Matching of Flow-Cytometry Histograms Using Information Theory in Feature Space

Qing Zeng, Ph.D.[1], Matthew Wand, Ph.D.[4], Alan J. Young, Ph.D. [3,5], James Rawn, M.D.[3], Edgar L. Milford, M.D.[4], Steven J. Mentzer, M.D.[3], Robert A. Greenes, M.D., Ph.D.[1]

[1]Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School

[2]Department of Biostatistics, Harvard School of Public Health

[3]Department of Surgery, Brigham and Women's Hospital, Harvard Medical School

[4]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

[5]Department of Biology & Microbiology, South Dakota State University

Abstract

Flow cytometry is a widely available technique for analyzing cell-surface protein expression. Data obtained from flow cytometry is frequently used to produce fluorescence intensity histograms. Comparison of histograms can be useful in the identification of unknown molecules and in the analysis of protein expression. In this study, we examined the combination of a new smoothing technique called SiZer with information theory to measure the difference between cytometry histograms. SiZer provides cross-bandwidth smoothing and allowed analysis in feature space. The new methods were tested on a panel of monoclonal antibodies raised against proteins expressed on peripheral blood lymphocytes and compared with previous methods. The findings suggest that comparing information content of histograms in feature space is effective and efficient for identifying antibodies with similar cell-surface binding patterns.

Introduction

Flow cytometry is a technique for quantitatively measuring the expression of individual molecules on cells. This technique, combined with monoclonal antibody technology, is a powerful tool with both research and clinical applications (1).

In flow cytometry, fluorescent labeled antibodies are used as probes, and fluorescence detectors are used to analyze large numbers of cells sequentially (2). Since antibodies will generally bind to their respective target molecule or "antigen" in a one-to-one ratio, the number of antibodies bound to a cell, and hence the number of fluorescent molecules present, will generally be proportional to the level of expression of that protein on the cell. Each antibody-bound cell emits a pulse of fluorescence which can be specifically detected by a cytometer. Fluorescence intensity histograms, which depict the distribution of cell surface antigen densities within the population of cells under study, can be thought of as a molecular "fingerprint" of protein expression.

With the recent development of high-throughput cytometers that employ digital signal process techniques, it has become feasible to obtain large amounts of accurate cytometry data that can be used for proteomics research. New techniques for the analysis of flow cytometry data, specifically comparison of fluorescence intensity histograms, could enable the identification of unknown molecules based on their patterns of cell-surface expression.

Techniques for comparison of fluorescence intensity histograms are not new. Most existing methods for analyzing cytometry histograms are designed to either test the hypothesis that two histograms are the same, or calculate the number of "positive" cells (cells that bind labeled antibodies) in a histogram (3-5).

Hypothesis testing methods such as the Kolmogorov-Smirnov (KS) method are capable of answering whether two histograms have a statistically significant difference (6). The statistically significant difference, however, is not necessarily biologically significant (7). These methods also do not offer a measurement of similarity beyond identity.

Other methods have been developed for counting the number of positive cells in a histogram (4, 8). While these methods are useful for certain clinical applications, the number of positive cells is not to establish identity. Two histograms with the same of number of positive cells may have very different shapes and reflect different binding patterns.

We had previously experimented by comparing cytometry histograms using Shannon's information theory, which provides a measurement of similarity and showed promising results in our preliminary study (9). Our original approach, however, had two limitations: (1) empirical parameter settings were employed to smooth noisy data; and (2) comparisons were performed on the entire histograms instead of selected or derived features, although certain features may be biologically more relevant than others.

To address these drawbacks, we combined a new curve characterization technique, SiZer (10), with an

information theory approach. SiZer provides cross-bandwidth smoothing, which transforms a curve into a family of curves with varying degrees of smoothness. Peaks are the most important features of a histogram and SiZer provides information on peaks in the form of a map which marks different parts of a curve as going up, down or flat. We applied information theory distance measurement on the SiZer-smoothed families of curves and on the SiZer maps.

To evaluate the effectiveness of the different approaches for identifying similar cell surface binding patterns, they were tested on a sample data set of peripherial blood lymphocytes (PBLs) labeled with a panel of diverse monoclonal antibodies. We calculated the information theory distance between the histograms in four different ways, using: (1) SiZer smoothed families, (2) SiZer maps, (3) histograms smoothed using a polynomial filter. and (4) unsmoothed histograms. Our evaluation showed that the distinguishing power of information theory distance is most enhanced when used with SiZer maps, which reflected the information content of histograms in a feature space.

Background

Information Theory

In recent years, Shannon's information theory (11) has been applied to a wide variety of problems, including image registration and gene sequence analysis (12, 13). In a previous paper, we reported its use for flow cytometry histogram comparison (9).

Given a discrete information source, the average information content or entropy reflects the unpredictability of the source. For a stream of symbols, the more random they are, the higher their entropy is. Both mutual information and distance can be used to measure the similarity between two sources.

We chose to use distance as the similarity measurement instead of conditional entropy and mutual information, because the distance is symmetrical between two sources, while conditional entropy and mutual information are not. To distinguish this specific distance measurement here from the general concept of "distance", we refer to it as the IT distance, calculated as in our previous work (9).

SiZer

SiZer ("Significant Zero Crossings of derivatives") is a new tool for distinguishing significant features (e.g., peaks) of a curve from noise. As the name suggests, SiZer uses estimates of the first derivative of the histogram following kernel smoothing (10). The significance of a feature in a histogram is also a function of the amount of smoothing applied to the histogram. To address this issue, SiZer performs cross-bandwidth smoothing, or multiple degrees (from minimum to maximum) of smoothing. After processing a histogram, it provides two types of output: (1) a family of smoothed histograms using different bandwidths, and (2) a map indicating which part of a histogram is increasing, decreasing, or flat based on a given smoothing bandwidth. Figure 1 shows a flow cytometry histogram. A family of 11 smoothed histograms for this histogram is displayed in Figure 2, and the corresponding SiZer map is shown in Figure 3.

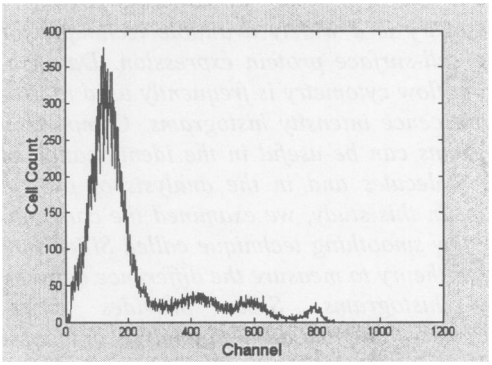

Figure 1. Unsmoothed flow cytometry histogram of antibody 18-106 with peripheral blood leucocytes. The x-axis represents the different levels (channels) of fluorescence intensity and y-axis represents the number of cell in a channel.

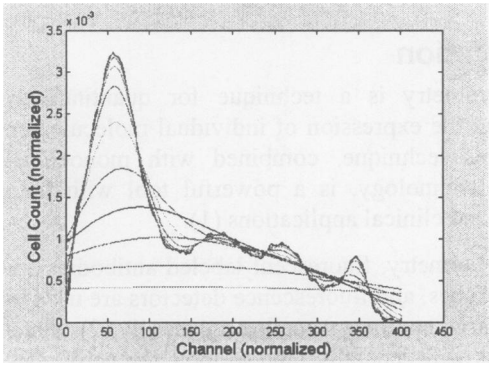

Figure 2. A family of 11 smoothed histograms of antibody 18-106 with PBLs, produced by SiZer. Each histogram corresponds to a particular smoothing bandwidth. The x-axis represents fluorescence intensity channels and y-axis represents the number of cell in the channels (both normalized).

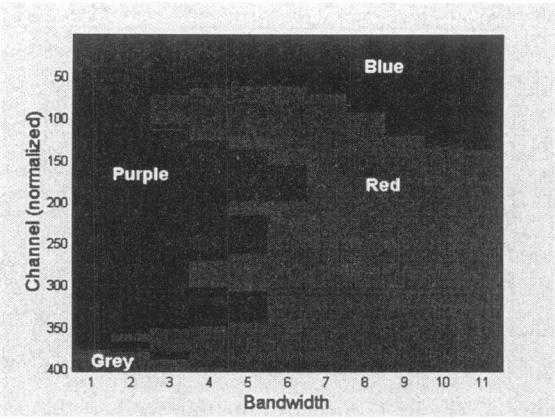

Figure 3. SiZer map of antibody 18-106 with PBLs. The x-axis represents the different smoothing levels and the y-axis represents the fluorescence intensity channels. The colors indicate that a histogram is: increasing (blue), decreasing (red), neither (purple); or that there is not enough sample to decide (grey).

Materials and Methods

Experimental animals

Randomly bred ewes were obtained and housed in accordance with Harvard guidelines on the care and use of experimental animals.

Test cell population

Peripheral blood leucocytes (PBLs) were collected from the jugular veins of the ewes by venipuncture under sterile conditions. Red blood cells were lysed using ammonium chloride and leucocytes harvested by centrifugation (14). All cell populations were washed 3 times in phosphate-buffered saline (PBS), and reacted with monoclonal antibodies as previously described (15).

Test antibodies

All test antibodies were from murine hybridomas produced over the last 15 years, directed against proteins normally expressed by subpopulations of sheep leucocytes. Antibodies were represented by the cell clone name of the murine hybridoma, and included *86D, 18-106, ST-1, H1-68, A1-107, 17-63, 2-87,2-128, 14-24, 14-109, 7C-2/38-65, ST-8, 6-87, 6-99, 2-128-1 (supernatant produced from a subclone of 2-128), FW4, and ERD2/114*. Three replicates were generated for each antibody. The molecular "target" of each of these antibodies is known and shown in Table 1. Antibodies 86D and 18-106 recognize the unique proteins TcR-γδ and CD3γ, which are individual components of the same multi-protein complex. Reactivity was detected by indirect immunofluorescence, using FITC-conjugated goat-anti-mouse Ig diluted 1:10 with PBS. For each histogram, a total of 50,000 cells was analyzed on a Coulter XL flow cytometer, and the fluorescence intensity was presented on log-scale histograms containing a total of 1,024 channels.

Table 1. Cell types, antibodies, and their respective molecular "targets" in the sample set.

Cell Type	Antibody	Molecular "Target"
PBL	86D	TCR-γδ
PBL	18-106	CD3γ
PBL	ST-1	CD5
PBL	H1-68	CD5
PBL	A1-107	CD5
PBL	2-128	CD21
PBL	14-24	CD21
PBL	14-109	CD21
PBL	7C-2/38-65	CD8
PBL	ST-8	CD8
PBL	6-87	CD8
PBL	6-99	CD8
PBL	2-128-1	CD21
PBL	FW4	CD29
PBL	ERD2/114	CD29

IT Distance Calculation

We calculated the IT distance between histograms in four different ways:

1. Using SiZer smoothed families

Each histogram was processed by SiZer into a family of 11 histograms with varying degrees of smoothness. When calculating the information distance, we viewed each smoothed histogram as a string of numbers and concatenated the 11 smoothed histograms into one string of numbers. The IT distance was then calculated between the two concatenated smoothed strings representing the two histograms.

2. Using SiZer maps

A SiZer map was generated for each histogram, which indicates if a point on the histogram is statistically significantly increasing, decreasing, neither, or that the direction cannot be determined because of the sample size. This information was originally

provided as a 2-dimensional array and coded in 4 numbers. For IT distance calculation, we represented the map as a string of numbers.

3. Using histograms smoothed with a Savitzky-Golay filter

For comparison purpose, we smoothed the histograms using a polynomial filter: the Savitzky-Golay (SG) filter as in our previous work (9). A polynomial order of 3 and a frame size of 41 were used, and IT distance was calculated for the smoothed histograms.

4. Using unsmoothed histograms

The information distance between unsmoothed histograms was also calculated as a control to evaluate the various smoothing methods.

Evaluation

In the PBL data set, there are 45 different histograms and 936 possible pairs of different histograms. Among them, 144 pairs recognize the same cell surface complexes and 792 pairs do not. The IT distances were calculated for all 936 pairs of histograms using the four methods described above. We constructed receiver operator characteristic (ROC) curves for the four methods and compared areas under the ROC (AROC) to see which methods are more effective in detecting the biological similarity underlying the histograms

Results

The ROC curves resulting from the four different methods differ but not dramatically (Figure 4). The AROC was the largest using the SiZer map and the smallest using the SG filter. The difference between using the SiZer map and using unsmoothed histogram was statistically significant ($p<0.01$), while the differences among the other three methods were not statistically significant ($p>0.05$).

Table 2. The AROCs and their respective 95% confidence intervals of the four ways we calculated information distances between histograms: using SiZer family, SiZer map, SG filter, and unsmoothed histogram.

Curve	Area	95% CI of Area
SiZer family	0.79	0.75 to 0.84
SiZer map	0.85	0.82 to 0.89
SG filter	0.77	0.72 to 0.81
Unsmoothed	0.78	0.73 to 0.82

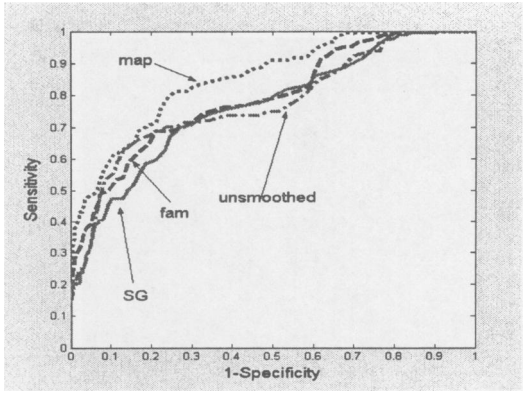

Figure 4. The ROC curves of the four methods by which we calculated information distances between histograms: using SiZer family, SiZer map, SG filter, and unsmoothed histogram.

Discussion

In this study, we experimented with combining a smoothing technique, SiZer, with information theory to analyze flow cytometry histograms for the purpose of identifying antibodies that recognize the same molecular "targets". Our results showed that when using SiZer maps as surrogates for the unsmoothed histograms to calculate information distance, we could better distinguish the histograms of antibodies with similar binding patterns from other histograms ($p<0.01$).

Although the AROC differences between using SiZer maps and other methods were less than 10%, there is significant gain in reducing computational complexity when using SiZer maps. The complexity of calculating information distance between two information sources or streams of symbols is of the order of the square of the size of the number of unique symbols in the sources ($O(n^2)$, *n: the number of unique symbols in the sources.*) In our calculation, we viewed cytometry histograms as strings of numerical symbols, which are the cell counts within florescence intensity channels. In a typical unsmoothed or smoothed histogram, there can be dozens to hundreds of different numbers/symbols. In a SiZer map, however, there are only 4 symbols indicating whether the curve at a particular point is rising, declining, flat or of undetermined direction (due to small sample size). When analyzing large data sets, this reduction in n may result in a major reduction in computing time.

Comparing with the method we previously reported, the use of SiZer required much less setting of smoothing parameters. It also allowed more objective analysis of the histogram data. Some cytometry histograms are "noisier" than others, as a result of which, when using a polynomial filter, we may need

to adjust certain parameters to achieve a desirable smooth effect. Since SiZer performs cross-bandwidth smoothing, no such adjusting is necessary.

The SiZer map focused on a particular feature of a histogram: the first-derivative characteristic. Biologists empirically know that this is the most important and biologically relevant feature. A histogram does have other features, yet not all features are biologically relevant. Analyzing data in the feature space allows us to separate them and to focus on the relevant ones.

In our study, only one data set was used, so the generalizability of our results has yet to be validated. This data set is relatively "clean", compared with some other histograms we have seen. With a data set that has more noise, the results may differ.

We would like to point out that histogram comparison cannot definitively determine the identity of target molecules. Related molecules may have different antigen density distributions within the cell population being studied. Similarly, it is also possible for some unrelated molecules to have density distributions (histogram morphology) for some cell populaitons. Comparisons utilizing multiple cell types, however, are likely to improve the resolution of this method.

We plan to explore more features of flow cytometry histograms in the future, since they might provide additional biologically relevant information. We are also in the process of obtaining larger and more diverse data sets for validation and future studies.

Conclusion

We studied a new approach that combines a smoothing technique, SiZer, with information theory to analyze flow cytometry histograms for the purpose of identifying antibodies that recognize the same molecular "targets". The results showed that this approach could improve performance and reduce computational complexity over the use of information theory with a polynomial filter.

References

1. Boddy L, Wilkins MF, Morris CW. Pattern recognition in flow cytometry. Cytometry 2001;44(3):195-209.

2. Givan AL. Principles of flow cytometry: an overview. Methods Cell Biol 2001;63:19-50.

3. Watson JV. A brief history of numbers and statistics with cytometric applications. Cytometry 2001;46(1):1-22.

4. Watson JV. Proof without prejudice revisited: immunofluorescence histogram analysis using cumulative frequency subtraction plus ratio analysis of means. Cytometry 2001;43(1):55-68.

5. Parikh HH, Li WC, Ramanathan M. Evaluation of an alternative to the Kolmogorov-Smirnov test for flow cytometric histogram comparisons. J Immunol Methods 1999;229(1-2):97-105.

6. Bagwell CB, Hudson JL, Irvin GL. Nonparametric flow cytometry analysis. J Histochem Cytochem 1979;27(1):293-6.

7. Lampariello F. On the use of the Kolmogorov-Smirnov statistical test for immunofluorescence histogram comparison. Cytometry 2000;39(3):179-88.

8. Lampariello F, Aiello A. Complete mathematical modeling method for the analysis of immunofluorescence distributions composed of negative and weakly positive cells. Cytometry 1998;32(3):241-54.

9. Zeng Q, Young AJ, Boxwala AA, Rawn J, Long W, Wand M, et al. Molecular Identification Using Flow Cytometry Histograms and Information Theory. Proc AMIA Symp 2001: 776-780.

10. Chaudhuri P, Marron JS. SiZer for exploration of structures in curves. Journal of the American Statistical Association 1999;94:807–823.

11. Shannon CE. A Mathematical Theory of Communication. The Bell System Technical Journal 1948;27(July,October):379-423, 623-656.

12. Schneider TD. Information content of individual genetic sequences. J Theor Biol 1997;189(4):427-41.

13. Wells WM, Viola P, Atsumi H, Nakajima S, Kikinis R. Multi-modal volume registration by maximization of mutual information. Med Image Anal 1996;1(1):35-51.

14. Young AJ, Hein, W.R., Hay, J.B. Cannulation of lymphatic vessels and its use in the study of lymphocyte traffic. In: Lefkovits I, editor. Manual of Immunological Methods: Academic Press; 1997. p. 2039-2059.

15. Young AJ, Marston WL, Dessing M, Dudler L, Hein WR. Distinct recirculating and non-recirculating B-lymphocyte pools in the peripheral blood are defined by coordinated expression of CD21 and L-selectin. Blood 1997;90(12):4865-75.