Molecular Identification Using Flow Cytometry Histograms and Information Theory

Qing Zeng, Ph.D.[1], Alan J. Young, Ph.D.[2], Aziz A. Boxwala, M.B.B.S., Ph.D.[1], James Rawn, M.D.[2], William Long, Ph.D.[3], Matthew Wand, Ph.D.[4], Mikhail Salganik, Ph.D.[4], Edgar L. Milford, M.D.[5], Steven J. Mentzer, M.D.[2], Robert A. Greenes, M.D., Ph.D.[1]

[1]Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School
[2]Department of Surgery, Brigham and Women's Hospital, Harvard Medical School
[3]Department of E.E. and C.S., Massachusetts Institute of Technology
[4]Department of Biostatistics, Harvard School of Public Health
[5]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

Abstract

Flow cytometry is a common technique for quantitatively measuring the expression of individual molecules on cells. The molecular expression is represented by a frequency histogram of fluorescence intensity. For flow cytometry to be used as a knowledge discovery tool to identify unknown molecules, histogram comparison is a major limitation. Many traditional comparison methods do not provide adequate assessment of histogram similarity and molecular relatedness. We have explored a new approach – applying information theory to histogram comparison, and tested it with histograms from 14 antibodies over 3 cell types. The information theory approach was able to improve over traditional methods by recognizing various non-random correlations between histograms in addition to similarity and providing a quantitative assessment of similarity beyond hypothesis testing of identity.

Introduction

Flow cytometry is a clinical and experimental technique for quantitatively measuring the expression of individual molecules on cells. Molecular expression is typically assessed using fluorochome-labeled monospecific antibodies and a laser-activated cytometer. In addition to quantifying the expression of known molecules, flow cytometry can also be used as a discovery tool to identify unknown molecules through comparison with known ones. Whether the target molecule is known or unknown, the reactivity of the antibody is represented by a frequency histogram of fluorescence intensities of the analyzed cells. The comparison of molecules is achieved through comparison of histograms. The technology of flow cytometry is well established. A major limitation of flow cytometry has been the existing methods of histogram comparison.

Attempts to compare histograms have typically relied upon statistical methods that compare only one feature of the histogram and do not provide adequate discrimination for investigators to cluster antibodies with like patterns of reactivity (1-3). For example, the Kolmogorov-Smirnov (KS) test compares the cumulative frequency distribution; KS can be useful if the histograms are identical; but it ignores many of the features (e.g. overall shape) of the histogram that may reflect molecular relatedness or similarity (2, 3). Thus, the application of KS may not identify antibodies that recognize the same molecule, but have slightly different histograms. The ability to characterize and compare non-identical histograms will determine the usefulness of flow cytometry as a knowledge discovery tool.

In our research, we have explored the application of information theory to flow cytometry histogram comparison. Unlike a statistical test for identity, the information theory approach recognizes any non-random correlation between histograms and can provide a quantitative assessment of similarity. We have found that the information theory approach captures features of flow cytometry histograms that are undetected by the traditional KS method.

Background

Flow Cytometry

The technique of flow cytometry, combined with monoclonal antibody technology, is an invaluable tool for both research and clinical applications (4). Fluorescent labeled antibodies can be used as a probe to measure target molecule expression on cell surfaces. 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, and can be visualized under fluorescence microscopy. Flow cytometry scales up the power of this technique, by analyzing large numbers of cells sequentially using automated fluorescence detectors. As each antibody-bound cell passes before the detectors, it will emit a pulse of fluorescence which will be specifically detected and

1067—5027/01/$5.00

recorded as a "fingerprint" of cellular protein expression. The total fluorescence emitted by a single cell is then calculated, and this value entered into a histogram. When a population of cells is analyzed in this way, the histogram generated will define a molecular "fingerprint" of protein expression for the target antigen. In the past, analog signal processing was used, which limited the precision available for quantitative analysis of flow cytometry data. With the development of digital signal processing hardware, it has become feasible to obtain more precise quantitative data. The goal of our studies was to design rigorous testing methods which could be applied to these histograms, to enhance the amount of information which can be obtained through this relatively common procedure.

KS test

The KS test makes a comparison between a sample cumulative distribution function $F_S(x)$, and some theoretical cumulative distribution function $F_T(x)$(5). The null hypothesis is that the sample was drawn from the population with specified theoretical distribution: H_0: $F(x) = F_T(x)$ *for all* x. The alternative is:

H_A : $F(x) \neq F_T(x)$ *for at least one* x. The test statistic is: $D = max\ (|F_S(x)-F_T(x)|)$. The null hypothesis is rejected at the α level of significance if D exceeds the relevant critical value.

Information theory

Information theory was developed on the foundation of Shannon's work in communication (6). In recent years it has been applied to areas including image registration and gene sequence analysis (7, 8).

Given a discrete information source A, the average information content or **entropy** H(A) can be described as the unpredictability of the source. For example, a stream of random symbols would have the highest entropy. The **conditional entropy** $H(A|B)$ is a measurement of the entropy of a information source A given another information source B. Similarly, the **conditional entropy** $H(A \cap B)$ is a measurement the entropy of information sources A and B which operate jointly or concurrently. Both **mutual information** and **distance** can be used to measure the similarity between two sources. The mutual information

$H(A;B) = H(A) - H(A|B)$,

and the **distance**

$D(A,B) = H(A \cap B) - H(A;B)$.

In the paper, to distinguish this specific distance measurement from the general concept of "distance", we refer to it as **IT distance**.

Material

Experimental Animals:

Randomly bred ewes were obtained and housed in accordance with Harvard guidelines on the care and use of experimental animals. Under general anesthesia, an efferent lymphatic draining the prescapular (subcutaneous) lymph node was cannulated and lymph collected (9). Thymus and lymph nodes were harvested after sacrifice, and cell suspensions prepared for analysis.

Test Cell Population:

Efferent lymph cells (ELU) were collected from a normal lymph node under sterile conditions, and cells harvested by centrifugation. Cell suspensions were made from unstimulated lymph nodes (LNU) and thymus (TH), and the cells harvested by centrifugation. All cell populations were washed 3 times in PBS, and reacted with monoclonal antibodies.

Test Antibodies:

All test antibodies are from murine hybridomas produced over the last 15 years, directed against proteins normally expressed by subpopulations of sheep leucocytes. Antibodies are represented by the cell clone name of the murine hybridoma, and included F10-150, Fw4, ERD2/117, FW3, ERD2/29, T2/51, T2/52, Du1-29, 36F, L2/16, M2/61, M3/89, T2/39, and 17D. The molecular "target" of each of these antibodies is known, and all recognize independent cell surface proteins with the exception of FW4 and ERD2/117, which are unique antibodies which recognize the $\beta 1$ integrin on sheep leukocytes. Reactivity was detected by indirect immunofluorescence, using FITC-conjugated goat-anti-mouse Ig diluted 1:10 with Phosphate Buffered Saline. A total of 5000 cells were analyzed on a Coulter XL flow cytometer, and the fluorescence intensity was presented on log-scale histograms containing a total of 256 channels.

Method

We have applied information theory to analyze relationships between flow cytometry histograms. The general approach is as follows:

1. Smooth the histograms
2. Calculate entropy and conditional entropy
3. Calculate mutual information and distance

Smoothing

We used smoothing of data to reduce high frequency fluctuations which are postulated to be noise. The Savitzky-Golay (polynomial) smoothing filter were used. The filter takes 2 parameters: the polynomial order and the frame size, which can be adjusted to

achieve the desired smoothing effect. We empirically chose a polynomial order of 3 and a frame size of 41 for our data set. Figure 1 shows the antibody F10-150 on the ELU cell type before and after smoothing.

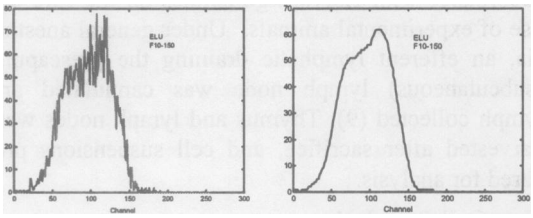

Figure 1 Antibody F10-150 with ELU cells, before and after smoothing.

Entropy

Each cytometry histogram contains a series of numbers N_i indicating the number of cells observed in each channel. We regard each number as a different symbol. For a unique N_i, its probability of occurrence in a histogram is defined as the number of channels with a cell count of N_i, divided by the total number of channels:

$P(N_i)$ = *Number of Channels with* N_i *Cells / Number of Channels*

The average information or **entropy** of a histogram A is:

$$H(A) = \sum P(N_i) \log (1/P(N_i))$$

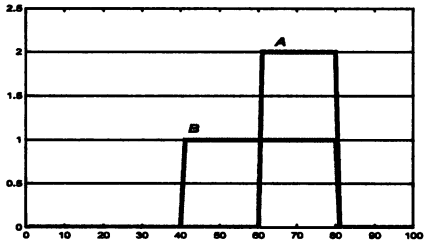

Figure 2 Simplified histogram example A, B.

Consider two histograms A and B in Figure 2 as simplified examples: A and B each have 100 channels. A has 20 channels with 2 cells and B has 40 channels with 1 cell (all other channels are 0). So for A, P(0) = 80/100 = 0.8, P(2) = 20/100 = 0.2, and H(A) = 0.8*log(1/0.8) + 0.2*(1/0.2) = 0.217. Similarly, for B, P(0) = 0.6, P(1) = 0.4 and H(B) = 0.292.

Conditional Entropy

To determine the conditional entropy between a histogram A with a series of numbers N_{ia} and a histogram B with a series of numbers N_{ib} (i = 1 to total channel number), we first have to calculate conditional and joint probabilities between N_{ia} and N_{ib}. In this study, all histograms have the same number of channels. The probability of N_{ia} given N_{ib} is defined as

$P(N_{ia}/N_{ib})$ = *(Number of Channels with* N_{ia} *cells in A and* N_{ib} *cells in A) / (Number of Channels with* N_{ib} *cells in B)*

The joint probability of N_{ix} and N_{iy} is calculated as:

$$P(N_{ia} \cap N_{ib}) = P(N_{ia}/N_{ib}) * P(N_{ib})$$

The **conditional entropy** of histogram A given histogram B is then calculated as follows:

$$H(A/B) = \sum P(N_{ia}, N_{ib}) \log (1/P(N_{ia}/N_{ib}))$$

To calculate the conditional entropy for A given B in Figure 2, we first need to calculate the conditional probabilities: P(0|0)=1, P(2|0)=0, P(0|1)=0.5, P(2|1)=0.5, and then the joint probablities: P(0,0)=0.6, P(2,0)=0, P(0,1)=0.2, P(2,1)=0.2. The conditional entropy H(A|B) = P(0,0)*log(1/p(0|0)) + P(2,0)*log(1/p(2|0)) + P(0,1)*log(1/p(0|1)) + P(2,1)*log(1/p(2|1)) = 0.120.

Mutual Information and IT Distance

Mutual information and IT distance can both be used as a measurement of the difference between two histograms. The more similar the two histograms are, the higher their mutual information and the smaller their IT distance will be. After calculating entropy and conditional entropy, it is straightforward to calculate mutual information and IT distance between histograms. The **mutual information** between two histograms is the entropy of a histogram A subtracted by the conditional entropy of another histogram A given B:

$$H(A;B) = H(A) - H(A/B) = H(B) - H(B/A)$$

The **IT distance** between one histogram A and another histogram B is the sum of the conditional entropy of B given A and given B:

$$D(A,B) = H(A \cap B) - H(A;B) = H(A) + H(B/A) - H(A) + H(A/B) = H(B/A) + H(A/B)$$

For the two histograms A and B in Figure 2, H(A;B) = 0.097 and D(A,B) = 0.292.

Evaluation

Both the information theory and the KS approach are used to perform two-way comparisons between pairs of the test antibodies, on each of the cell types. Although mutual information was also calculated, we chose to use IT distance as the information theory measurement. (Will be explained in the Discussion.) The test statistic D of the KS test was used as the KS score. We ranked the measurement results in ascending order and analyzed the measurement distributions. The scores and rankings of the comparison

between FW4 and ERD2-117 (the only two antibodies that recognize the same antigen) were compared.

Results

The KS and IT distance measurements were calculated for all 273 possible two-way comparisons: 14 test antibodies over the 3 cell types (ELU, LNU, TH). The mean, minimum and maximum of the KS and IT distance scores are shown in Table 1. The distributions of the scores are shown in Figures 3 and 4.

Table 1 The mean, minimum and maximum of the IT distance and KS measurements for the 273 comparisons between 14 antibodies.

	Mean	Min	Max	Std. Dev.
IT distance	2.205	0.915	3.615	0.538
KS	0.428	0.050	0.960	0.203

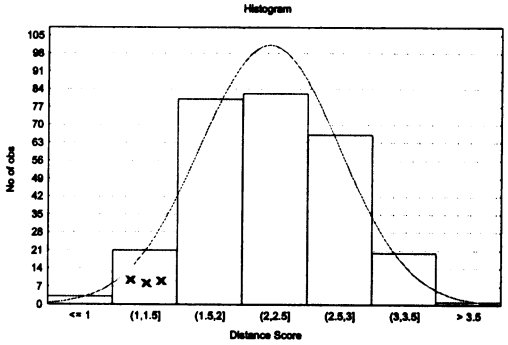

Figure 3 Distribution and normal fitting curve of the IT distance measurements. The "x" indicates the bins where the FW4 and ERD2-117 comparison scores belong.

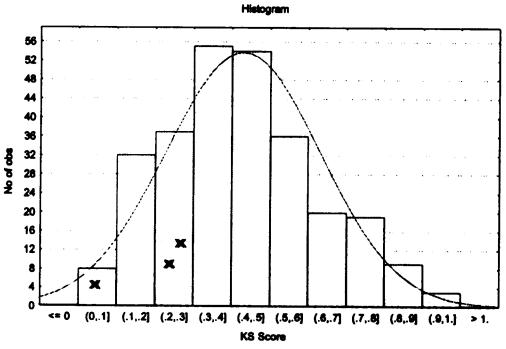

Figure 4 Distribution and normal fitting curve of the KS measurements. The "x" indicates the bins where the FW4 and ERD2-117 comparison scores belong.

The scores and rankings of the comparisons between FW4 and ERD2-117 over the 3 cell types are shown in Table 2. Since for both measurements, a lower score indicates better matching, a lower ranking also indicates better matching. To provide a better sense of the scoring for the FW4 and ERD2-117 comparison relative to the distribution of all scores, we marked the bins where the 3 scores for FW4 and ERD2-117 belong to in "x". (Figures 3 and 4)

Table 2 IT distance and KS scores and rankings (of the scores) for FW4 and ERD2-117 comparison on ELU, LNU and TH.

	IT Distance		KS	
	Score	Ranking	Score	Ranking
ELU	1.26	7	0.22	44
LNU	1.09	10	0.21	41
TH	1.36	13	0.07	4

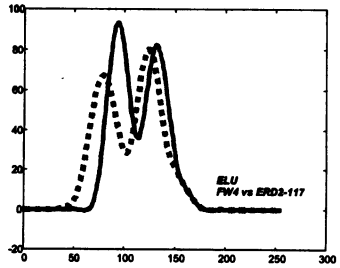

Figure 5 Antibodies FW4 and ERD2-117 with ELU cells. Rankings: IT distance – 7/273, KS – 44/273.

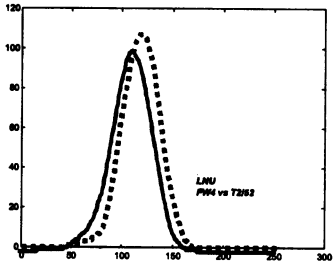

Figure 6 Antibodies FW4 and T2/52 with LNU cells. Rankings: IT distance – 4/273, KS – 14/273.

As shown in Table 2, IT distance gave lower rankings to the comparisons between the antibodies FW4 and ERD2-117 (which recognize the same antigen) than KS did. In other words, IT distance recognized the similarity between them better than KS did (The KS test rejected all three pairs as identical.). This is an improvement, especially because the difference between FW4 and ERD2-117 histograms is considered very large for antibodies that recognize the same antigen (Figures 5) and there are other test antibodies with very similar histograms but not recognizing the same antigen (Figure 6).

Also worth noting is information theory's capability of recognizing non-similar, yet non-random correlation between histograms. On ELU, F10-150 and DUI-29 histograms have a symmetrical correlation

that domain experts deemed to require further examination and their comparison received a much better matching score (ranked 26/273) from IT distance than from KS (ranked 65/273) (Figure 7).

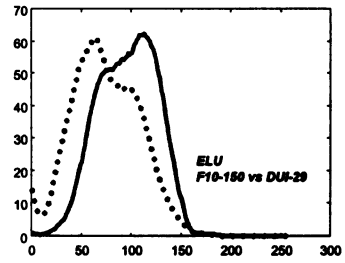

Figure 7 Antibodies FW10-150 and DUI-29 with ELU cells. Rankings: IT distance – 26/273, KS – 65/273.

Discussion

The information theory approach has shown promise in matching flow cytometry histograms and can potentially be used for molecular identification. Compared to KS, information theory has two major strengths: (1) Information theory is able to capture various non-random correlations between histograms in addition to similarity; (2) Information theory is able to provide a quantitative assessment of similarity whereas KS is a statistic test for identity. These strengths are especially valuable, because antibodies with comparable patterns of reactivates often produce similar, but non-identical histograms.

We found IT distance to be a better measurement for histogram "similarity" than mutual information. The IT distance between two identical histograms is always zero, while their mutual information can be anything, depending on their original entropy. This makes it harder to infer "similarity" from the mutual information measurement. For example, it is possible for two identical histograms to have lower mutual information than two other non-identical histograms.

For our data set, it was necessary to smooth the histograms. The entropy of smoothed histograms is lower because in the raw data, part of the entropy is contributed by the random noise. If the random noise is not filtered out, it can mask the real difference between the histograms.

The information theory approach has some limitations. It does not differentiate the similarity correlation from other types of correlation. Although there are cases where it is desirable to recognize those other types of correlation, their implications can be very different and should not always be treated the same way. Compared to KS, IT distance or mutual information measurements are more expensive computationally which can be an issue in real-time molecular identification.

In this study, we have tested a relatively small number of antibodies and cell types with only two antibodies recognizing the same antigen. To quantitatively validate information theory as better than KS, a larger sample size is needed. Currently, we are generating more sample data for further testing.

We have also begun exploring other means of applying information theory to histogram analysis. For example, peakness (second derivative) of a histogram is an important feature that domain experts use in visual histogram comparison. We plan to combine information theory with peakness.

References

1. Bagwell CB, Hudson JL, Irvin GL. Nonparametric flow cytometry analysis. J Histochem Cytochem 1979;27(1):293-6.
2. Lampariello F. On the use of the Kolmogorov-Smirnov statistical test for immunofluorescence histogram comparison. Cytometry 2000;39(3):179-88.
3. Young IT. Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. J Histochem Cytochem 1977;25(7):935-41.
4. Smith KB, Ellis SA. Standardisation of a procedure for quantifying surface antigens by indirect immunofluorescence. J Immunol Methods 1999;228(1-2):29-36.
5. Daniel WD. Biostatistics: A Foundation for Analysis in the Health Sciences: John Wiley & Sons, Inc.; 1995.
6. Shannon CE. A Mathematical Theory of Communication. The Bell System Technical Journal 1948;27(July,October):379-423, 623-656.
7. Schneider TD. Information content of individual genetic sequences. J Theor Biol 1997;189(4):427-41.
8. 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.
9. 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.