The probability binning algorithm and frequency difference gating

Hans Jürgen Hoffmann, Department of Pulmonary Medicine, Aarhus University Hospital, DK

The theory of probability binning and frequency difference gating will be discussed on the basis of three articles that describe the probability binning algorithm in univariate analyses, the extension into a multivariate space, and application to complex data sets, but will not address the mathematical arguments that the theory is based on.

Probability Binning Comparison: A Metric for Quantitating Univariate Distribution Differences (Cytometry 45:37–46, 2001)

Mario Roederer, Adam Treister, Wayne Moore, Leonore A. Herzenberg

Background: Comparing distributions of data is an important goal in many applications. For example, determining whether two samples (e.g., a control and test sample) are statistically significantly different is useful to detect a response, or to provide feedback regarding instrument stability by detecting when collected data varies signifi-cantly over time.
Methods: We apply a variant of the chi-squared statistic to comparing univariate distributions. In this variant, a con-trol distribution is divided such that an equal number of events fall into each of the divisions, or bins. This ap-proach is thereby a mini-max algorithm, in that it mini-mizes the maximum expected variance for the control distribution. The control-derived bins are then applied to test sample distributions, and a normalized chi-squared value is computed. We term this algorithm Probability Binning.
Results: Using a Monte-Carlo simulation, we determined the distribution of chi-squared values obtained by compar-ing sets of events derived from the same distribution. Based on this distribution, we derive a conversion of any given chi-squared value into a metric that is analogous to a t-score, i.e., it can be used to estimate the probability that a test distribution is different from a control distribu-tion. We demonstrate that this metric scales with the difference between two distributions, and can be used to rank samples according to similarity to a control. Finally, we demonstrate the applicability of this metric to ranking immunophenotyping distributions to suggest that it in-deed can be used to objectively determine the relative distance of distributions compared to a single control.
Conclusion: Probability Binning, as shown here, pro-vides a useful metric for determining the probability that two or more flow cytometric data distributions are different This metric can also be used to rank distributions to identify which are most similar or dissimilar. In addition, the algorithm can be used to quantitate contamination of even highly-overlapping populations.

Probability Binning Comparison: A Metric for Quantitating Multivariate Distribution Differences (Cytometry 45:47–55, 2001)

Mario Roederer, Wayne Moore, Adam Treister, Richard R. Hardy, Leonore A. Herzenberg

Background: While several algorithms for the compar-ison of univariate distributions arising from flow cyto-metric analyses have been developed and studied for many years, algorithms for comparing multivariate dis-tributions remain elusive. Such algorithms could be useful for comparing differences between samples based on several independent measurements, rather than differences based on any single measurement. It is conceivable that distributions could be completely dis-tinct in multivariate space, but unresolvable in any combination of univariate histograms. Multivariate com-parisons could also be useful for providing feedback about instrument stability, when only subtle changes in measurements are occurring.

Methods: We apply a variant of Probability Binning, de-scribed in the accompanying article, to multidimensional data. In this approach, hyper-rectangles of *n* dimensions (where *n* is the number of measurements being com-pared) comprise the bins used for the chi-squared statistic. These hyper-dimensional bins are constructed such that the control sample has the same number of events in each bin; the bins are then applied to the test samples for chi-squared calculations.

Results: Using a Monte-Carlo simulation, we determined the distribution of chi-squared values obtained by compar-ing sets of events from the same distribution; this distri-bution of chi-squared values was identical as for the uni-variate algorithm. Hence, the same formulae can be used to construct a metric, analogous to a t-score, that estimates the probability with which distributions are distinct. As for univariate comparisons, this metric scales with the difference between two distributions, and can be used to rank samples according to similarity to a control. We apply the algorithm to multivariate immunophenotyping data, and demonstrate that it can be used to discriminate distinct samples and to rank samples according to a bio-logically-meaningful difference.

Conclusion: Probability binning, as shown here, pro-vides a useful metric for determining the probability with which two or more multivariate distributions represent distinct sets of data. The metric can be used to identify the similarity or dissimilarity of samples. Finally, as demon-strated in the accompanying paper, the algorithm can be used to gate on events in one sample that are different from a control sample, even if those events cannot be distinguished on the basis of any combination of univari-ate or bivariate displays.

Frequency Difference Gating: A Multivariate Method for Identifying Subsets That Differ Between Samples (Cytometry 45: 56–64, 2001)

Mario Roederer and Richard R. Hardy

Background: In multivariate distributions (for example, in 3- or more color flow cytometric datasets), it can become difficult or impossible to identify populations that differ be-tween samples based only on a combination of univariate or bivariate displays. Indeed, it is possible that such differences can only be identified in "n"-dimensional space, where "n" is the number of parameters measured. Therefore, computer assisted identification of such differences is necessary. Such a method could be used to identify responses (i.e., by com-paring cell samples before and after stimulation) in exquisite detail by allowing complete analysis of the collected data on only those events which have responded.

Methods: Multivariate Probability Binning can be used to compare different datasets to identify the distance and sta-tistical significance of a difference between the distributions. An intermediate step in the algorithm provides access to the actual locations within the n-dimensional comparison which are most different between the distributions. Gates based on collections of hyper-rectangular bins can then be applied to datasets, thereby selecting those events (or clusters of events) that are different between samples. We term this process Frequency Difference Gating.

Results. Frequency Difference Gating was used in several test scenarios to evaluate its utility. First, we compared PBMC subsets identified by solely by immunofluorescence staining: based on this training data set, the algorithm automatically generated an accurate forward and side-scatter gate to identify lymphocytes. Second, we applied the algorithm to identify subtle differences between CD4 memory subsets based on 8-color immunophenotyping data. The resulting 3-dimensional gate could resolve cells subsets much more frequent in one subset compared to the other; no combination of two-dimensional gates could accomplish this resolution. Finally, we used the algorithm to compare B cell populations derived from mice of dif-ferent ages or strains, and found that the algorithm could find very subtle differences between the populations.

Conclusion. Frequency Difference Gating is a powerful tool that automates the process of identifying events com-prising underlying *differences* between samples. It is not a clustering tool; it is not meant to identify subsets in multidimensional space. Importantly, this method may reveal subtle changes in small populations of cells, changes that only occur simultaneously in multiple dimen-sions in such a way that identification by univariate or bivariate analyses is impossible. Finally, the method may significantly aid in the analysis of high-order multivariate data (i.e., 6-12 color flow cytometric analyses), where identification of differences between datasets becomes so time-consuming as to be impractical.