Nonparametric Flow Cytometry Analysis
C. B. BAGWELL, J. L. HUDSON and G. L. IRVIN, III

Surgical Immunology Laboratory, Veterans Administration Hospital and the Departments of Surgery and Microbiology, University of Miami School of Medicine, Miami, Florida 33125

Received for publication June 16, 1978

A nonparametric statistical test for the analysis of flow cytometry derived histograms is presented. The method involves smoothing and translocation of data, area normalization, channel by channel determination of the mean and S. D., and use of Bayes' theorem for unknown histogram classification. With this statistical method, different sets of histograms from numerous biological systems can be compared.

Flow cytometry is becoming an increasingly popular technique for measuring parameters such as cellular DNA content, RNA content, protein content, immunofluorescence, cell and nuclear size, light scatter, and fluorescence anisotropy. These parameters are usually displayed as either one or two dimensional histogram with the measured parameter(s) being on the independent axis and the number of events on the dependent axis.

One frequent problem associated with flow cytometry analysis is describing differences between one group of histograms and another group. The generally accepted method of comparison has been one of arbitrarily selecting a "typical" histogram from each set and qualitatively describing differences between the histograms. This method suffers from such serious drawbacks as selection bias, lack of information about channel to channel variability, qualitative rather than a quantitative description of differences, and lack of predictive capability.

In 1977 the Kolmogorov-Smirnov (KS) test was applied to the comparison of two different histograms (5). The test was advantageous because it made no assumptions concerning the distribution of variables or distribution of sampling error and thus was described as a nonparametric statistical analysis.

The test described in this paper is a natural extension of the 2 sample KS test. The KS test as described is only suitable for the comparison of two histograms. The method described here extends this analysis to two groups of histograms. It also is capable of finding small statistically significant areas in histogram comparisons that may have special biological meaning.

The general procedure is outlined as follows:

1. Smooth and translocate all histograms to the same first mode channel (optional).
2. Normalize the relevant area in each histogram to 1.
3. Calculate the mean and standard deviation for the number of events in each channel.
4. Perform a channel-by-channel statistical test such as the Student's (Gossett) t-test.
5. Select desired zone (descriptor) where there is a statistical difference between the two histograms. Use Bayes' theorem to give the probability that the unknown histogram belongs to one of the two groups.

This procedure is nonparametric in the sense that (like the KS test) it does not attempt to reduce the data to a few parameters (eg. the number of cells in G0-G1, S, and G2-M); therefore, it can be applied to any two groups of histograms. The type of statistical test that is employed in Step 4 is optional and if there are serious reservations in assuming the sampling error is normally distributed, one could use a test such as the KS test instead of the Student's t-test.

MATERIALS AND METHODS

Smoothing and translocation: Smoothing the data is optional and is rarely necessary. In fact, one must be careful not to use a smoothing routine that will result in significant loss of structural information. A good review of current smoothing techniques has recently been published (1). The smoothing routine that we normally use is a variable window parabolic smoothing routine. It was specifically designed to smooth histograms less in regions of high information density (low channel numbers) than in regions of low information density (high channel numbers). The result of a typical smoothing operation on actual data is shown in Figures 1A and B. We often use the smoothing routine because it makes our automatic peak searching algorithms much more reliable. Also, after smoothing, one can easily interpolate between points, which is necessary for histogram translocation.

In the case of DNA histograms with a distinct G0-G1 peak, we normally translocate the histograms to a first modal channel of 40. We have found that there is a small degree of shifting that occurs from one run to another. By translocating all the DNA histograms to a single position, we minimize this shifting variability. If this shifting is not compensated, there will be a large reduction in the sensitivity of this analysis.

Another technique for minimizing histogram shifting is the use of internal standards such as beads, chicken erythrocytes, or sorted octoploid cells. One can then use a translocation routine to move all the standards to the same channel number. It is best to use a standard that has approximately the same staining properties as the sample.
FIG. 1. A) Raw histogram. DNA histogram of 90 hr cultured PHA stimulated human peripheral lymphocytes (HPL). B) Smoothed histogram. The raw histogram is smoothed with a variable window parabolic smoothing routine.

As seen in Figure 2, the histogram (Figure 1B) has been translocated to channel 40.

Normalize to constant area: Once all the histograms in each series have been smoothed and translocated, they are converted to probability density functions so that sets of histograms can be averaged. In order to exclude unwanted debris or clumping, one selects the area of interest in the histograms such as from channels 25 to 100 in our example. The normalizing algorithm determines the area from the lower boundary at channel 25 to the upper boundary at 100 and then divides every channel accumulation by the selected area. The result is a normalization of the histogram such that the area from the lower boundary to the upper boundary is equal to 1. Every histogram in the two sets is treated in the same fashion (see Fig. 3).

Mean and standard deviation: After all the histograms have been normalized to a constant area in the range specified, an arithmetic mean and S. D. are found for each channel number. The locus of means for a particular data set is referred to as the mean histogram. With this procedure we form two types of histograms which we arbitrarily will refer to as type A and type B (see Fig. 4). The variability of each data point is shown as a horizontal line representing one S. D. above and below the mean (see Figure 4).

Statistical test: The two tailed Student's (Gossett) t-test (3) was used because of the relatively few samples in each set (<10). The $P$ values are represented in histogram form. Those $P$ values that are greater than $P = 0.05$ are not shown and those values which are less than $P = 0.05$ are plotted on a log scale as shown in Figure 5. The end result is that one can visually inspect the graph for areas of statistical significance. In the example shown in Figure 5, there are 3 major areas of significant difference: G0-G1, late S, and G2-M. It is interesting to note that this procedure can detect small differences in S which would be diluted when comparing only percent S.

Descriptor selection and use of Bayes' theorem: As seen at the $P < 0.05$ level in Figure 5, there are three statistically significant zones. The area underneath the curve in any one of these zones is referred to as a descriptor. The only restriction in selecting a descriptor zone is to stay within a region where the mean histograms do not intersect. As seen in Figure 5, by selecting the area under the curve from channels 86 to 95, we can then determine the probability density functions (see Figure 6) of the descriptor for both sets of histograms.

With unknown histograms we perform the same operations of smoothing, translocation, normalization, and descriptor selection and use Bayes' theorem to determine the likelihood that the histogram is
of analysis. The control group of DNA histograms were from PHA stimulated human peripheral lymphocytes (HPL) and the test group of histograms were from PHA stimulated HPL treated with 200 μg of thymocyte extract. The analysis demonstrated significant differences between control and test groups in G0-G1, late S, and G2-M suggesting that the extract was perturbing the cycling characteristics of stimulated human peripheral lymphocytes. Because of the nonparametric nature of this test, there is no quantitation of these differences.

in class A or class B. Bayes' theorem requires knowledge of the probability of occurrence of both type A and type B. If this information is not known, it is generally assumed that the probability of occurrence for each type is 0.5 (2). Figure 7 shows the result of applying Bayes' theorem to each member of our training set and it correctly classified each histogram.

**DISCUSSION**

The general procedure described here for determining areas of significant difference between two groups has a number of biological applications. Probably, the most important application is the quantitation of significant changes in flow cytometric histograms associated with experimental manipulation. The data used in this paper (4) is a good example of this type
Nonparametric flow cytometry analysis can also be used in another type of biological application. By using a zone of significant difference between two groups as a descriptor, Bayes' Theorem can be used to classify histograms. For example, if we had a group of histograms from "normal" tissue and a group of histograms from "abnormal" tissue and a significant difference was found between the two groups, Bayes' theorem could be used to classify unknown histograms in either "normal" or "abnormal" categories. There are a number of clinical situations where this diagnostic capability would be advantageous.

This procedure can be modified in several ways depending on the investigator's needs. The program can theoretically be extended to two or more dimensions. Instead of area normalization, there would be volume normalization for two parameters. The degree of statistical significance could be indicated by density plots or color topography maps. Also Bayes' theorem can be extended to any number of parameters. The number of features or descriptors from the two groups of histograms can theoretically be increased which may result in an enhancement of specificity for population identification.

The chief limitation of the sensitivity of this method is the variability in the coefficient of variation (C.V.). This variability is mostly a function of the state of the flow cytometer (e.g. partial orifice block, laser tuning, dirty optics, sample flow rate, or sheath flow rate). Thus, it is always a good practice to run a standard before and during experimental runs.

Finally, the number of numeric manipulations involved in this statistical procedure is quite small for one parameter. We use a desktop microprocessor unit (Tektronix 4051 Beaverton, Oregon) for all our work. An analysis of the size presented in this paper (four histograms in each group) takes less than 5 min. Most of the new microprocessor units available today can effectively handle two parameter analyses. For three or more parameters, a larger scale computer is necessary.

More information concerning the computer software that executes the above statistical test can be obtained by writing the authors or Coulter Electronics Inc., Hialeah, Florida.

SUMMARY

A nonparametric statistical analysis was presented that involves smoothing, translocation, area normalization, statistical examination of differences, and use of Bayes' theorem. This statistical test can be used for any type of histogram and can be readily extended to two or more parameters.

ACKNOWLEDGMENTS

The authors wish to gratefully acknowledge the valuable assistance of A. Brunsting and F. Briese and the generous support of Coulter Electronics Inc.

LITERATURE CITED