

## Flow Cytometry Data Analysis: Comparing Large Multivariate Data Sets Using Classification Trees

Joseph Norman

Section on Medical Informatics  
Medical School Office Building X-215  
Stanford University  
Stanford, California 94305  
norman@camis.stanford.edu

### Abstract

This paper describes a method to compare flow cytometry data sets, which typically contain 50,000 six-parameter measurements each. By this method, the data points in two such data sets are divided into subpopulations using a binary classification tree generated from the data. The  $\chi^2$  test is then used to establish the homogeneity of the two data sets based on how their data are distributed across these subpopulations. Preliminary results indicate that this comparison method is sufficiently sensitive to detect differences between flow cytometry data sets that are too subtle for human investigators to notice.

### Introduction

#### Flow Cytometry

Flow cytometry is a powerful laboratory technique for analyzing biological cells [Parks et al., 1989]. This technique allows for the simultaneous measurement of several characteristics on a cell-by-cell basis. For example, an investigator can use flow cytometry to measure the levels of expression of three different surface proteins on each of 50,000 white blood cells. Fluorescence-activated cell sorting (FACS), an extension of flow cytometry, physically separates subpopulations of viable cells from one another. The acronym FACS is commonly used to refer to flow cytometry in general, even if cells are not actually sorted.

Flow cytometry uses fluorescent reagents to label molecules expressed on cell surfaces or contained within cells. Each reagent is a fluorescent dye joined to an antibody specific for a particular cellular molecule. For a typical experiment, a population of cells is stained with a set of three or four carefully chosen reagents. The stained cells are suspended in a liquid medium and sent, one cell at a time, through the sensing region of a FACS instrument. There, each cell is exposed to several beams of laser light of wavelengths appropriate to excite the fluorescent reagents. Photosensitive detectors then measure the light emitted by each cell. Each detector is tuned for a specific range of

wavelengths of emitted light, corresponding to the emission spectrum of one of the reagents. Thus, the intensity of the signal from each detector indicates the amount of the corresponding reagent bound to the cell analyzed. There are also detectors that measure light scattered by each cell. The signals from the detectors are digitized and recorded.

In this way, each cell is characterized by several measurements. Most instruments in the Stanford FACS facility have two detectors to measure scattered light and four to measure fluorescent emissions, producing six measurements per cell. From 10,000 to 50,000 cells are analyzed from each population studied, resulting in large multivariate data sets. Analyzing these data is a challenging task, typically done by expert investigators who use a combination of qualitative and quantitative methods.

#### Data Analysis

Data set comparison is an important task in the analysis of flow cytometric data. Most interpretations of FACS results rest on an assertion that one population of cells differs from another population of cells. Investigators infer such differences between cell populations from differences between data sets that describe samples of those populations. However, investigators do not use a quantitative test of similarity to compare multiparameter FACS data sets.

In many cases, the fact that two FACS data sets represent different cell populations is evident from visual inspection of a few well-chosen, two-parameter plots of the data. But visual inspection of data plots is not always satisfactory. Even expert investigators sometimes disagree about whether two data sets are different, based on a set of two-dimensional projections of the data. Moreover, there are many subtle differences between data sets that are not visually apparent, such as the difference between two samples of the same cell population run consecutively on the same FACS instrument or two samples of the same type of cell prepared and analyzed by different investigators on different days.

To address this analysis task, we have developed a method to compare FACS data sets that detects both large and small differences. One of our goals was to use a statistically sound method that would allow us to infer some level of significance from our measurement of homogeneity between data sets; we have made some progress in that direction.

## Related Work

Researchers at other institutions have implemented techniques to compare one-parameter flow cytometry data distributions and to classify multiparameter flow cytometric data. The Kolmogorov-Smirnov test, a nonparametric test for comparing two univariate data distributions, was first applied to flow cytometric data by Young [1977]. Cox and colleagues discussed the use of this test, as well as parametric tests, to compare distributions of single-parameter FACS data [1988]. Simple forms of cluster analysis have been applied to flow cytometric data: Demers and coworkers used the technique to identify different species of aquatic microorganisms [1992]. Neural networks based on multiparameter data have also been used to classify aquatic microorganisms [Frankel et al., 1989]. Neural network analysis has been used to identify features in single-parameter flow cytometry data that correlate with risk of breast cancer relapse [Ravdin et al., 1993].

Our work developing computer-based support for flow cytometry has addressed experiment planning, instrument modeling, and data analysis. The PENGUIN system was developed in our laboratory to facilitate the use and sharing of declarative domain knowledge stored in relational databases [Barsalou et al., 1991]. One of our colleagues completed preliminary work on a distributed, object-oriented system to perform a variety of FACS tasks, including instrument control, protocol design, data analysis, and data visualization [Matsushima, 1993].

## Method

Our strategy for comparing flow cytometric data sets was to reduce the multivariate data to categorical data. We defined for each pair of data sets a set of subpopulations based on the measured parameters, and converted the long lists of multiparameter measurements to short lists of frequency counts for those subpopulations. The data sets could then be tested for homogeneity using a simple  $\chi^2$  test.

This approach is suggested by the nature of flow cytometric data. Existing parametric tests for the homogeneity of two multivariate data samples assume that the samples have multinormal distributions [Piterbarg and Tyurin 1993]; FACS data is decidedly non-normal (for example, the marginal distributions are often multimodal), which makes it necessary to use a different approach.

Consider a data set  $A = \{A_1, A_2, \dots, A_m\}$ , where each  $A_i$  is a multiparameter measurement

$\{A_{i1}, A_{i2}, \dots, A_{id}\}$ , and a similar data set  $B = \{B_1, B_2, \dots, B_n\}$ . These data sets have  $m$  and  $n$  data points respectively (typically 50,000 data points each) and contain  $d$ -dimensional data (typically 6). Based on the union of  $A$  and  $B$ ,  $\{A_1, \dots, A_m, B_1, \dots, B_n\}$ , we can divide the data set space into  $q$  regions, as described below.

Let  $X_k$  be the number of data points  $A_i$  in the  $k$ th region, and  $Y_k$  the number of data points  $B_j$  in the  $k$ th region. Now we can express each data set as a list of frequency counts for the  $q$  regions. Thus  $A = \{X_1, \dots, X_q\}$ , and  $B = \{Y_1, \dots, Y_q\}$ . Under the null hypothesis that  $A$  and  $B$  are samples drawn from the same population, we expect

$$E_{Ak} = \frac{m}{m+n} (X_k + Y_k)$$

data points from data set  $A$  in the  $k$ th region. Similarly,

$$E_{Bk} = \frac{n}{m+n} (X_k + Y_k)$$

for data set  $B$ . The test statistic is then given by

$$x^2 = \sum_k \left[ \frac{(X_k - E_{Ak})^2}{E_{Ak}} + \frac{(Y_k - E_{Bk})^2}{E_{Bk}} \right]$$

which approaches  $\chi^2$  for large  $q$ .

We reject the null hypothesis that  $A$  and  $B$  are homogeneous if  $x^2$  exceeds the tabulated value of  $\chi^2$  for  $q-1$  degrees of freedom and the chosen confidence level  $\alpha$ . [Ott, 1993]

The data set space is divided into regions by means of a binary classification tree that is drawn by splitting the space recursively along the parameter axes at their median data points. The parameter with the longest 5th to 95th percentile range of data is chosen for the next split of each region, and the region is further subdivided until a specified minimum number of data points remain in each region.

A data-driven classification technique is used because the data set space is very sparsely populated. There are many more points in six-dimensional space (considered at a nine-bit resolution, as FACS fluorescence signals are digitized) than there are points in a data set. Subpopulations defined by dividing the space into a regular grid would be too coarse to resolve small differences between data sets.

The comparison method was implemented in C++ on Sun SPARCstations under UNIX.

## Evaluation

In order to evaluate the comparison method we collected and compared several data sets. Different sample preparations and FACS instrument conditions were chosen to demonstrate several varieties of difference:

- Variation within a single data set

- Variation between data sets collected from identically prepared and analyzed cells
- Variation between data sets collected from differently prepared and analyzed cells

The differences in cell preparation and analysis included changes in factors suspected to influence data slightly, but known not to introduce gross changes in the visual appearance of data sets. These factors included:

- Cell sample condition
- Instrument calibration
- Laser realignment

### Sample Preparation

Four groups of cells (A–D) were prepared as described below. White blood cells were harvested from the spleens of two genetically identical BALB/c mice, washed and resuspended. The cells for groups A through D were kept together in a single test tube.

The four groups of cells were run sequentially through the FACS machine. Before any cells were analyzed, the FACS instrument was calibrated. Four samples were then drawn from the test tube for the four group A data sets. The tube was removed from the FACS machine and vigorously shaken, after which three more data sets were collected (group B). The instrument was recalibrated, and three group C data sets were collected. One of the lasers was intentionally misaligned and realigned; three more samples were analyzed for group D. The sample treatments are summarized in Table 1.

Group	Treatment
A	Unstained
B	Unstained, shaken
C	Unstained, after recalibration
D	Unstained, after laser realignment

**Table 1.** Cell sample treatments. Three 50,000-point data sets were collected from each treatment group.

### Data Set Comparison

Two parameters were measured for each data set, forward scatter and obtuse scatter. These light scatter measurements correlate with cell size and shape; investigators often use these parameters to monitor the quality of data collection. Three data sets were collected for each cell group; each data set contained 50,000 data points (representing 50,000 cells).

Data sets were compared pairwise both within each treatment group and between the treatment groups. For this set of comparisons, the minimum number of data points per classified region was set at 1000, resulting in 64 subpopulations for each comparison.

## Results

Preliminary results show that the classification method is sufficiently sensitive to distinguish between data sets from the same treatment group and data sets from different treatment groups. The comparison results are summarized in Table 2. Within-group comparisons for groups A, B, C, and D gave  $\chi^2$  values between 67 and 142, whereas between-group comparisons gave values ranging from 499 to 3144.

Pairwise comparison of three 50,000-point data sets extracted from a single 300,000-point set gave an average comparison value of 60.

For the  $\chi^2$  distribution with 63 degrees of freedom, values above 88 are significant at the 5% level.

	A	B	C	D
A	121	3144	2525	499
B		142	615	2679
C			67	1568
D				91

**Table 2.** Comparison results. The tabulated values are averaged comparison values for pairs of data sets (three values for each within-group comparison, nine for each between-group comparison). These numbers are roughly comparable to  $\chi^2$  values for 63 degrees of freedom.

## Discussion

The comparison method performs well in distinguishing different levels of difference from one another. The order of magnitude of the comparison value indicates whether two compared data sets belong to the same treatment group or not.

However, the statistical interpretation of the comparison value as a proper  $\chi^2$  value is questionable. Data sets separated only by seconds of collection time, such as the different data sets in group A, gave comparison values in excess of the cutoff value for statistical significance at the 5% level. Our comparison measure may still have statistical validity, perhaps with a distribution different from that of  $\chi^2$ . Further investigation should elucidate this actual distribution.

This comparison method takes advantage of the multivariate nature of the data. Using a classification tree allows all data parameters to be taken into account, unlike some other methods which only compare univariate distributions.

There are many possible ways to divide data sets to define subpopulations; we simply chose one that seemed reasonable. But perhaps rather than simply splitting the parameter with the widest data distribution at its median value, the method could search for divisions that maximize some measure of the amount

of order in the data (as in the ID3 classification algorithm [Ginsberg, 1994]).

Data visualization played an important role in the course of this work. The ability to see the regions drawn by the classification algorithm was essential in evaluating its appropriateness. The advantages of static data displays such as printouts could be expanded through the use of dynamic, interactive displays such as discussed by Becker and colleagues [1987].

### Acknowledgments

The author thanks Lawrence Fagan and Leonore Herzenberg for their guidance and support, Michael Walker and Tze Lai for their statistical advice, and Toshiyuki Matsushima for his technical assistance. This work was supported by National Library of Medicine grants 2-R01-LM04336-04-A1 and LM-050305, and the Medical Scientist Training Program.

### References

- Barsalou, T., W. Sujansky, L. A. Herzenberg, and G. Wiederhold. 1991. Management Of Complex Immunogenetics Information Using an Enhanced Relational Model. *Computers and Biomedical Research* 24:476–498.
- Becker, R. A., W. S. Cleveland, and A. R. Wilks. 1987. Dynamic Graphics for Data Analysis. *Statistical Science* 2:355–395.
- Cox, C., J. E. Reeder, R. D. Robinson, S. B. Suppes, and L. L. Wheelless. 1988. Comparison of Frequency Distributions in Flow Cytometry. *Cytometry* 9:291–298.
- Demers, S., J. Kim, P. Legendre, and L. Legendre. 1992. Analyzing Multivariate Flow Cytometric Data in Aquatic Sciences. *Cytometry* 13:291–298.
- Frankel, D. S., R. J. Olson, S. L. Frankel, and S. W. Chisholm. 1989. Use of A Neural Net Computer System for Analysis of Flow Cytometric Data of Phytoplankton Populations. *Cytometry* 10:540–550.
- Ginsberg, Matthew. 1993. *Essentials of Artificial Intelligence*. San Mateo, CA: Morgan Kaufman.
- Matsushima, T. 1993. Constructing a distributed object-oriented system with logical constraints for fluorescence-activated cell sorting. In Proceedings of the First International Conference on Intelligent Systems for Molecular Biology, 266–274. Menlo Park, CA: AAAI Press.
- Ott, R. Lyman. 1993. *An Introduction to Statistical Methods and Data Analysis*. Belmont, CA: Wadsworth.
- Parks, D. R., L. A. Herzenberg, and L. A. Herzenberg. 1989. Flow Cytometry and Fluorescence-Activated Cell Sorting. In Paul, W. E. (ed), *Fundamental Immunology*, 2nd ed. New York: Raven.
- Piterbarg, V. I., and Y. N. Tyurin. 1993. Testing for Homogeneity of Two Multivariate Samples: a Gaussian Field on a Sphere. *Mathematical Methods of Statistics* 2:147–164.
- Ravdin, P. M., G. M. Clark, J. J. Hough, M. A. Owens, and W. L. McGuire. 1993. Neural Network Analysis of DNA Flow Cytometry Histograms. *Cytometry* 14:74–80.
- Young, I. T. 1977. Proof without Prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. *J Histochem Cytochem* 25: 935–941.