see Fig. 2.5) is a map showing where differences between neighboring nodes occur. This can either be done as an average, where the U-Matrix value of each node is the average distance it is from all its neighbors, or it can be done for each neighboring edge, so that each edge has the value of the distance between two nodes. This has been used in attempts to automatically cluster data sets [Kohonen et al., 2001].

The hit histogram (see Fig. 2.5) shows what nodes in the SOM data points land on. After the organization process has finished, each data point in the data set is placed on the SOM Node it is closest to (its BMU). The number of data points each node has is displayed in the hit histogram, and which data points belong to which nodes is used for selecting subsets of the data from the SOM. Hit histograms are useful for comparing different subpopulations of the data [Vesanto, 1999]. By making an SOM of several subpopulations together, displaying the separate hit histograms of the subpopulations shows where the subpopulations fit on the SOM, and thus what characteristics they have compared to the other subpopulations.

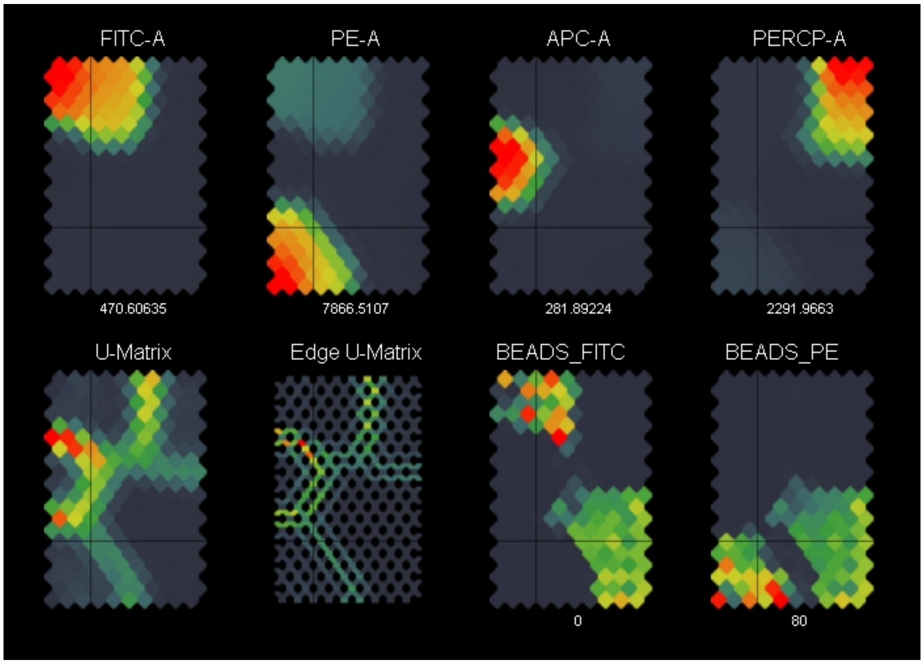

Figure 2.5: An SOM of several FCM calibration beads data sets. Both the U-Matrices for nodes and for edges are displayed. The hit histograms for two data sets are shown. The first (BEADS_FITC) has FITC+ beads and unstained beads (low in all four dimensions). There are none on the selected node. The second (BEADS_PE) has PE+ beads and unstained beads. There are 80 on the selected node.

2.3 Wind-Rose Charts used for summarizing FCM data

Scarberry [Scarberry, 2009] used WRCs to visualize FCM data in her MS Thesis written under the direction of Dr. Hertzberg (Biostatistics Department, Emory University). The FCM data had the amounts different cell populations from transplant patients in the PIP study. There were three groups of patients: controls, transplants receiving thymoglobulin and transplants not receiving thymoglobulin. Each patient was measured at several different times. She tested different methods of visualizing the data in SAS/Graph^{TM}software [SAS Institute Inc., 2011], namely: radar plots, wind-rose charts, star charts and pie charts. Scarberry concluded that radar plots were the most effective tools for displaying these data, but much of this was due to the limitations in how WRCs were implemented in SAS/Graph software.

After this, Dr. Hertzberg and Lisa Elon (Biostatistics Department, Emory University) began using the stars function in R [R Development Core Team, 2008], which produced WRCs that better fit their visualization needs. They placed the different time points from the patients next to each other and were able to quickly see how patients were changing over time and how that compared to other patients.

Before I started developing IFC Soft, my first task for the PIP study was to take the R script made by Lisa Elon for generating the WRCs of the clusters from the FCM data, automate it, and make it handle missing data points.

2.4 Self-Organizing Maps used with Flow Cytometry

In spite of the ability of Self-Organizing Maps to cluster and analyze multidimensional data, they have rarely been used in analyzing FCM data. In 1996, SOMs were used to display the results of different clustering methods used on FCM data of phytoplankton [Wilkins et al., 1996]. In the paper, SOMs were never used generating clusters, but merely for displaying how other methods clustered the data. In 2001, in another study of phytoplankton, SOMs were used for clustering FCM data [Grégori et al., 2001]. A small SOM (three by four nodes), was created from a data set of calibration beads and different types of phytoplankton. Separate hit histograms were displayed for the beads and the different types of phytoplankton, showing which nodes captured which type of phytoplankton or bead. This was shown to be an effective way of separating and counting the different types of phytoplankton.

In 2010, for a study on brain tumors, SOMs were used to analyze cell samples from the tumors [Sun et al., 2010]. Their methodology is similar to the one I use in this paper, though they focused on automated clustering of the patients, while my focus is on visualization and interaction. They only had one sample from each patient and instead of using FCM data, they used Microfluidic Image Cytometry, which produces similar data, but for fewer cells (on the order of thousands). They collected cells from 19 different patients and produced an SOM for the cells from all the patients. From the SOM, they generated hit histograms for each of the different patient samples showing how the cells in each sample compared. For each sample, the Neighborhood Frequency Vector (sum of the frequency of a node in and its neighbors in the hit histogram) was computed for all the nodes on the SOM. This was then used in an unsupervised clustering algorithm to cluster the patients. They were able to characterize the hit histograms for each of these clusters and found that they clusters correlated to survival of the patient.

Chapter 3

IFC Soft

3.1 About

As part of this project, an open-source web-application was developed to implement our strategies for visualizing and analyzing flow cytometry data. The program was written in Java and JavaFX, which were chosen for their combined features of GUI design and multi-threaded processing. IFC Soft is currently under active development on version 0.4. It can be run online at `http://www.ifcsoft.com` and the source code is available at `http://code.google.com/p/ifcsoft/`.

3.2 Features

The current version of IFC Soft (v0.4), supports importing and exporting of Comma Separated Variable (CSV) files. Thanks to code donated by Tjibbe Donker from the FloCK project [Donker, 2009], IFC Soft can also directly import FCS files.

For visualization, IFC Soft uses SOMs and WRCs along with rudimentary implementations of histograms and scatter plots.

SOMs can be calculated from any combination of data sets (as long as they have the same number of dimensions) using either the incremental or batch SOM algorithm. The different dimensions of the input data can be weighted or disabled for the construction of the SOM. Algorithm parameters, such as node array size, number of iterations and minimum neighborhood size can be set as well.

For displaying the SOM, the node array is displayed separately for each of the dimensions (as in Fig.2.4), and again for the U-Matrix and hit histograms for each of the data sets used in calculating the SOM. These are displayed against a black background with the node color ranging from dark blue (lowest) up to bright red (highest). This color scheme was chosen to conform to other displays of SOMs and to maximize the effectiveness of viewing on a screen. When the cursor is held over one of the copies of the SOM array, cross-hairs are shown on all of them to allow rapid comparison of the same node in all the copies of the SOM array (see Fig. 2.5).

Groups of nodes can be selected (Fig. 3.1) and the program will display the average value of the nodes under each dimension and the percentage of each data set contained in those nodes under the appropriate hit histograms. The percentages of each data set's points in the given selection can be saved internally for use in WRCs or further SOMs, and the actual data points that fit in the selected area can be saved internally or to a file.

WRCs can be used to display the data produced by selecting clusters in the SOM (Fig. 3.2). Like the SOM, they are placed against a black background to allow for easier reading on screen. The values of the data point can either be represented by the areas of the pie wedges, or by the length of the pie wedges. Using the length works better for comparing larger values, while using area makes comparing smaller values easier. The pie wedges are all exploded (moved away from the center) by a small amount to make it easier to see small wedges.

Though IFC Soft is written specifically for handling FCM data, I attempted to make it useful for analyzing other multi-dimensional data sets.

Some modifications were made to IFC Soft to handle the large number of data files for this paper (eg. allowing re-sizing of wind-rose charts). These changes will be incorporated into a future version of IFC Soft.

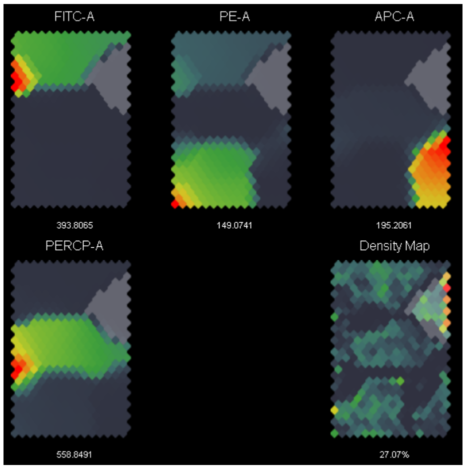

Figure 3.1: SOM with nodes selected

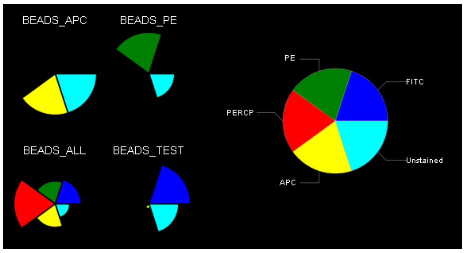

Figure 3.2: Wind-Rose Chart of summary data. Each wedge area represents the percentage of beads of the given type.

Chapter 4

Using IFC Soft to Analyze FCM data

4.1 Protective Immunity Project (PIP)

A study on renal transplant patients was done as part of the Protective Immunity Project (PIP) grant (a biodefense contract from the National Institute of Allergy and Infectious Diseases to study the basic biology underlying protective immunity in renal transplant recipients). Potential renal transplant candidates were approached for consent. In addition, a set of control subjects was recruited from the set of their kidney donors and healthy volunteers. Peripheral blood mononuclear cells (PBMCs) were collected from each patient in a CPT tube at the time of transplant (baseline) and at months 3, 6, 9, 12, 18, and 24. After centrifugation and red cell lysis, mononuclear cells were isolated and stained with monoclonal antibodies suitable for polychromatic flow cytometry. PBMC samples were processed by the FCM machine and FCS files were generated. In FlowJo [TreeStar Inc., 2011], the CD8 T-lymphocytes were selected as those having positive CD8 and CD3, and negative in CD4. In addition to the PBMC samples, demographic and clinical data (e.g., occurrence of infection, transplant medications) are collected on subjects for two years. All data were deidentified.

4.2 Lung Transplant Data Set

The lung transplant study provided a data set of FCM data for analysis. Patients undergoing lung transplantation were enrolled in the Emory Immune Monitoring Protocol. This is an IRB approved protocol in which informed consent is obtained. At regular time points (2 weeks and 1, 2, 3, 6, 9, 12 Months), post transplant peripheral blood was obtained from each patient in a CPT tube. After centrifugation and red cell lysis, mononuclear cells were isolated and stained with monoclonal antibodies suitable for polychromatic flow cytometry. Three panels were run on the blood samples (each panel giving an FCS file with a different combination of biological markers), but only the 8c panel is discussed here. In FlowJo [TreeStar Inc., 2011], CD8 T-Lymphocytes were selected as those having positive CD8 and CD3 and exported to FCS files. Data from ten patients were used in the analysis in this paper, five having good graft function (having a Forced Expiratory Volume (FEV1) of more than 80 percent peak post transplant lung function at 18 mnths post transplant) and five with poor graft function (having less than 50 percent of peak pst transplant function at 18 months). All data were deidentified.

4.3 Pre-Processing

The medical researchers (Dr. Aneesh Mehta for the PIP study and Dr. David Neujahr for the lung transplant study) took the raw FCM data from the blood samples and, using manual gating in FlowJo [TreeStar Inc., 2011], selected the

CD8 T-lymphocytes. Dr. Mehta used five channels in his gating process (FSC, SSC, CD3, CD4 and CD8) leaving four unexplored channels in the data (CD27, CD28, CCR7 and CD45ra). Dr. Neujahr used four channels (FSC, SSC, CD3 and CD8), leaving seven unexplored channels in each of the three panels run. I will only demonstrate the analysis of panel 8c for the lung transplant study, which had Perforin, CD127, HLADR, CD4, CCR7, CXCR3 and Granzyme B as the seven unexplored channels. The CD8 lymphocytes from each patient were exported from FlowJo into separate FCS files (one per blood sample).

The FCS files of the data were loaded into IFC Soft. Since often the biological markers weren't measured using the same channel and FlowJo outputted the data columns in different orders, the columns had to be rearranged, so that the biological markers matched. Once fixed, the data were saved into CSV files.

4.4 SOM of All Cells

To compare all the CD8 T-lymphocytes in the blood samples, for each study I combined all the data sets from the study and made an SOM of all the cells combined using the channels mentioned above (Fig. 4.1 for lung transplant study, and Fig. 4.2 for PIP study). This is only meaningful if the scales on all the data sets are the same and if the machines are calibrated properly. If the machines are not calibrated, then cells of the same type in different runs of the machine will probably get mapped to different parts of the SOM. I used the Batch SOM algorithm, since that was recommended as a faster alternative in SOM book by Kohonen [Kohonen et al., 2001], and set the node array size to 15x15. If the node array was much smaller, nodes could start combining different clusters of cells, which would prevent accurate cluster selection. If the node array was much larger, then besides being harder to read, small differences in scales could cause the same population in different files to be mapped to different nodes; having fewer nodes forces them into the same spot on the SOM. The algorithm was set to have the minimum neighborhood distance be 2, since in my experimentation, if it was smaller than that, small populations with very large values would get a node all to themselves and would throw off the scaling of the channels (eg. Fig. 4.3).

After building the SOM on all the cells, the hit histograms for each blood sample were displayed (Fig. 4.4 for lung transplant study, and Figs. 4.5 and 4.6 for PIP study).

The blood samples from the ten patients in the lung study reasonably fit on one page and can be compared. This display reveals the large variation both between different patients and in individual patients over time, though general differences between the groups of patients are hard to find. The amount of information displayed at once is too much to be able to quickly identify trends between the patients. While the ten patients in the Lung study did fit well on a single page, the hit histograms for the PIP project (163 blood samples) did not fit well on one page. It would take several pages to be able to be able to line up the times and show all the data. In order to find patterns between the different classes of patients, I had to do something to simplify the data further than the SOM already had. The next step was selecting clusters of the cells.

4.5 Selecting Clusters of Cells

In order to simplify the data for further visualization and analysis, I wanted to pick different sections of the SOMs as representing separate clusters of the cells. While the U-Matrix clearly marked borders between populations in the FCM beads (Fig. 2.5), it did not prove useful in this FCM data (eg. Fig. 4.7). This seems to be due to the variance within each natural cluster and the overlapping of clusters in the raw FCM data.

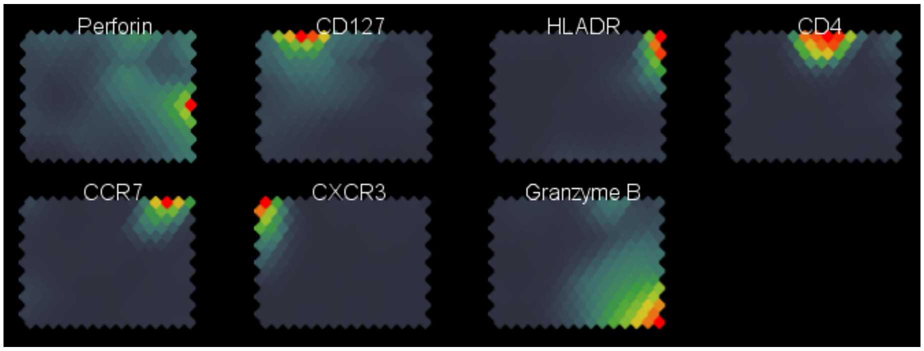

Figure 4.1: SOM of all the CD8 T-lymphocytes from the lung transplant study

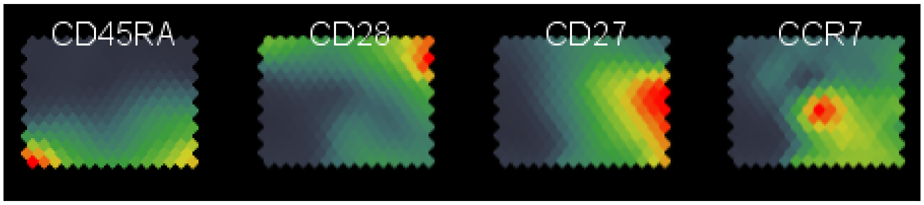

Figure 4.2: SOM of all the CD8 T-lymphocytes from the PIP study

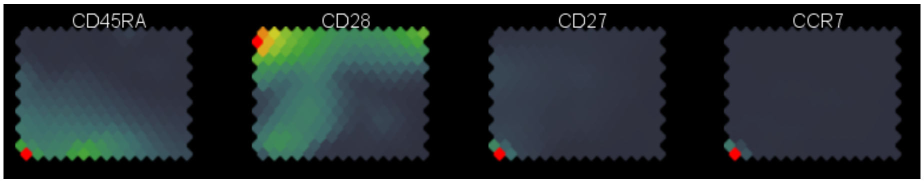

Figure 4.3: SOM of all the CD8 T-lymphocytes from the PIP study made with a minimum neighborhood distance of 1. Notice that a small, extreme population in the bottom left node throws off the scales of CD27 and CCR7.

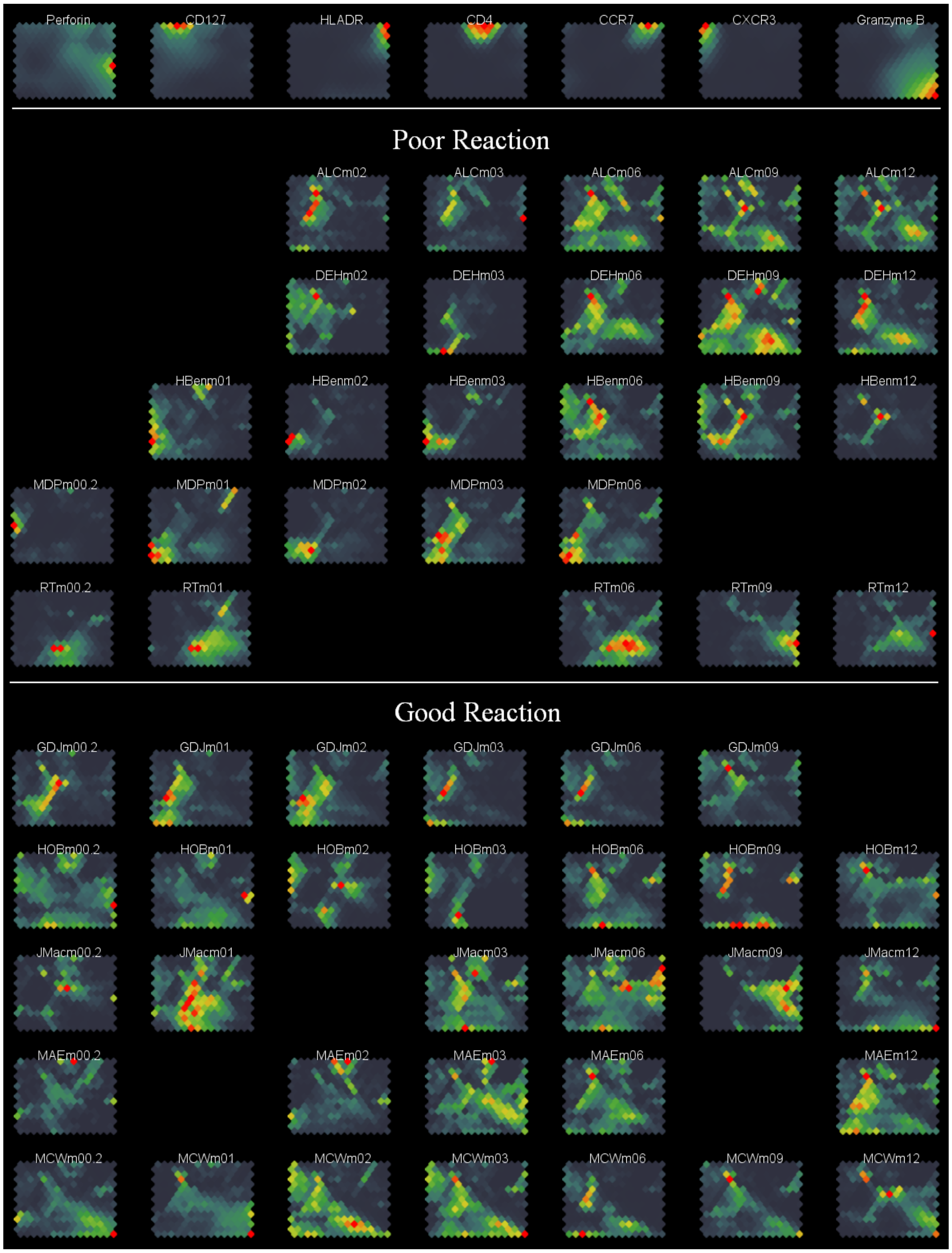

Figure 4.4: SOM of all the CD8 T-lymphocytes from the lung transplant study with hit histograms. The columns are time points (M00.2 = Week 2, M01 = Month 1, etc.), and the rows are patients. The top five patients reacted poorly to the transplant and the bottom five reacted well.

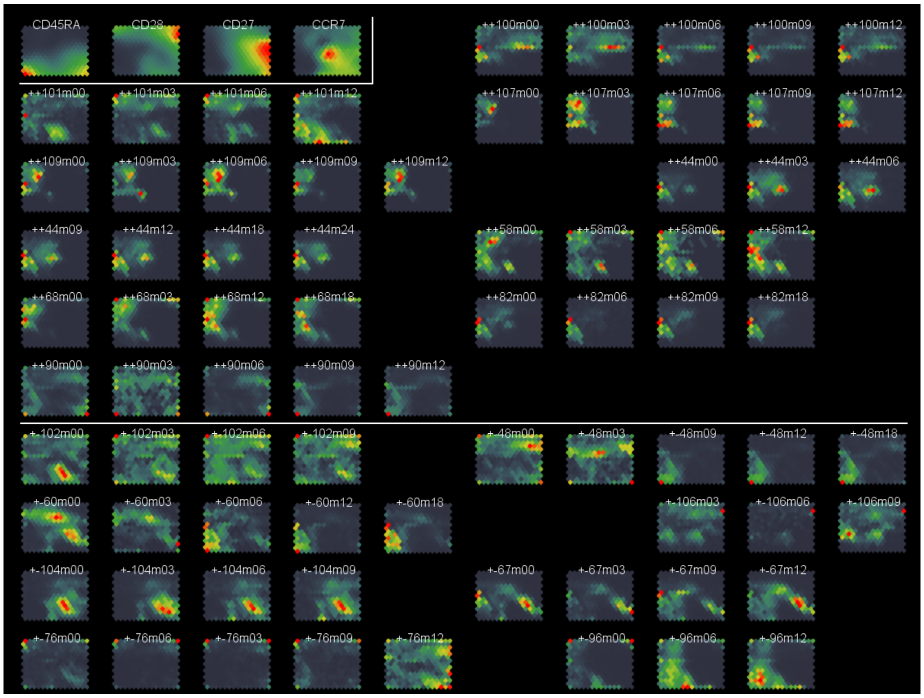

Figure 4.5: SOM of all the CD8 T-lymphocytes from the PIP study with hit histograms. Each patient has all their blood samples in order by time and the patients are divided into groups by Donor/Recipient CMV Status. This graph: +/+ and +/-. Fig. 4.6 has -/+, -/-, and control.

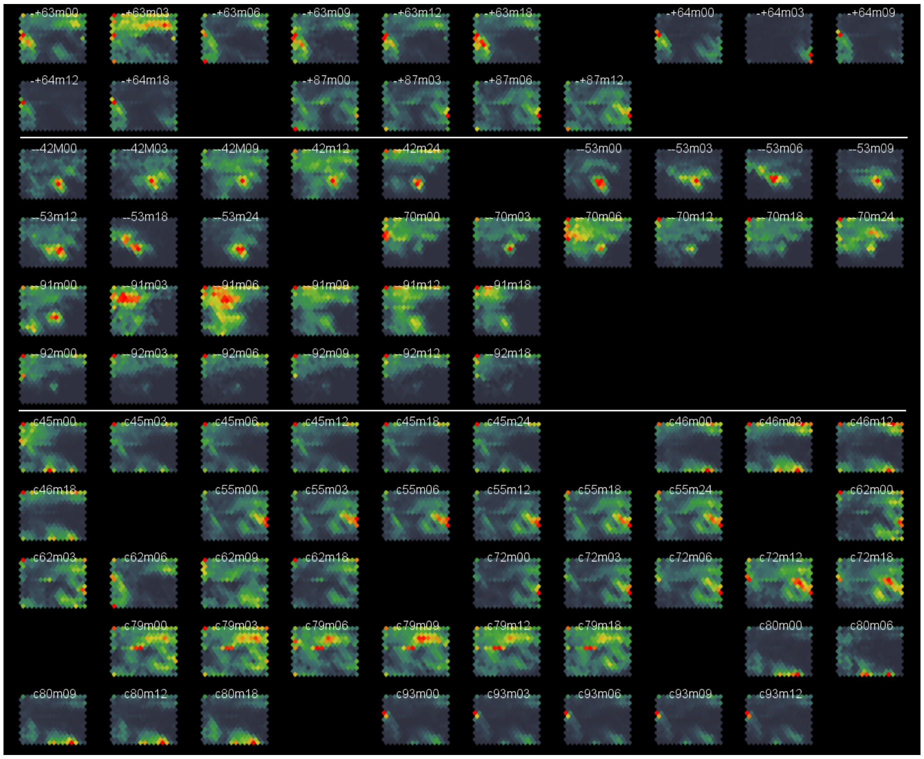

Figure 4.6: SOM of all the CD8 T-lymphocytes from the PIP study with hit histograms. Each patient has all their blood samples in order by time and the patients are divided into groups by Donor/Recipient CMV Status. This graph: -/+, -/-, and control. Fig. 4.5 has +/+ and +/-.

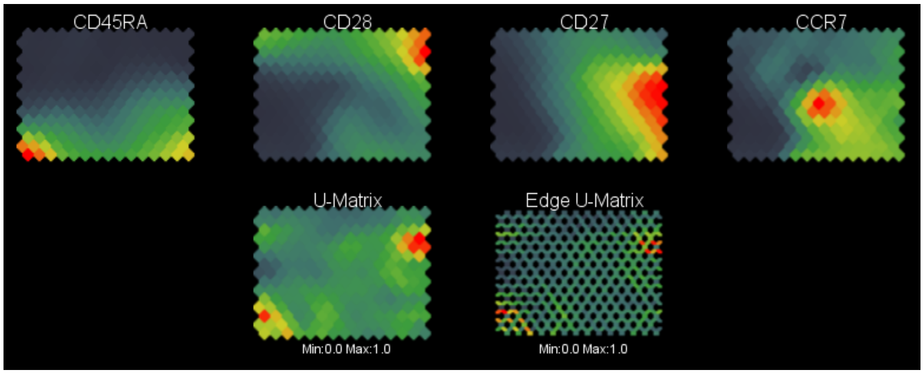

Figure 4.7: U-Matrix and Edge U-Matrix for the PIP SOM.

Instead of using the U-Matrix, I manually selected clusters based on areas where different markers were present or absent (See Fig. 4.9 for clusters on the lung transplant data and Fig. 4.8 for the PIP data clusters). For example, in the PIP data in Fig. 4.8, there are two areas that are low in CD45RA and high in CD28. One is high in both CD27 and CCR7 (Early diff. central mem.) and the other is low in CD27 and CCR7 (Late diff. effector mem. 28+). In the PIP study, after selecting the clusters I had one of the medical researchers (Dr. Mehta) determine medically relevant names for the clusters. For the lung transplant data, I just used names based on which channels were high or low.

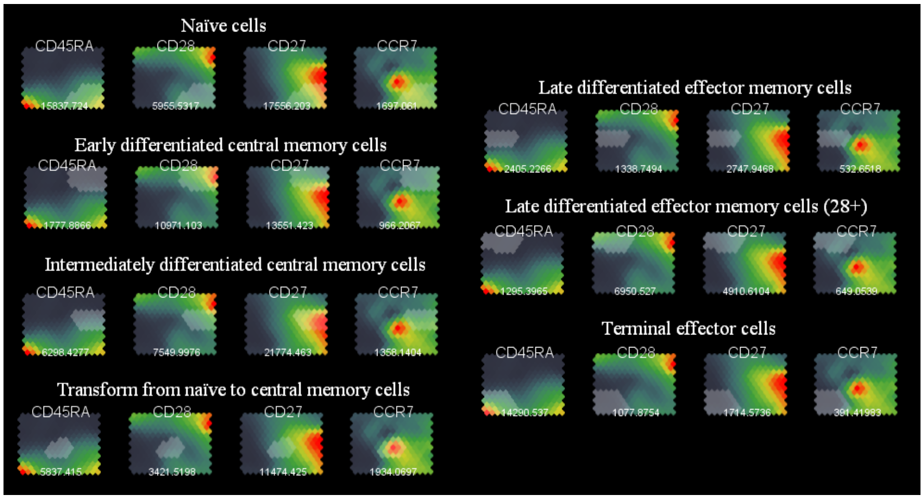

Figure 4.8: Clusters selected from PIP cells

With each cluster I selected, I saved all the data points in the cluster to a file and had IFC Soft save the percentage of each blood sample's cells that were in the cluster.

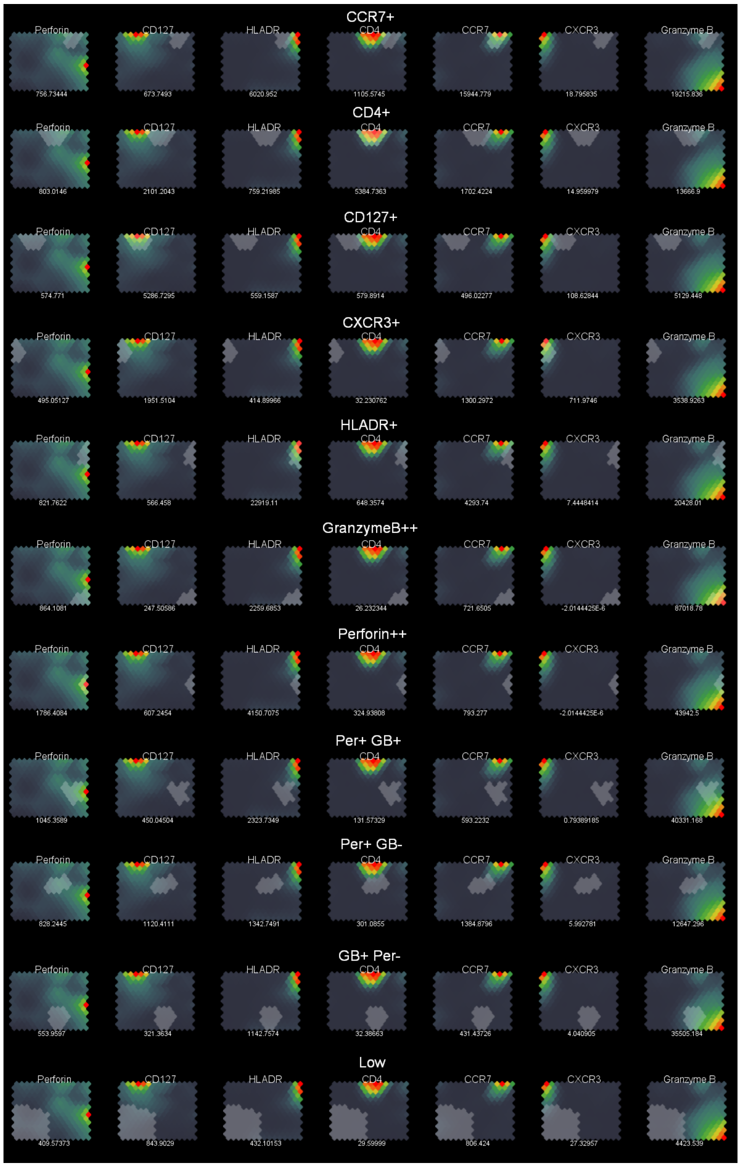

Figure 4.9: Clusters selected from lung study cells

4.6 Using Wind-Rose Charts to Compare Cluster Summaries

I displayed the statistics of how much of each cell type the blood samples had on Wind-Rose Charts (Fig. 4.10 for lung study, and Figs. 4.11 and 4.12 for pip study). Some patterns can be seen with these charts (eg. the amount of "naïve -> central mem" cells in the "-/-" samples in the PIP data). While it is easier to try to find patterns in the data on the wind rose charts than it was on the SOM of all cells, it is still quite difficult, especially with all the data points in the PIP study. There still seems to be too much information for quickly discovering patterns, so I saved this summary data out to a CSV file to prepare it for a new SOM.

4.7 Displaying Summary Results on an SOM

Since IFC Soft does not yet support categorical data, the best way to separate categories is to have separate files with the data points for each category. I took the summary files of the clusters from the original SOM and created separate files with just the data points for each patient type and time.

I loaded these files back into IFC Soft and made an SOM of all the blood sample summaries. (Fig. 4.13 for lung transplant study, Fig 4.14 for PIP). In this SOM, the data points are blood samples from a given patient at a given time. They are organized by how much of their cells were in each cluster type. The hit histograms show where the data points from each group ended up, so therefore what that patient's blood samples were like.

Finally, with this visualization, patterns are relatively easy to find. For example, in the lung transplant study SOM, there is an area circled at the top-right (Fig. 4.13). This area represents blood samples that had high percentages

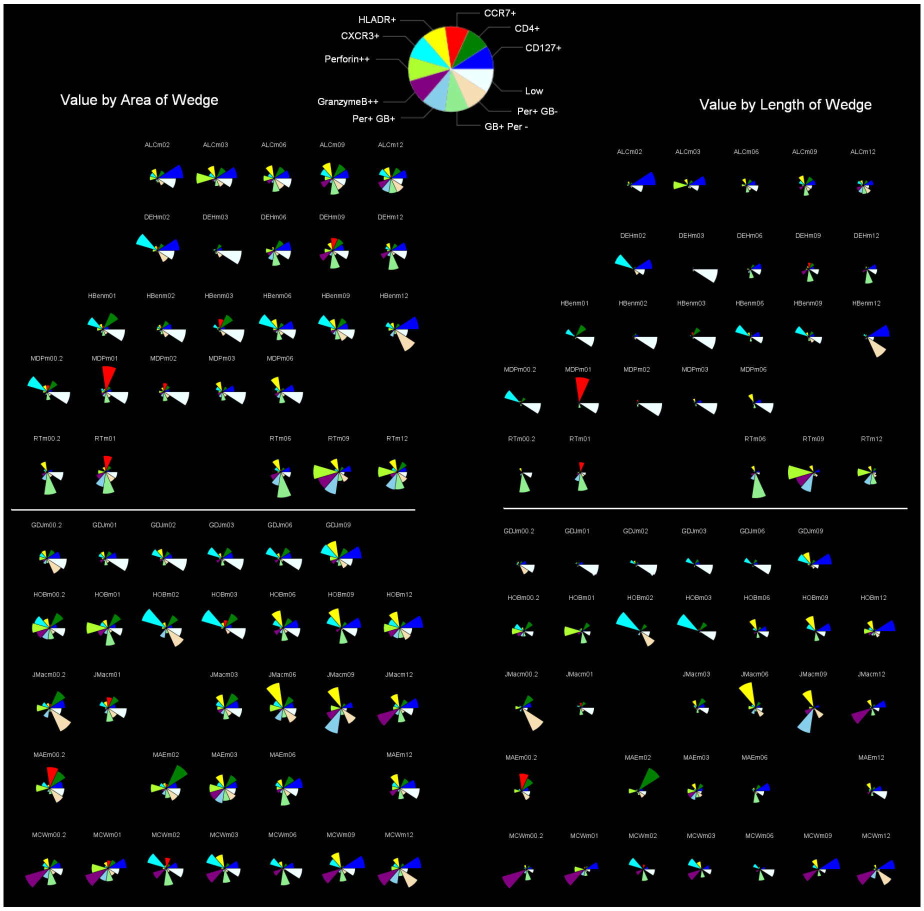

Figure 4.10: Wind Rose charts from lung transplant study

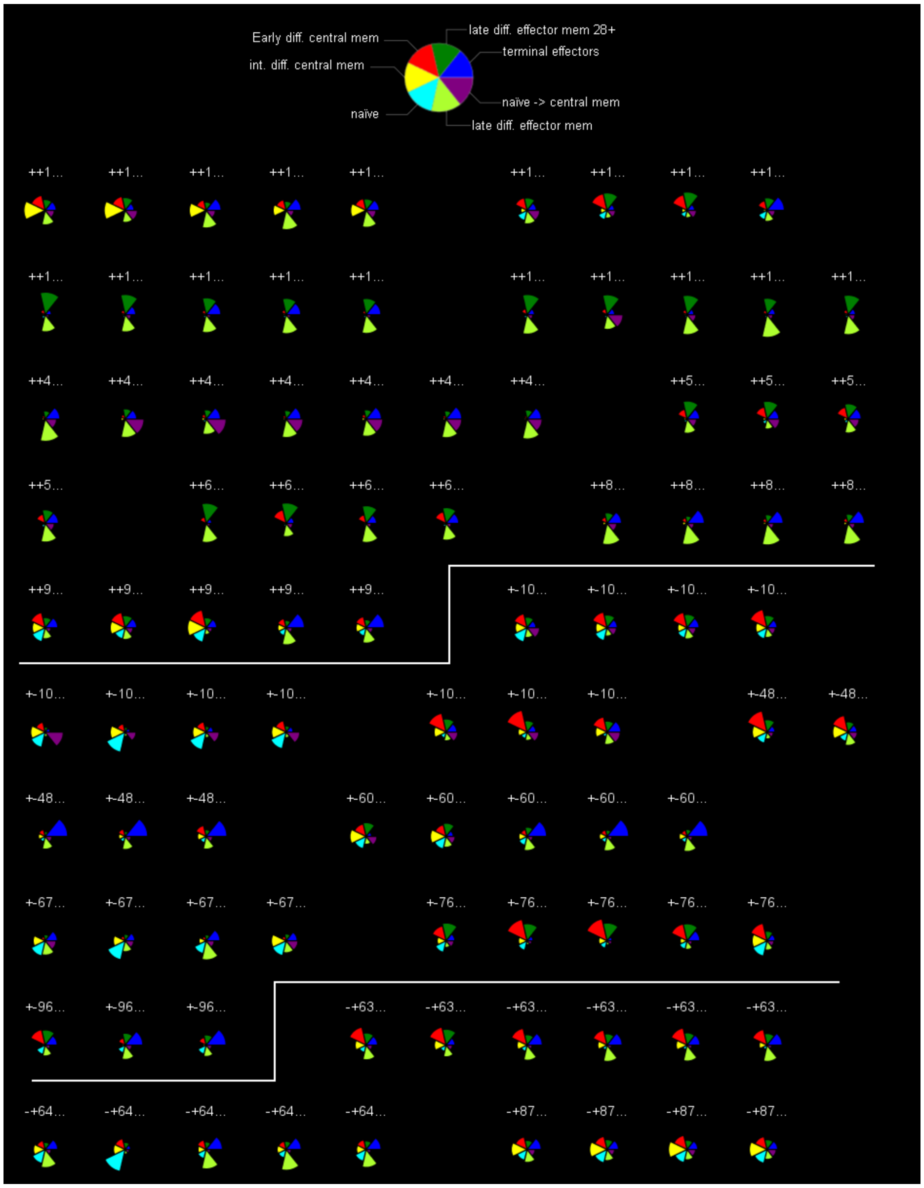

Figure 4.11: Wind Rose Chart for PIP study (+/+, +/- and -/+)

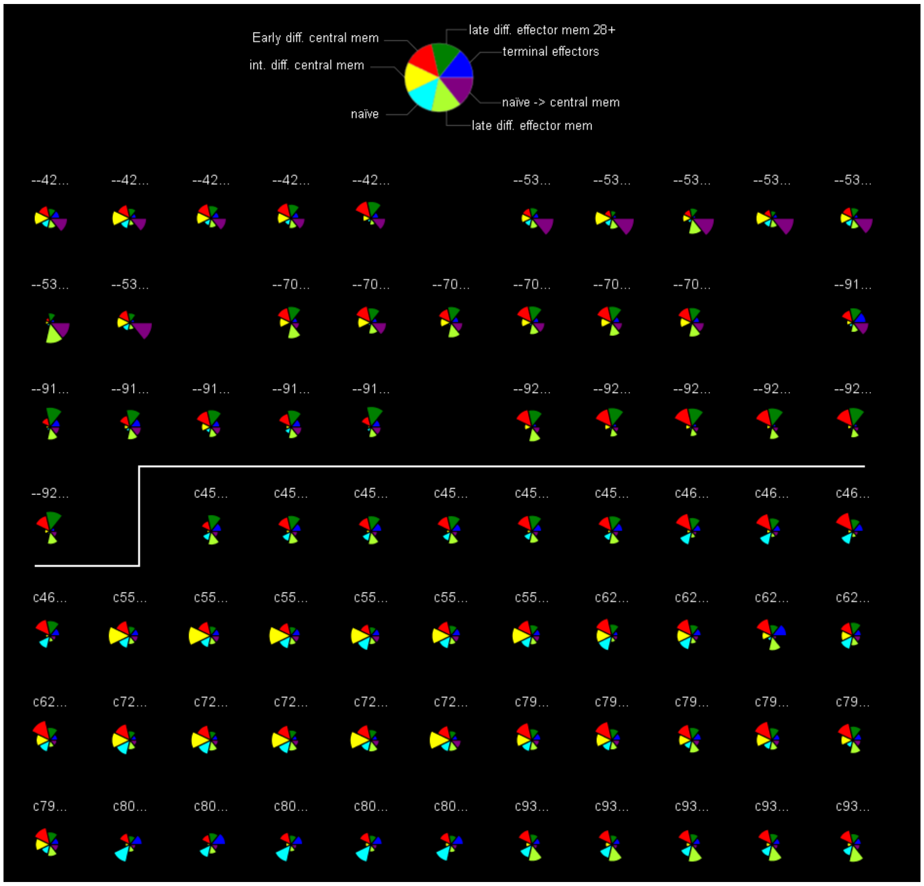

Figure 4.12: Wind Rose Chart for PIP study (-/- and control)

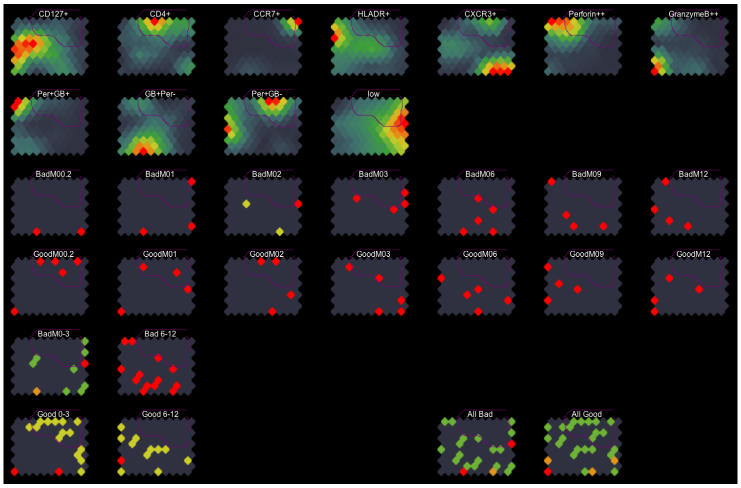

Figure 4.13: SOM of all patient\time summaries from the lung transplant study. Notice that the circled area only contains early data points from patients with a good reaction to the transplant.

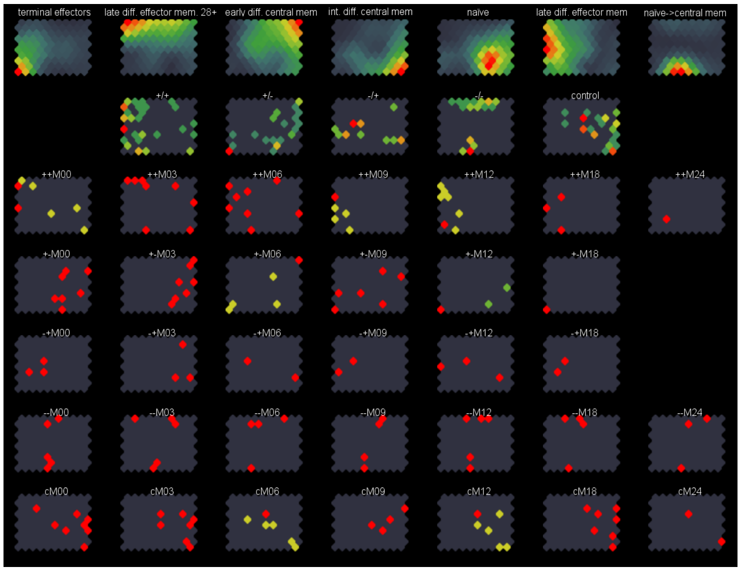

Figure 4.14: SOM of all patient\time summaries from the PIP study divided by donor\patient CMV type

of CD4+ cells and Per+GB- cells, along with being low in CD27+ cells, among others. Only patients who had good reactions (good graft function) to the transplant had blood samples that had this general amount of each cluster. On top of this, it's only those patients in the first 3 months that have blood samples in that area. This then is a strong candidate for a predictor of whether a patient is going to have a good reaction to the transplant. Further testing should be done of new patients to see if those who have initial blood samples of this type will also react well to their transplant.

In the SOM for the PIP data, (Fig. 4.14), the only patients who had blood samples that fit in the top left corner were +/+ patients and -/- patients. These had high numbers of late differentiated effector memory 28+ cells and late differentiated effector memory cells. In addition to this SOM display, I made a separate division of the PIP patients based on whether they got CMV after their transplant (See Fig. 4.15).

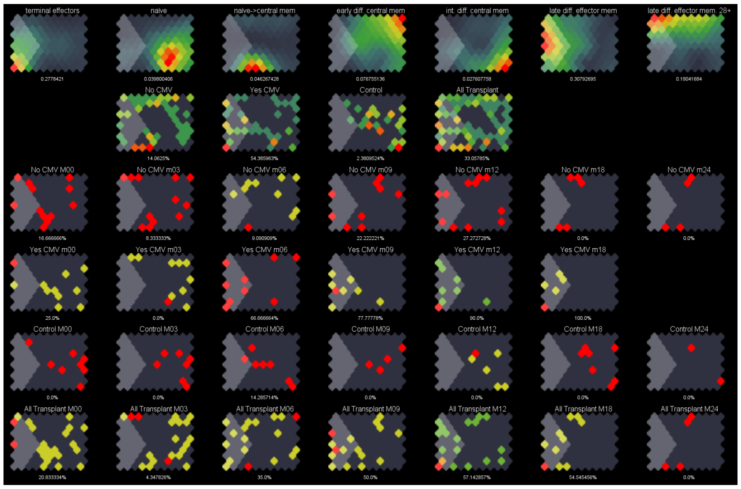

Figure 4.15: SOM of all patient\time summaries from the PIP study divided by whether the patient got CMV. "All Transplants" is just the sum of the "Yes CMV" and "No CMV" patients.

In this SOM for the PIP data (Fig. 4.15), there is a very clear trend in the patients who got CMV. I highlighted an area where the blood samples had a high percentage of late differential effector memory cells or terminal effector cells. Almost no blood samples from controls belonged in this region, and occasional patients who didn't get CMV during the study belonged in the region. Among patients who did get CMV during the study, very few belonged in that region at Months 0 and 3, but in months 9-18, almost all blood samples fit in this region.

Now that we have observed these patterns, we can go back to the Wind-Rose Charts and the original SOM and find the patterns there. For the PIP data, Fig. 4.16 shows all the cells from blood samples at the different times for two transplant groups: The patients who got CMV and those who did not (all PIP patients are on Fig. 4.17).

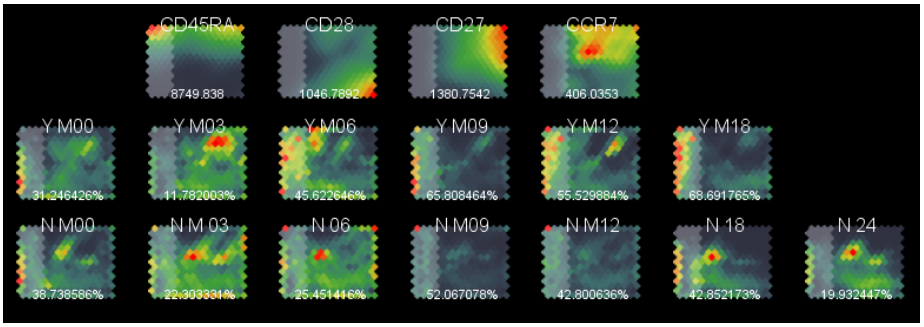

Figure 4.16: SOM of all cells with combined hit histograms of those who got CMV (Y) and those who did not (N). Late differential effector memory cells and terminal effector cells are selected.

Notice that the combined Y patients go from having few cells in the selected area to many, while the N's are relatively constant in that area (the jumps in months 9 and 12 seem to be from patient 90 who had a much larger number of cells than the others did). The individual patients who got CMV can be examined on Fig. 4.18. Almost all these patients show a jump to at least 50% of cells in this region. One exception is patient 104, which stays low for all its

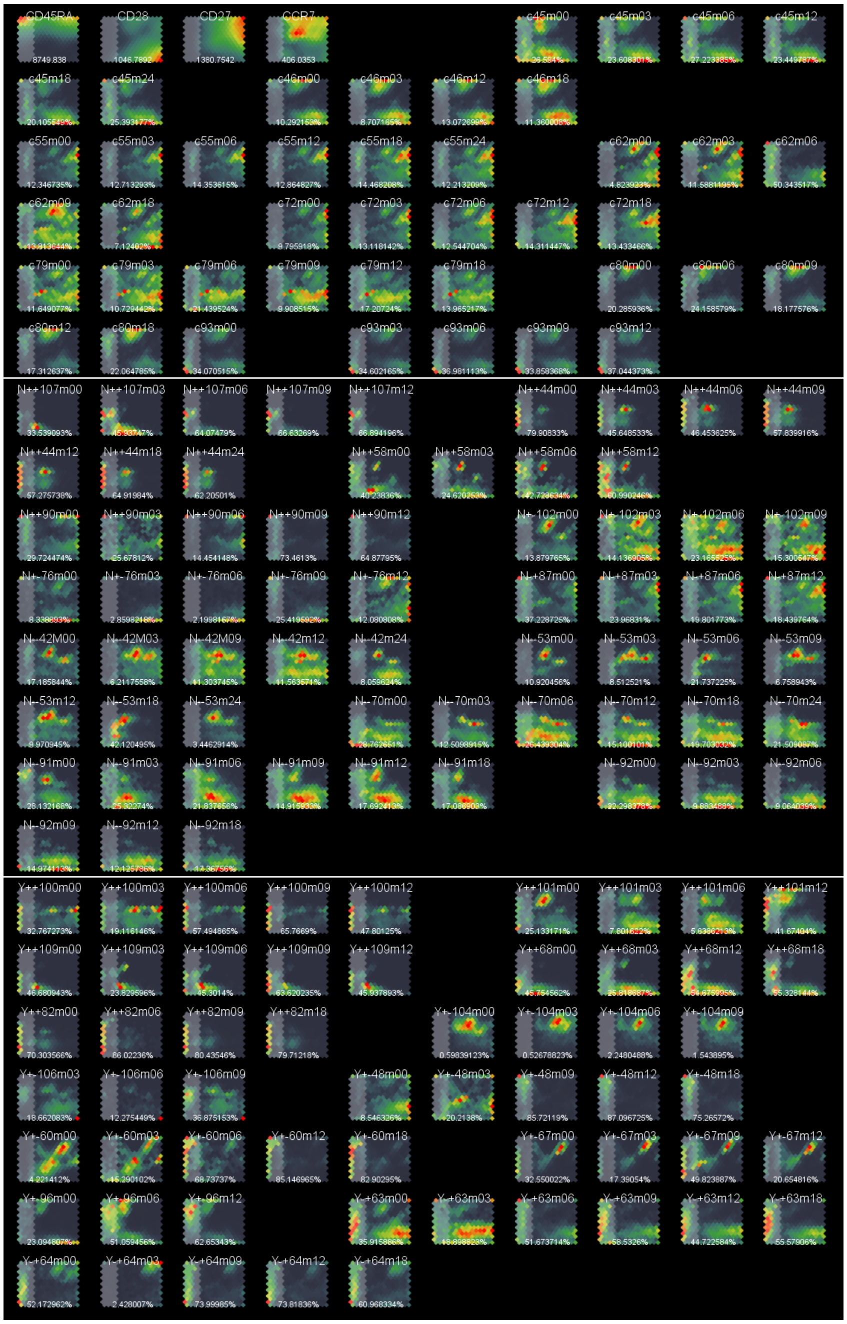

Figure 4.17: SOM of cells with hit histograms for each patient at each time grouped by controls, those who got CMV (Y) and those who did not (N). Late differential effector memory cells and terminal effector cells are selected.

times, but that patient didn't get CMV until month 12, after the last blood sample was taken. In general, these patients seem to show this jump in the month they showed CMV or in the next measurement.

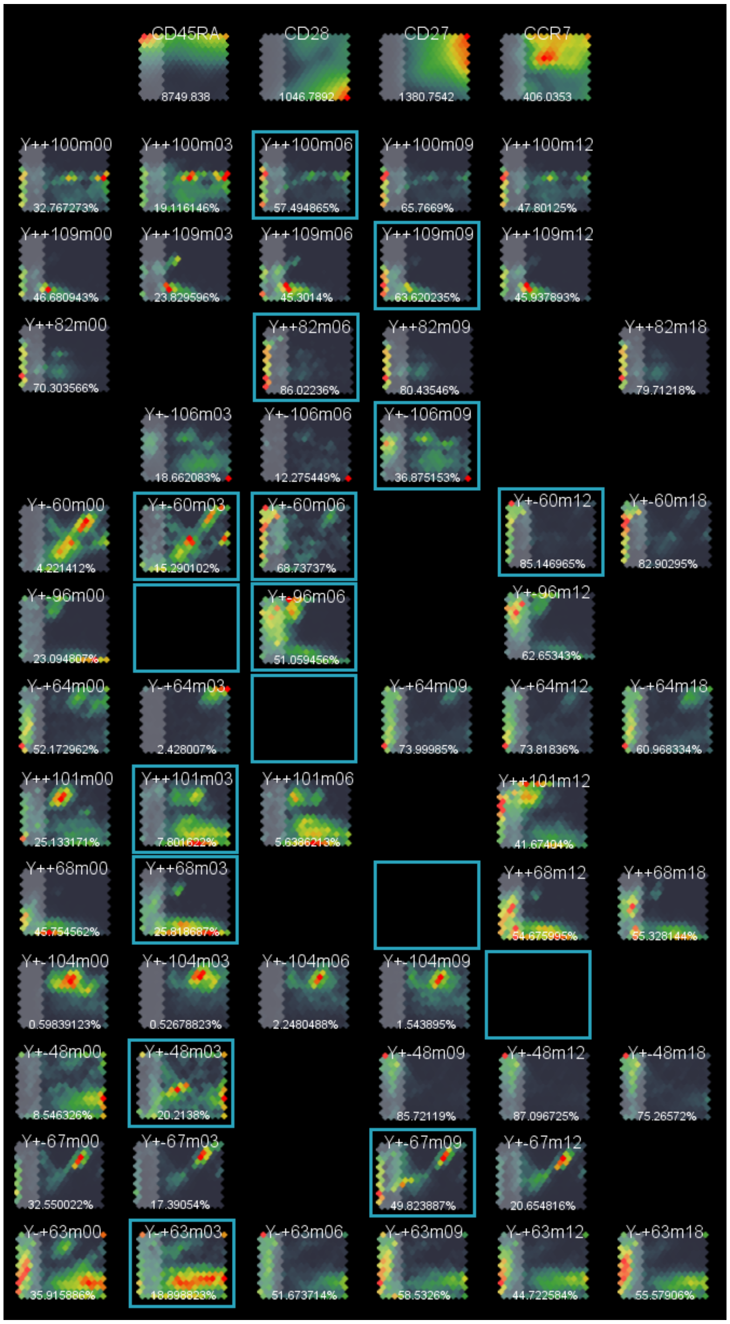

Figure 4.18: SOM of cells with hit histograms for each patient who got CMV at each time. Late differential effector memory cells and terminal effector cell are selected. A blue rectangle is around each time a patient had CMV. The jump in the selected cells generally occurs when they get CMV.

Chapter 5

Discussion

SOMs and WRCs are effective tools for finding patterns in FCM data from different patients at different times. The patterns found in this paper seem to be genuine patterns in the data I was given. In particular the gain in late differential effector memory cells and terminal effector cells in patients who got CMV (see Fig. 4.18) had already been found by the medical researchers [Mehta et al., 2011]. Whether the other patterns found will hold with further data on other patients still needs to be tested.

The time needed to perform the tedious traditional FCM analysis prevents the clinical use of FCM data. With SOMs and WRCs, this analysis can be performed much faster and allow clinical use. Once the relevant cells (in this study CD8 T-lymphocytes) have been exported and aligned for use in IFC Soft, all the clustering and graph production for one study can be done in a single day. This could be done much faster with improvements to IFC Soft (eg. handling categorical divisions internally, saving program sessions, exporting graphs, etc.). If SOMs can be used to pick out the important population of cells for the study, as was done in FlowJo (eg. CD8 T-lymphocytes), then doing this could speed up analysis as well.

Based the visualization and analysis done for this paper, the most effective order to process the data after getting it into IFC Soft is to first use an SOM of all the cells for selecting clusters, then examine an SOM of all the patient/time cluster summaries to find general trends in the data. From there the WRCs and hit histograms of the SOM on all the cells can be examined for further details. These patterns can be used to produce signature hit histograms of different types of patients. These SOMs and hit histogram signatures can be compared with future patients as an aid to diagnosis or confirmation of the signature. The clusters of cells that were selected in IFC Soft were saved out into individual csv files which can be converted into FCS files and loaded back into FlowJo [Simm, 2010] for further investigation. WRCs of the cluster data and SOMs of all the cells are still not easy for finding general patterns, but once patterns are found, they provide more detail in a concise way.

While this is in some ways similar to the methodology used in the cancer study [Sun et al., 2010], my methodology emphasizes visualization and interaction with the data rather than automatic clustering of cells and patients. Both their methodology and mine were based on making an SOM of all the cells of the patients combined, but while they continued by using automated clustering algorithms for the patients on the resulting SOM, I focused on visualization and interaction with the data. I made use of WRCs and a second level of SOMs to allow for rapid visual exploration of trends and patterns in the data. If their automated methodology could be integrated with IFC Soft, then it could be an even more powerful tool in analyzing FCM data.

There is still much that can be done with IFC Soft to improve its use in visualizing and analyzing FCM data. The interface is rudimentary and from what I've seen of others trying to use it, it is not very intuitive. In order for future researchers to make use of this methodology, the software must be made easy to use. Additionally, many new features could be added to further IFC Soft's use with FCM data. Some of the most important features would be the internal handling of categorical data, automatic clustering methods, ability to save SOMs and sessions, and alternate displays of the SOM nodes (in particular, more sophisticated alternatives to hit-histograms).

Though the traditional techniques for analyzing FCM data are used effectively, they are tedious and possibly miss out on multi-dimensional patterns in the data. SOMs and WRCs provide faster, less tedious and more informative means of interacting with the data and provide valuable alternative means of visualization and analysis.

Bibliography

Esa Alhoniemi anb Johan Himberg, Juha Parhankangas, and Juha Vesanto. Som toolbox 2.0. http://www.cis.hut.fi/somtoolbox/, 2005.

Ali Bashashati and Ryan R. Brinkman. A survey of flow cytometry data analysis methods. *Adv. Bioinformatics*, 2009, 2009.

BD Biosciences. Special Order BD LSRFortessa$^{\text{TM}}$Cell Analyzer Technical Specifications. http://www.bdbiosciences.com/documents/SO_LSRFortessaCellAnalyzer_TechSpec.pdf, May 2009.

Tjibbe Donker. FloCK: Flow-cytometry Clustering by K-means. http://theory.bio.uu.nl/tjibbe/flock/, May 2009.

Michael Eisenstein. Cell sorting: Divide and conquer. *Nature*, 441:1179–1185, 2006. URL http://www.nature.com/nature/journal/v441/n7097/full/4411179a.html.

Michael Friendly. Milestones in the history of thematic cartography, statistical graphics, and data visualization: 1800 - 1849. http://euclid.psych.yorku.ca/SCS/Gallery/milestone/sec5.html, August 2009a.

Michael Friendly. Milestones in the history of thematic cartography, statistical graphics, and data visualization: 1850-1899. http://euclid.psych.yorku.ca/SCS/Gallery/milestone/sec6.html, August 2009b.

G Grégori, A Colosimo, and M Denis. Phytoplankton group dynamics in the bay of marseilles during a 2-year survey based on analytical flow cytometry. *Cytometry*, 44(3):247–56, 2001. ISSN 0196-4763. URL `http://www.biomedsearch.com/nih/Phytoplankton-group-dynamics-in-Bay/11429775.html`.

Robert L. Harris. *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, Inc., New York, NY, USA, 1999. ISBN 0195135326.

D.R. Helsel and R. M. Hirsch. Statistical Methods in Water Resources Techniques of Water Resources Investigations, Book 4, chapter A3. `http://pubs.usgs.gov/twri/twri4a3/pdf/chapter16a.pdf`, 2002.

T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001. ISBN 3540679219.

Holden T Maecker, J Philip McCoy Jr, and the FOCIS Human Immunophenotyping Consorium. A model for harmonizing flow cytometry in clinical trials. *Nature Immunology*, 11:975–978, 2010.

MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

A. K. Mehta, M. McCausland, D. J. Lo, J. Joseph, J. Cheeseman, R. Elbein, T. Gourley, P. Mulupuri, J. Miller, A. D. Kirk, R. Ahmed, and C. P. Larsen. Poster abstracts: Abstract# 1226 poster board #-session: P78-iii chronic kidney disease and cmv drive patient specific pre- transplant immune deviations toward terminal differentiation and exhaustion. *American Journal of Transplantation*, 11:388, 2011. ISSN 1600-6143. doi: 10.1111/j.1600-6143.2011.03534.x. URL `http://dx.doi.org/10.1111/j.1600-6143.2011.03534.x`.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

SAS Institute Inc. SAS Product Documentation. `http://support.sas.com/documentation/`, March 2011.

Meredith N. Scarberry. Graphical display methods for exploiting the dimensionality of flow cytometry data. Master's thesis, Emory University, 2009.

Richard H. Scheuermann, Yu Qian, Chungwen Wei, and Inaki Sanz. Immport flock: Automated cell population identification in high dimensional flow cytometry data. *Adv. Bioinformatics*, 182, 2009.

Maciej Simm. The Daily Dongle: Roll Your Own FCS Files Part 2. `http://flowjo.typepad.com/the_daily_dongle/2010/12/roll-your-own-fcs-files-part-2.html`, December 2010.

Jing Sun, Michael D Masterman-Smith, Nicholas A Graham, Jing Jiao, Jack Mottahedeh, Dan R Laks, Minori Ohashi, Jason DeJesus, Ken-ichiro Kamei, Ki-Bum Lee, Hao Wang, Zeta T F Yu, Yi-Tsung Lu, Shuang Hou, Keyu Li, Max Liu, Nangang Zhang, Shutao Wang, Brigitte Angenieux, Eduard Panosyan, Eric R Samuels, Jun Park, Dirk Williams, Vera Konkankit, David Nathanson, R Michael van Dam, Michael E Phelps, Hong Wu, Linda M Liau, Paul S Mischel, Jorge A Lazareff, Harley I Kornblum, William H Yong, Thomas G Graeber, and Hsian-Rong Tseng. A microfluidic platform for systems pathology: multiparameter single-cell signaling measurements of clinical brain tumor specimens. *Cancer Res*, 70 (15):6128–38, 2010. ISSN 1538-7445. URL `http://www.biomedsearch.com/nih/Microfluidic-Platform-Systems-Pathology-Multiparameter/20631065.html`.

TreeStar Inc. FlowJo: Flow Cytometry Analysis Software. `http://www.flowjo.com/`, March 2011.

Juha Vesanto. Som-based data visualization methods. *Intelligent Data Analysis*, 3:111–126, 1999.

Malcolm F. Wilkins, Lynne Boddy, Colin W. Morris, and Richard Jonker. A comparison of some neural and non-neural methods for identification of phytoplankton from flow cytometry data. *Computer Applications in the Biosciences*, 12(1):9–18, 1996.

Wolfram Research, Inc. Mathematica Edition: Version 8.0, 2010.