Biometrical Journal **47** (2005) 5, 740--754 DOI: 10.1002/bimj.200310142

Classifying Antibodies Using Flow Cytometry Data: Class Prediction and Class Discovery

M. P. Salganik[*, 1], **E. L. Milford**[2], **D. L. Hardie**[3], **S. Shaw**[4], and **M. P. Wand**[5]

[1] Department of Biostatistics, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA, 02115, USA

[2] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA

[3] MRC Centre for Immune Regulation, Institute of Biomedical Research, University of Birmingham, Birmingham 2TT, UK

[4] Experimental Immunology Branch, National Cancer Institute, National Institute of Health, Bethesda, MD 20892-1360, USA

[5] Department of Statistics, School of Mathematics, University of New South Wales, Sydney, 2052, Australia

Received 21 October 2003, revised 14 January 2005, accepted 16 March 2005

Summary

Classifying monoclonal antibodies, based on the similarity of their binding to the proteins (antigens) on the surface of blood cells, is essential for progress in immunology, hematology and clinical medicine. The collaborative efforts of researchers from many countries have led to the classification of thousands of antibodies into 247 clusters of differentiation (CD). Classification is based on flow cytometry and biochemical data. In preliminary classifications of antibodies based on flow cytometry data, the object requiring classification (an antibody) is described by a set of random samples from unknown densities of fluorescence intensity. An individual sample is collected in the experiment, where a population of cells of a certain type is stained by the identical fluorescently marked replicates of the antibody of interest. Samples are collected for multiple cell types. The classification problems of interest include identifying new CDs (class discovery or unsupervised learning) and assigning new antibodies to the known CD clusters (class prediction or supervised learning).

These problems have attracted limited attention from statisticians. We recommend a novel approach to the classification process in which a computer algorithm suggests to the analyst the subset of the "most appropriate" classifications of an antibody in class prediction problems or the "most similar" pairs/groups of antibodies in class discovery problems. The suggested algorithm speeds up the analysis of a flow cytometry data by a factor 10–20. This allows the analyst to focus on the interpretation of the automatically suggested preliminary classification solutions and on planning the subsequent biochemical experiments.

Key words: Classification; Monoclonal antibodies; Flow cytometry; Dissimilarity measure; Kernel smoothing; SiZer; Class discovery; Class prediction; Unsupervised learning; Supervised learning.

1 Introduction

During the past 20 years thousands of monoclonal antibodies have been discovered to bind specifically to 247 molecules (antigens) on the surface of certain blood cells (leukocytes). Each antigen was given a unique cluster of differentiation (CD) number (e.g. Schlossman et al., 1995; Kishimoto et al.,

[*] Corresponding author: e-mail: salganik@hsph.harvard.edu, Phone: +01 617 432 3689, Fax: +01 617 432 3755

1997; Mason et al., 2002) that denotes both the particular antigen and the cluster of antibodies of different molecular structure which bind exclusively to this antigen. The discovery of new CDs is essential for the progress in immunology, hematology and clinical medicine. The current applications of CD-based technology are numerous and include leukemia/lymphoma phenotyping, monitoring the progression of HIV/AIDS, and diagnosing transplant compatibility (Stuart and Nicholson, 2000; Givan, 2001). It is estimated (Zola and Swart, 2003) that only 10–20% of the existing antigens have been discovered and that about 100 new CD clusters will be designated in December 2004 at the 8th HLDA (Human Leucocyte Differentiation Antgiens, www.hlda8.com) workshop.

Classifying newly discovered antibodies involves resolving two problems which, following the terminology used in a different setting by Golub et al. (1999), we will call "class prediction" and "class discovery". In the class prediction analysis, a newly discovered antibody is assigned to the most appropriate known CD cluster. The class discovery analysis, on the other hand, studies the antibodies that do not fit in any of the known CD clusters and thus aims to identify new ones. Class prediction and class discovery analysis are often denoted as "supervised" and "unsupervised" learning in the statistical literature. In both class discovery and class prediction analysis it is widely recognized that analyzing panels of flow cytometry data (conventionally called "blind panels" in the biological literature) is often very useful (e.g. Spiegelhalter and Gilks, 1987; Gilks, Oldfield and Rutherford, 1989; Gilks and Shaw, 1995; Miyazaki et al. 1997; Mason et al, 1997; Hilgert and Drbal, 2002).

A blind panel may be viewed as an $(n + n_0) \times p$ array of experiments. The columns correspond to p different blood cell populations. The experiment in the i-th row and j-th column results in a large univariate sample of fluorescence measurements for thousands of cells from population j. For experiments $i = 1, \ldots, n$, the cells in each of the cell populations $j = 1, \ldots, p$ are stained by a large number of replications of antibody i. For control experiments $i = n + 1, \ldots, n + n_0$, the cells are not stained. The binding of a fluorescently marked antibody to an antibody-specific antigen on the surface of a cell increases the cell's fluorescence. Therefore, an increase in the fluorescence of cell population j relative to the "baseline" fluorescence observed in the control experiments, which is caused by staining the cells with fluorescently marked antibody i, suggests that antigens specific to this antibody are expressed on the surface of the type j cells. The similarity in the pattern of distributions of fluorescence intensities across multiple cell populations $j = 1, \ldots, p$ observed for a pair or group of antibodies suggests a possible identity in their antigens. For example, Figure 1 shows kernel density estimates of log-fluorescence for $p = 20$ cell populations (TONSIL.T, ..., K562) stained by the fluorescently marked antibodies (antibodies B031 and B032 belong to the CD139 cluster; antibodies B034 and B036 belong to the CD55 cluster) and distributions of a baseline log-fluorescence observed in the control experiment. The discussion of the log-fluorescence scale is presented in Section 2.1. The fluorescence of 5 of the 20 cell populations (TONSBDEN,GCB, HPB.ALL, MEM.B, H.MY) increased after staining by the antibodies B031 and B032, which suggests that CD139 antigens were expressed on the surface of these cell populations. The fluorescence of 16 cell populations increased after staining by the antibodies from the CD55 cluster. Overall, the patterns of fluorescence distributions observed for the antibodies from the same CD are substantially similar. The goal of the analysis of a blind panel is to detect such similarities for a pair or group of antibodies within the panel. Additional biochemical experiments are then implemented to evaluate whether the antibodies inducing a similar pattern of fluorescence in multiple cell populations bind to the same antigen. A typical blind panel contains a mixture of known and unknown CD specificities. The dimensions are typically in the range $100 \leq n \leq 500$, $1 \leq n_0 \leq 5$ and $20 \leq p \leq 100$ which means that the number of fluorescence samples is in the thousands. This makes visual search for similarities among the antibodies difficult.

Initial approaches to analyzing the blind panels characterized an individual flow cytometry experiment by the estimated percent of the "positive" cells that had antibody-dependent antigen on their surface. We discuss this terminology in Section 2.1. In subsequent analyses the results of an individual experiment were summarized by the mean (Gilks and Shaw, 1995; Kishimoto et al., 1997) or the mean and the standard deviation values (Hallam et al., 1997) of fluorescence intensity. These values are then used to assign a dissimilarity score $D(i_1, i_2)$ to each pair of antibodies (i_1, i_2). Low dissimilar-

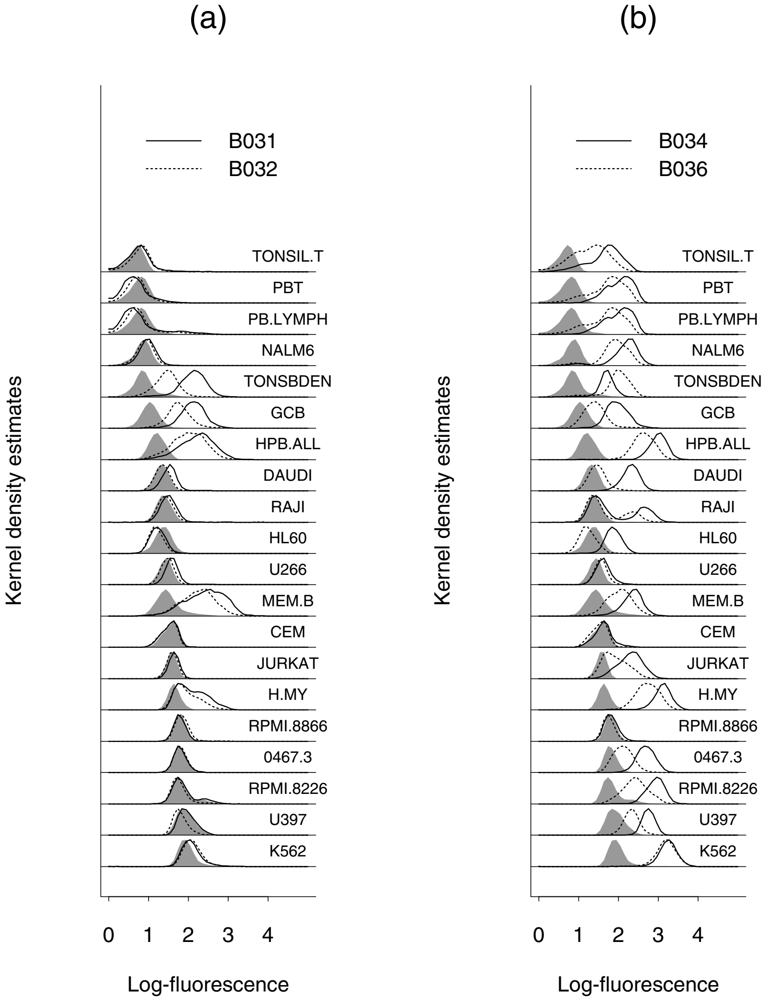

Figure 1 Kernel density estimates (KDE) of log-fluorescence intensity distributions for unstained cell populations (gray polygons), and cell populations stained by antibodies (solid and dashed lines) binding to CD139 (a) and CD55 (b) antigens. The KDE curves are scaled so that their maximum value is equal to one. We used the bandwidth $h = 0.05$. The details of KDE estimation are presented in Section 2.2.4. The description of the panel can be found in Mason et al. (1997).

ity scores indicate high similarity among the antibodies. The matrix of dissimilarity scores is used for hierarchical clustering of antibodies, and the summary of the resulting clusters is displayed in the form of a dendrogram. A data analyst uses the dendrogram to visually identify antibodies that are clustered together and may therefore have identical specificity. The usefulness of this conventional approach has been demonstrated by its application to the analysis of large panels of flow cytometry data (Shaw et al., 1995; Kishimoto et al., 1997), but is bounded by the well-know deficiencies of the hierarchical clustering. The structure of a dendrogram is often too sensitive to the intercluster dissimilarity definition (i.e. the choice between group average, nearest neighbor and further neighbor methods) and small changes in the data. It is often too difficult to identify clusters based on visual inspection of the dendrogram or to even estimate the number of clusters in a dataset. Hierarchical clustering always assigns the object into one of the non-overlapping clusters even if the data do not contain enough information to separate the "true" CD clusters, which may lead to incorrectly formed clusters. These limitations of a hierarchical clustering are well-known from the statistical literature (e.g. Kaufman and Rousseeuw, 1990; Hastie, Tibshirani and Friedman, 2001) and have been mentioned in papers describing the analysis of blind panel data (e.g. Gilks and Shaw, 1995; Kishimoto et al., 1997). It is also evident that in the class prediction problem the re-grouping of the antibodies of known CD specificity by hierarchical clustering is unnecessary because the biologically correct clustering of these antibodies (i.e. their CD number) is already known. In addition to that important information may be lost when the distribution of log-fluorescence is summarized by the mean or mean and standard deviation. For instance in Figure 1 the distributions of the log-fluorescence of the RAJI cells stained by B034 and B036 antibodies have similar bi-modal distributions, that the conventional approach could miss.

We propose a different approach to analyzing blind panels, in which evaluating the similarity between fluorescence patterns is based on visual inspection of a subset of the stacks of fluorescence distributions (as shown in Figure 1) and a corresponding subset of the stacks of SiZer maps (Chaudhuri and Marron, 1999). SiZer maps are described in Section 2.2.4 and shown in Figure 2. The role of the automatic algorithm is limited to suggesting to the analyst the subset of the ‘‘most appropriate’’ classifications of antibody (in class prediction problems) or the ‘‘most similar’’ pairs/groups of the antibodies (in class discovery problems). This approach is similar in spirit to that used in the search for information on the Internet, where rapid search engines help focus the user’s attention on the subset of possibly relevant objects.

In Section 2 we discuss five possible definitions of the dissimilarity score between a pair of antibodies: the score described by Hallam et al. (1997), and four others which, to the best of our knowledge,

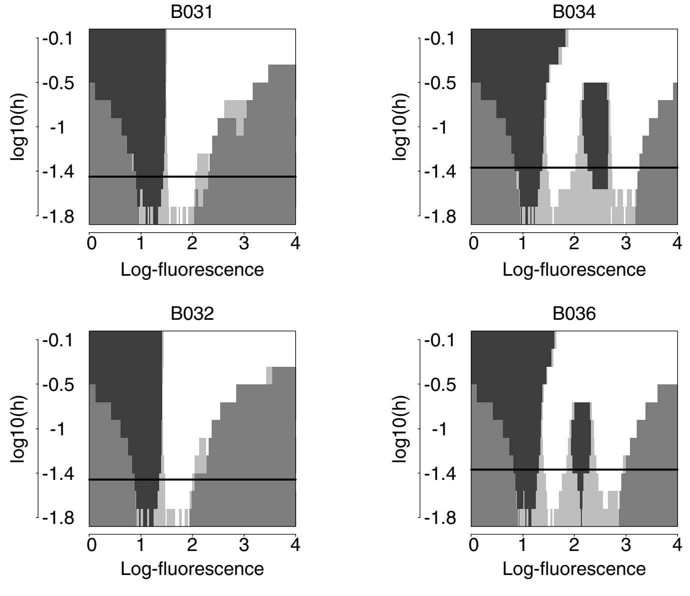

Figure 2 SiZer maps for densities of antibodies B031,B032, B034 and B036 and cell population RAJI. The shading marks increasing (black), decreasing (white), or flat (lighter gray) regions of function $E\{\hat{f}(x;h)\}$. The darker gray shading shows regions with sparse data. The horizontal lines show values for the Sheather-Jones selected bandwidths.

have never been used in the analysis of blind panel data. We found that one of these four new definitions of a dissimilarity score, which is based on the correlation between the adjusted mean values of log-fluorescence, provides an attractive alternative to Hallam's definition. This correlation-based definition of a pair-wise dissimilarity score is biologically justified, easily interpretable, and substantatively identical to the definition of a dissimilarity measure routinely used in the analysis of gene expression data.

We found it useful to distinguish between the class prediction and class discovery settings and discuss the suggested approaches to the solution of these problems in Sections 3 and 4. Section 5 presents our conclusions and suggestions for future research.

2 Dissimilarity Scores

2.1 A mixture model of fluorescence distributions

A cell's fluorescence is conventionally measured on the logarithmic scale. We denote as

$$x = \log_{10} \frac{FL}{FL_{\min}}$$

the logarithm of the ratio of observed fluorescence FL to the minimum detectable value of fluorescence $FL_{\min}$ and call it log-fluorescence. The choice of $FL_{\min}$ is arbitrary and for simplicity we will assume that it is identical for all cell populations, which is warranted since in our datasets difference between the values of the thresholds across the cell populations are small. Gilks and Shaw (1995) show how to use beads (small commercially available "microspheres" with known fluorescence intensities) to re-scale the log-fluorescence values when $FL_{\min}$ differs between cell populations. In our discussion we ignore the error caused by binning the log-fluorescence values by flow cytometer and treat these values as continuous random variables. The probability density function for cell population j stained by antibody i is denoted by f_{ij}.

For the given antibody i the cell population j may contain cells with and without antibody-specific antigens. We follow a convention accepted in the flow cytometry literature and use the term "negative" for cells without antigens and the term "positive" for cells with antigens. The distribution of log-fluorescence may be described by a simple mixture model (e.g. Lampariello and Aiello, 1998; Watson, 2001)

$$f_{ij}(x) = (1 - \alpha_{ij}) f_{ij}^{-}(x) + \alpha_{ij} f_{ij}^{+}(x) \tag{1}$$

where $f_{ij}^{-}(x)$ and $f_{ij}^{+}(x)$ are the probability density functions of log-fluorescence for the subpopulations of negative and positive cells to the antibody i in the population j and α_{ij} is the proportion of cells in the positive subpopulation. The log-fluorescence of positive cells is on the average higher but in many cases the densities $f_{ij}^{-}(x)$ and $f_{ij}^{+}(x)$ overlap.

For antibodies (i_1, i_2), which bind to the same antigen, the proportions of positive cells α_{i_1j} and α_{i_2j} and the shapes of the densities $f_{i_1j}^{+}(x)$ and $f_{i_2j}^{+}(x)$ are often very similar. We found that the locations of $f_{i_1j}^{+}(x)$ and $f_{i_2j}^{+}(x)$ are also very similar for many pairs (i_1, i_2) of identical specificity. However, in the relatively rare case when one of the antibodies has substantially lower affinity (i.e. the replicates of this antibody are detached from their antigens relatively easily) or is limiting in concentration, the location of $f_{i_1j}^{+}(x)$ may shift towards the low values of log-fluorescence.

As a first approximation it may be assumed that this shift is approximately the same for all cell types j so that

$$f_{i_2j}^{+}(x) \approx f_{i_1j}^{+}(x - \delta_{i2,i1} + \epsilon_{i1,i2,j})\,. \tag{2}$$

Here $\delta_{i2,i1}$ denotes a systematic shift caused by the difference in the affinity of antibodies and $\epsilon_{i1,i2,j}$ denotes a random shift in the location of $f_{ij}^{+}(x)$. For a given pair (i_1, i_2), the shifts $\epsilon_{i1,i2,j}$ may be considered independent across the cell types $j = 1, \ldots, p$. An example of a systematic shift in the location of $f_{ij}^{+}(x)$ may be seen in Figure 1, where the mean log-fluorescence of the positive cells stained by antibody B031 is slightly higher than the mean log-fluorescence of the positive cells stained by antibody B032, and the mean log-fluorescence of positive cells stained by antibody B036 is slightly higher than the mean log-fluorescence of positive cells stained by antibody B034.

The distribution of the log-fluorescence for negative cells is approximately the same as the distribution of the fluorescence in the population of unstained cells denoted by $f_{0j}^{-}(x)$, so that

$$f_{ij}^{-}(x) \approx f_{0j}^{-}(x)\,. \tag{3}$$

2.2 Definitions of dissimilarity scores

The pair-wise dissimilarity between antibodies is quantified by the dissimilarity score $D(i_1, i_2)$ assigned to a pair (i_1, i_2). The five scores considered here all have an important property: the rank of the dissimilarity score for the pair (i_1, i_2) in the set of possible pairs formed from the n antibodies is not affected by the linear transformation (i.e. scaling and shifting) of the log-fluorescence values. In the absence of the missing data these scores may be expressed as

$$D(i_1, i_2) = \frac{1}{p} \sum_{j=1}^{p} d_j(i_1, i_2) + c\,, \tag{4}$$

where dissimilarity score $d_j(i_1, i_2)$ quantifies the dissimilarity between antibodies i_1 and i_2, based on the distributions of the (possibly transformed) values of log-fluorescence for cell type j and c is the conveniently chosen constant. The choice of the constant does not affect the classification solution. We choose the value $c = 1$ for the correlation based dissimilarity score, described in Section 2.2.2 and the value $c = 0$ for all other definitions of the dissimilarity score. If the data for some of the experiments are missing, then the calculation is done using the cell populations for which data are available. The proportion of missing values in the well-designed panel of flow cytometry experiments is extremely small.

We will denote as $\bar{x}_{ij}$ and s_{ij} the mean value and standard deviations of observed log-fluorescence in the cell population j observed in the experiment i. The estimated mean value of a baseline log-fluorescence $\bar{x}_{0j}$

$$\bar{x}_{0j} = \frac{1}{n_0} \sum_{i=n+1}^{n+n_0} \bar{x}_{ij}$$

was calculated by averaging the mean values of log-fluorescence observed in the control experiments.

2.2.1 Mean-based dissimilarity score

The simplest definition of a dissimilarity score

$$d_j^M(i_1,\, i_2) = (\bar{x}_{i_1 j} - \bar{x}_{i_2 j})^2$$

is based on the cell-wise comparison of the magnitude of mean log-fluorescence values.

2.2.2 Dissimilarity score based on the correlation of baseline-adjusted means

Figure 1 suggests that some adjustment for the difference in the background fluorescence across the cell populations may be beneficial. A simple adjustment may be achieved by subtracting the "baseline" values of log-fluorescence from the observed values

$$\bar{x}_{ij}^{a} = \bar{x}_{ij} - \bar{x}_{0j}$$

and defining a dissimilarity score based on the sample correlation coefficient $r(i_1, i_2)$ between the adjusted mean values $\bar{x}_{i_1 j}^{a}$, $\bar{x}_{i_2 j}^{a}$

$$D^{\mathrm{COR}}(i_1, i_2) = 1 - r(i_1, i_2)$$

so that that a high correlation corresponds to a low dissimilarity. The correlation coefficient was estimated based on the non-missing values. In the absence of the missing data

$$r(i_1, i_2) = \frac{\sum_{j=1}^{p} (\bar{x}^a_{i_1 j} - \bar{x}^a_{i_1})\,(\bar{x}^a_{i_2 j} - \bar{x}^a_{i_2})}{\sqrt{\sum_{j=1}^{p} (\bar{x}^a_{i_1 j} - \bar{x}^a_{i_1})^2}\,\sqrt{\sum_{j=1}^{p} (\bar{x}^a_{i_2 j} - \bar{x}^a_{i_2})^2}}$$

with $\bar{x}^a_i = \frac{1}{p} \sum_{j=1}^{p} \bar{x}^a_{ij}$ and correlation-based definition of the dissimilarity belongs to the family of definitions, described by Eq. (4) with $c = 1$ and cell-dependent component d_j of a dissimilarity score defined by

$$d_j^{\mathrm{COR}} = -\frac{p(\bar{x}^a_{i_1 j} - \bar{x}^a_{i_1})\,(\bar{x}^a_{i_2 j} - \bar{x}^a_{i_2})}{\sqrt{\sum_{j=1}^{p} (\bar{x}^a_{i_1 j} - \bar{x}^a_{i_1})^2}\,\sqrt{\sum_{j=1}^{p} (\bar{x}^a_{i_2 j} - \bar{x}^a_{i_2})^2}}$$

2.2.3 Mean- and standard deviation-based similarity score

Hallam et al. (1997) suggest a dissimilarity score based on the normalized sample means $\bar{x}_{ij}$ and standard deviations s_{ij}

$$d_j^{\mathrm{MSD}}(i_1, i_2) = \left(\frac{\bar{x}_{i_1 j} - \bar{x}_{i_2 j}}{d_{\bar{x}_{ij}}}\right)^2 + \left(\frac{\bar{s}_{i_1 j} - \bar{s}_{i_2 j}}{d_{\bar{s}_{ij}}}\right)^2$$

where $d_{\bar{x}_{ij}} = \frac{1}{n} \sum_{i=1}^{n} |\,\bar{x}_{ij} - \bar{x}_{.j}\,|$, $\bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_{ij}$ and $d_{\bar{s}_{ij}} = \frac{1}{n} \sum_{i=1}^{n} |\,\bar{s}_{ij} - \bar{s}_{.j}\,|$, $\bar{s}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \bar{s}_{ij}$.

2.2.4 SiZer-based dissimilarity score

The SiZer map, introduced by Chaudhuri and Marron (1999), visualizes essential features of the family of kernel density estimates. Kernel density estimation is described by Scott (1992), Wand and Jones (1995), and Bowman and Azzalini (1997). It uses observations $X_1, \ldots, X_L$ from a random sample of a univariate random variable X with smooth probability density $f(x)$ for estimating f through

$$\hat{f}(x; h) = \frac{1}{Lh} \sum_{l=1}^{L} K\left(\frac{x - X_l}{h}\right),$$

where $K(z)$ is the kernel function and h is a bandwidth parameter of the kernel. In our calculation we used the Gaussian kernel $K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$. The expected value of an estimate corresponding to a particular bandwidth h is

$$E\{\hat{f}(x; h)\} = \int K_h(x - y)\, f(y)\, \mathrm{d}y\,,$$

where $K_h(u) = h^{-1} K\left(\frac{u}{h}\right)$. Kernel estimation therefore provides a spatially averaged image of $f(.)$ and spatial resolution of this image is regulated by the parameter h. If the function f contains local maxima with different widths (spatial scales), then it may be beneficial to look at the the family of density estimates $\hat{f}(x; h)$ that correspond to different levels of spatial resolution h. Chaudhuri and Marron (1999) describe the statistical methodology which, for a given location x and bandwidth value h, evaluates whether $E\{\hat{f}(x; h)\}$ is increasing or decreasing. The original goal of the SiZer approach was to visualize locations with the SIgnificant ZERo (SiZer) crossings of the derivative of $E\{\hat{f}(x; h)\}$, i.e. the locations x, where for a given bandwidth h the derivative of $E\{\hat{f}(x; h)\}$ changes its sign and the function $E\{\hat{f}(x;h)\}$ therefore has a local extremum. Zeng et al. (2002) suggest that SiZer maps can be used to measure the dissimilarity between distributions.

Figure 2 shows SiZer maps for the distributions of log-fluorescence for antibodies B031, B032, B034 and B036 and cell population RAJI. It illustrates how the similarities of log- fluorescence distributions (Figure 1) for antibodies binding to the same antigen (i.e B031 and B032, B034 and B036) translate into the similarities of the corresponding SiZer maps.

The horizontal axis on these images shows the values of the log-intensity, while the vertical axis shows the values of the smoothing parameter *h*. The colors on the map mark the regions where derivative $\hat{f}'(x;h)$ is significantly positive (black), significantly negative (white), or cannot be distinguished from zero (lighter gray). The regions marked by darker gray correspond to areas where the data are too sparse to make conclusions. The horizontal line on each of the maps shows the profile corresponding to the bandwidth chosen by the Sheather-Jones plug-in method (Jones, Marron, and Sheather, 1996). The locations of intensity on horizontal line $h = h_0$ preceeded by black shading (increasing function) and followed by white shading (decreasing function) contain local maxima of the function $E\{\hat{f}(x;h=h_0)\}$.

Figure 2 was obtained using software developed by J. S. Marron to accompany Chaudhuri and Marron (1999). The evaluation of significant increases and decreases of the derivative is based on the “number of independent blocks” approximation discussed by these authors, and on their default significance level of $\alpha = 0.05$. Figure 2 is an image of a $N \times M$ matrix S of ordinal values for N values of the bandwidth and M values of intensity. A dissimilarity score $d_j^{\mathrm{SZ}}(i_1, i_2)$ is defined by the proportion of locations where the maps disagree

$$d_j^{\mathrm{SZ}}(i_1, i_2) = \text{proportion of locations,} \quad \text{where } S_{i1} \neq S_{i2} \ .$$

This definition of dissimilarity is simpler than the one used by Zeng et al. (2002), who use an entropy-based measure of a dissimilarity between SiZer maps.

Plotting the stacks of SiZer maps for an antibody of interest and multiple cell populations provides a useful compliment to the display of fluorescence distributions.

2.2.5 Square-root difference dissimilarity score

Bowman and Azzalini (1996) describe use of a statistic, which in our notation, may be written as

$$t_j(i_1, i_2, x) = \sqrt{\hat{f}_{i_1 j}(x)} - \sqrt{\hat{f}_{i_2 j}(x)}\,,$$

for evaluating of the dissimilarity between the estimates of probability densities $f_{i_1 j}(x)$ and $f_{i_2 j}(x)$ at arbitrary location x. They also discuss the statistic $\int t_j(i_1, i_2, x)^2 \, \mathrm{d}x$ as a possible measure of disagreement between the distributions. We suggest a somewhat different measure

$$d_j^{\mathrm{SRD}}(i_1, i_2) = \max_x \, |t_j(i_1, i_2, x)|\,,$$

as the dissimilarity score between the distributions observed for a cell population *j*. We used the mean of the values of the smoothing parameter *h*, suggested by a normal reference density rule (Wand and Jones, 1995; Bowman and Azzalini, 1997) for individual samples in all of the comparisons.

2.3 Comparison of the dissimilarity scores

The simplest way to evaluate the identity of the CD specificities of antibodies i_1 and i_2 is to suggest that these antibodies belong to the same CD cluster if their dissimilarity score is less or equal than the threshold value $D(i_1, i_2) \leq D_T$ and otherwise belong to different CD clusters.

We compared the efficiency of this rule for different definitions of dissimilarity scores by applying it to datasets that included antibodies of known specificity. The original blind panels, which we denote

HLDA5 and HLDA6, were studied by the 5th and 6th workshops on Human Leucocyte Differentiation Antgiens. The datasets are described by Shaw et al., (1995) and Mason et al. (1997). Unfortunately, only the scaled mean values of log-fluorescence were available for the HLDA5 panel. We expanded the dataset, described by Mason et al. (1997) adding to the panel the originally collected data for the additional 15 antibodies of known specificity. We also excluded 10 "dim" antibodies from HLDA5 and 14 "dim" antibodies from the HLDA6 dataset. The antibody i was defined as "dim" if for any of the cell populations j the fraction of the positive cells α_{ij} in the panel was less than 0.5. For the HLDA5 dataset we used the estimates of the percentages of the positive cells given by Shaw et al. (1995). For the HLDA6 dataset we crudely estimated the percentage of the positive cells as the percentage of the cells that had a fluorescence intensity higher than the 95% quantile of fluorescence intensity for the unstained cell populations. We also excluded from the HLDA5-based dataset the data for cell populations that had missing log-fluorescence values for 10% or more of the cell populations in the panel and antibodies that had 10% or more of the missing mean log-fluorescence values for the selected subset of cell populations. The resulting HLDA5 subset included mean values of log-fluorescence for $p = 110$ cell populations stained by $n = 271$ antibodies from $C = 150$ CD classes. The 36385 pairs formed from these antibodies included 240 pairs of antibodies with identical specificity and 36345 pairs of antibodies with different specificity. The HLDA6 subset included the log-fluorescence samples for experiments with $p = 20$ cell populations stained by $n = 55$ antibodies of known specificity from $C = 33$ CD classes. The 1485 pairs formed from these antibodies included 38 pairs of antibodies with identical specificity and 1447 pairs of antibodies with different specificity.

For a given value of the threshold D_T we defined the sensitivity of the identity assignment rule as the proportion of pairs of antibodies with identical CD specificity that had a dissimilarity score $D(i_1, i_2)$ less or equal than D_T. We defined the specificity of the assignment as the proportion of pairs of antibodies

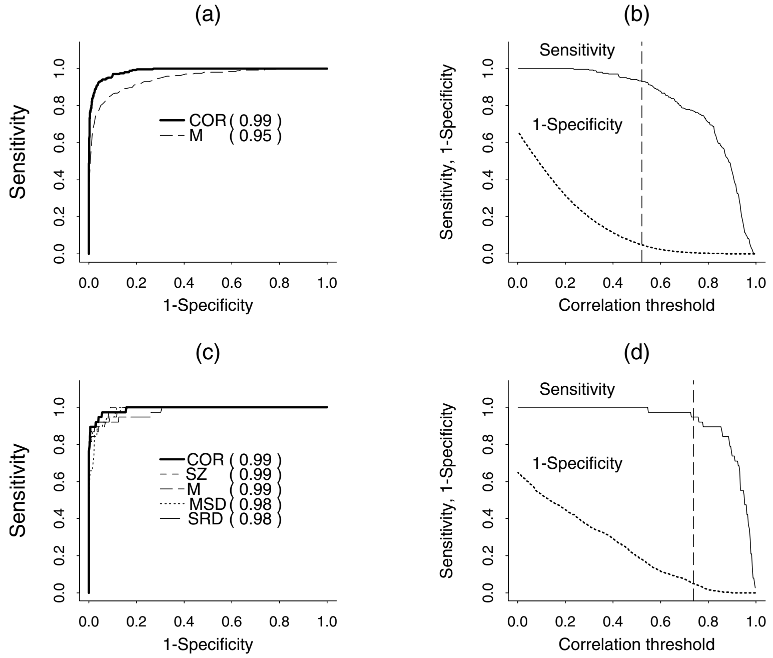

Figure 3 The analysis for the HLDA5 (upper row) and HLDA6 (lower row) datasets: ROC curves (a and c), and sensitivity/specificity of the correlation based rule (b and d). The vertical line in the second column of graphs shows the correlation threshold that leads to the specificity of 0.95.

with a different CD specificity for which the value of the dissimilarity score $D(i_1, i_2)$ exceeded the threshold. We visualized the trade-off between the sensitivity and specificity of the assignment rule by plotting the receiver operating curve (ROC), which shows the values of sensitivity and specificity for all possible values of the threshold. A discussion of the usefulness of the ROC curves for evaluating the efficiency of a classification rule may be found in Metz (1978), Hanley and McNeil (1982) and Hastie, Tibshirani and Friedman (2001). Figures 3a and 3c show the ROC curves for the HLDA5 and HLDA6 data. We were unable to calculate the MSD, SZ and SRD dissimilarities because the raw data for HLDA5 dataset were not available. The numbers in the legend are the values of the area under the ROC curve.

In both datasets we found that the correlation-based definition of dissimilarity was at least as efficient as the other four definitions. It is easily interpretable, potentially less sensitive to the systematic shift $\delta_{i2,i1}$ (as discussed in Section 2.2.2), and very similar to the definition of the dissimilarity that is successfully used in analyzing gene expression data (e.g. Paramigiani et al., 2003). We therefore recommend using the correlation-based definition of dissimilarity as the default option, checking its efficiency against the other suggested definitions of dissimilarity score by comparing the ROC curves for a subset of antibodies of known specificity in the dataset of interest.

To simplify the interpretation of the scores, we also recommend displaying the correlation scores $r(i_1, i_2)$ instead of the dissimilarity scores. The corresponding rule for evaluating the identity of CD specificities assigns identical specificity to the antibodies if their correlation score $r(i_1, i_2)$ exceeds the threshold r_T and it otherwise concludes that CD specificity of antibodies is different. The choice of the correlation threshold corresponding to the specificity of 0.95 leads to the sensitivity of 0.93 for the HLDA5 dataset and sensitivity of 0.95 for the HLDA6 dataset. Figures 3(b) and 3(d) show the values of the specificity and sensitivity resulting from the application of the correlation-based assignment rule to the HLDA5 and HLDA6 datasets.

3 Suggested Approach to a Class Prediction Analysis

We define the average dissimilarity $AD(i_1, CD_l)$ between the individual antibody i_1 that needs to be classified and antibodies of known specificities from a certain cluster CD_l as

$$AD(i_1, CD_l) = \frac{\sum\limits_{i_2 \in CD_l} D(i_1, i_2)}{N_{CD_l}} .$$

Here l is the index of the CD cluster and N_{CD_l} is the number of antibodies in the cluster. We suggest to inspect visually the similarity between the profiles $\bar{x}^a_{i_1 j}$ of the adjusted mean log-fluorescence values of the antibody i_1 with the corresponding profiles for antibodies contained in the subset of CD clusters with the m_0 smallest values of the $AD(i_1, CD_l)$ scores. Gilks and Shaw (1995) present multiple examples of similar expression profiles of the $\bar{x}_{ij}$. An inspection of the profiles may lead to the additional reduction of the number of the CD classes screened out for the visual comparisons of the sets of fluorescence distributions/SiZer maps for their antibodies with the sets of fluorescence distributions/SiZer maps for antibody i. Figure 4(b) provides an example where visual evaluation of the profiles suggests that only assignment to the cluster with the smallest $AD(i_1, CD_l)$ score is of interest.

The value m_0 may be crudely estimated by the simple "one out" analysis of the panel of n antibodies of known specificity representing C known clusters. In this analysis, each of the of the n antibodies is removed from the panel and it is assumed that correct classification of the remaining $n-1$ antibodies is known. For each of the removed antibodies i, we record whether the correct CD assignment is included in the subset of the m CD clusters with the smallest average dissimilarity scores, varying m in the range $m = 1, \ldots, C$. For each value of m, the localization error rate is defined as the proportion of the antibodies for which the correct CD assignment is not included into the set of the m most similar clusters. The value m_0 that leads to the small localization error is then conservatively chosen. We anticipate that in many cases the value $m_0 = 10$ may be used as the default.

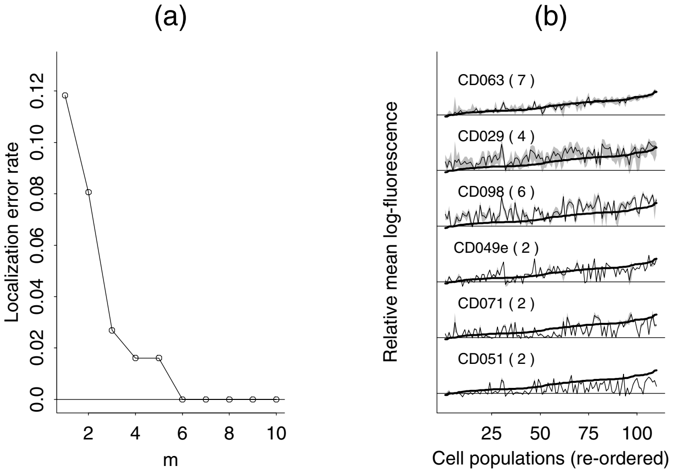

Figure 4 (a): Localization error rate for the assignment of the antibody to the set of m clusters with smallest average dissimilarity scores. (b): The profiles of $\bar{x}^a_{ij}$ for the particular antibody i_1 and antibodies from six CD clusters with the smallest $AD(i_1, CD_l)$ scores. The indexes of cell populations j are sorted by the $\bar{x}^a_{i_1j}$ values.

Figure 4(a) shows the result of the “one-out” analysis for a subset of HLDA5 data. Excluded from the subset were antibodies of unknown specificities, antibodies with “dim” fluorescence patterns as defined in Section 2.3, and samples for antibodies from CD clusters that contained less then two antibodies. The resulting dataset contained the values of the scaled mean log-fluorescence for $p = 110$ cell populations stained by $n = 168$ antibodies from $C = 65$ classes. The details of the scaling of the log-fluorescence values can be found in Gilks and Shaw (1995). For the suggested default choice of a dissimilarity measure based on the correlation scores, assignment to the “most similar” class was correct for 87.6% of antibodies and the correct CD cluster was always included in the set of $m_0 = 6$ classes. Figure 4(b) compares the profile of $\bar{x}^a_{i_1j}$ values for an antibody from the CD063 cluster with the profiles of antibodies from the $m_0 = 6$ CD clusters with the smallest $AD(i_1, CD_l)$. The thick solid line shows the $\bar{x}^a_{i_1j}$ values and the thin solid line shows the mean of the corresponding values for the antibodies within each of the particular CD clusters. The gray shading shows the region that includes the minimum and maximum values of $\bar{x}^a_{i_1j}$ within the cluster. The label above the traces contains the number of the CD cluster and number of the antibodies of known specificity representing this cluster in the dataset. The values were rescaled by multiplying all of the values reported by Gilks and Shaw by the single scaling factor.

There is a remarkable fit between the profile for the antibody of interest and the profiles of the antibodies from its cluster (CD063), while the similarity with the profiles of antibodies from the remaining clusters is noticeably weaker.

4 Suggested Approach to a Class Discovery Analysis

The panel used for the class discovery analysis contains the n_1 antibodies of known specificity and the n_2 antibodies of unknown specificity. The goal of the class discovery analysis is to identify the pairs/groups among the n_2 antibodies with unusually high similarity/low dissimilarity scores. We assume that the class discovery analysis follows the class prediction analysis and that none of the antibodies of unknown specificity belongs to any of the known CD clusters. We note that identification of a single pair (i_1, i_2) of antibodies with the identical CD specificity by analyzing the blind panel data with subsequent confirmation of the identical specificity by biochemical analysis may be sufficient for discovery of the CD. The cost of screening this pair out of the subsequent biochemical analysis is therefore very high. Any of the dissimilarity definitions discussed in Sections 2.2.1–2.2.5 may be used to detect antibodies with an unusually high similarity of fluorescence distributions. We will illustrate the approach by using the correlation-based dissimilarity score $D(i_1, i_2) = 1 - r(i_1, i_2)$, and by displaying the correlation score $r(i_1, i_2)$.

We recommend selecting the pairs (i_1, i_2) with the values of $r(i_1, i_2)$ exceeding the threshold value r_T for the subsequent visual inspection of the stacks of densities of log-fluorescence distributions/SiZer maps. The effort required for visual inspection of images similar to the ones presented on Figures 1 and 2 is negligible in comparison to the effort invested in collecting panels of flow cytometry data and the effort invested in designing and implementing the subsequent biochemical experiments. We therefore believe that setting the correlation threshold at a relatively low value is justified and suggest using the value of the 95% quantile of the distribution of the correlation score for the pairs of antibodies with different specificity as the threshold r_T.

As an example, Figure 5 presents an analysis of the panel of antibodies $(n = 95,\ p = 20)$ which includes a group of $n_1 = 25$ antibodies from known CD clusters and $n_2 = 70$ antibodies of unknown specificity. This dataset is very similar to the setting of the HLDA6 experiment as described by Mason et al. (1997). Similarly to the analysis presented in Section 3, we added 15 antibodies of known

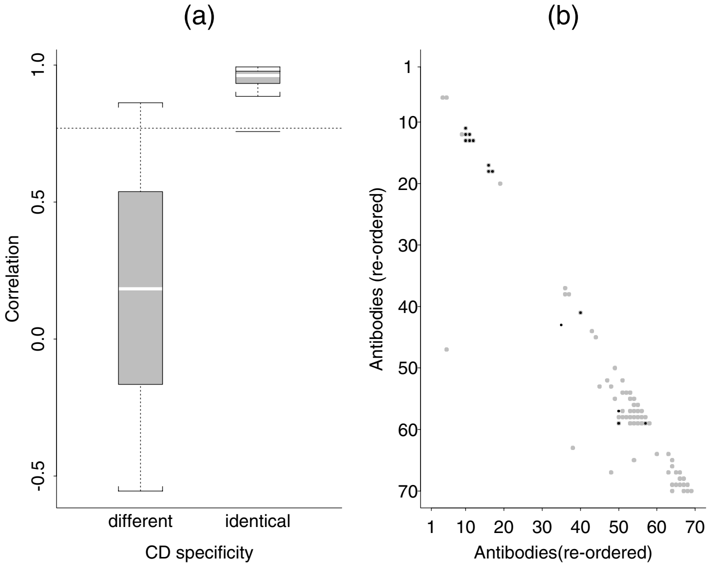

Figure 5 (a): Boxplots of the correlation score distributions for the pairs of identical and different CD specificity formed by 25 antibodies of known specificity. The horizontal line shows the suggested threshold value of the correlation $r_T = 0.77$ which corresponds to the 95% quantile of the distribution of the correlation scores among the pairs of antibodies with different specificities. (b): Correlation scores for 70 antibodies to be classified. The shaded symbols mark the pairs of antibodies with correlation scores exceeding the threshold. The dots mark pairs with biochemically confirmed identical CD specificity of antibodies. The re-ordering of antibodies is described in Section 4.

specificity to the dataset and eliminated "dim" antibodies. Figure 5(a) shows the boxplot for the distribution of correlation scores observed in a sample of 24 pairs of antibodies of identical specificity and the boxplot for distribution observed in a sample of 276 pairs of antibodies from different CDs. The horizontal line shows the suggested threshold value of the correlation $r_T = 0.77$ which corresponds to the 95% quantile of the distribution of the correlation scores among the pairs of antibodies with different specificity. The $n_2 = 70$ antibodies of unknown CD specificity form 2415 pairs which include 74 pairs with a correlation score that exceeds the threshold. The CD specificity of 29 of the 70 antibodies to be classified in our illustrative example is known. Twelve out of the 14 currently confirmed pairs of antibodies with identical CD specificity were included into the set of 74 pairs screened out by our approach for future inspection.

In our opinion clustering of antibodies may play a limited but useful role in the class discovery analysis. It is well known (e.g. Hastie, Tibshirani and Friedman, 2001) that, for an indexed set of objects with a defined pair-wise dissimilarity score, the dissimilarity between the objects with the adjacent indexes may be decreased by hierarchical clustering of the objects and re-indexing them so that their order is consistent with the order of the objects in the dendrogram. This re-ordering is not unique. The structure of the dendrogram depends on how intercluster dissimilarity is defined, with possible definitions including single-linkage, complete linkage and group average (e.g. Kaufman and Rousseeuw, 1990). In addition, for any given dendrogram and n objects there are 2^{n-1} possible orderings of the objects consistent with the structure of a tree. This non-uniqueness of the ordering is unimportant for our purposes. However, for the unique definition of the suggested approach, we recommend using the group average linkage method and the default ordering of the elements in a dendrogram used in the S-plus package. For this ordering, the subtree with the tighter cluster is placed to the left at each merge. We recommend re-ordering antibodies of unknown specificity in this manner and displaying the map of the re-ordered matrix of the correlation scores. The pairs of the correlation scores with the values exceeding the threshold are then highlighted. Figure 5(b) shows the map of the lower diagonal portion of the symmetric matrix of correlation scores for 70 antibodies of unknown specificity from our illustrative dataset. The gray shading marks the pairs of antibodies with correlation scores exceeding the $r_T = 0.77$ threshold. The dots mark the pairs of antibodies with the confirmed identity of their CD specificities. In this example the thresholding reduced the number of pairs to be investigated by a factor of 30 and clustering-based re-ordering provided a convenient grouping of the 74 selected pairs of antibodies for subsequent visualization.

5 Conclusion

We have discussed an important biomedical problem of classifying newly discovered monoclonal antibodies based on panels of flow cytometry data (blind panels), where the objects of classification are characterized by a set of samples of univariate iid observations. To the best of our knowledge the problem has attracted only limited attention from statisticians (Spiegelhalter and Gilks, 1987; Gilks, Oldfield and Ratherford, 1989; Gilks and Shaw, 1995).

We suggest an approach to analyzing blind panel data that is similar in spirit to previously used approaches but differs in important implementation details. We demonstrated the usefulness of separating the analysis of the blind panel data into class prediction and class discovery, suggested several new definitions of dissimilarity scores that can be used for quantifying the pair-wise dissimilarity between antibodies, and recommended a flexible semi-automatic approach to classification analysis. In this approach the automatic algorithm is used as a screening tool that suggests to the analyst the subset of the "most appropriate" classifications of antibody in class prediction analysis, or the "most similar" pairs/groups of the antibodies in class discovery analysis. It focuses the attention of the analyst on the visual analysis of the raw data for the subset of the most relevant objects. We also emphasized the usefulness in this analysis of the statistical methodology (SiZer maps) developed by Chaidhurri and Marron (1999).

We demonstrated the potential usefulness of the simplest approaches to the class prediction and class discovery analysis by illustrative analysis of several blind panels. The suggested algorithm speeds up the analysis of a flow cytometry data by a factor of 10–20. Class prediction and class discovery analysis of blind panel have many unique features, such as a large number of possible classes within the panels, that contain a small number of objects (e.g. 2–10) per class, the moderate length ($p = 20 - 100$) of the list of features describing an individual object, and a complex characterization of the individual features by kernel density estimates of fluorescence distributions or SiZer maps. We therefore anticipate that many different approaches to the development of an optimal statistical methodology for the design and analysis of the blind panel experiments will be developed and tested in the future.

We also expect greater involvement from statisticians in designing and analyzing flow cytometry experiments. Interested readers are referred to the classic book on flow cytometry (Shapiro, 2003), several papers written by statisticians (Eudey, 1996; Baggerly, 2001) and statistical software developed by A. Rosini and his group (http://www.analytics.washington.edu/downloads/rflowcyt) for the Bioconductor project.

Acknowledgements The authors are grateful to Steven Mentzer, Howard Shapiro, and Heddy Zola for stimulating discussions on the biological foundations of the blind panel experiments. We also acknowledge Donald Halstead, Inge Koch, Peter Macardle, James Rawn, Mario Roederer, Bernadette Swart, Kathryn Tinckam, Qing Zeng, Heddy Zola, Jim Ware and the anonymous referees for their useful comments and suggestions. The work was partially supported by the grant from the Flow Cytometry Facility, Renal Division, Brigham and Women's Hospital and the grant from the University of New South Wales Faculty Research Grants Program.

References

Baggerly, K. A. (2001). Probability binning and testing agreement between multivariate immunofluorescence histograms: extending the Chi-Squared test. *Cytometry* **45**, 141--150.

Bowman, A. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis.* Clarendon Press, Oxford.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**, 807--823.

Eudey, T. L. (1996). Statistical considerations in DNA flow cytometry. *Statistical Science* **11**, 320--334.

Gilks, W. R., Oldfield L., and Rutherford, A. (1989). Statistical analysis. *Leucocyte Typing IV* (eds. Knapp W. *et al).* Oxford University Press, Oxford, 6--11.

Gilks, W. R. and Shaw, S. (1995) Statistical analysis. *Leucocyte Typing V* (eds. Schlossman S. *et al.).* Oxford University Press, Oxford, 8–13.

Givan, A. L. (2001). *Flow Cytometry: First Principles.* Wiley-Liss, New York.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P. Coller, H., Loh, M. L., Downing, J. R., Caliguri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene monitoring. *Science* **286**, 531--537.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29--36.

Hallam, J., Ying Z., Gail, H. M., and Shaw, S. (1997). T-cell blind panel: evolving strategy for finding similarities between monoclonal antibodies on the basis of flow cytometric analysis of multiple cell subsets in peripheral blood. *Leucocyte Typing VI* (eds. Kishimoto, T. *et al.).* Garland Publishing, Inc., New York, 82--85.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Chapman and Hall, London.

Hilgert, I. and Drbal, K. (2002). Non-lineage panel-analysis by cytofluorometry. *Leucocyte Typing VI* (eds. Mason, D. *et al).* Oxford University Press, Oxford, 459--462.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**, 401--407.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 44--45.

Kishimoto, T., Kikutani, H., von dem Borne, A. E. G. Kr., Goyert, S. M., Mason, D. Y., Miyasaka, M., Moretta, L., Okumura, K., Shaw S., Springer, T. A., Sugamura, K., and Zola, H. (eds) (1997). *Leucocyte Typing VI.* Garland Publishing, Inc., New York.

Lampariello, F. and Aiello, A. (1998). Complete mathematical modeling method for the analysis of immunofluorescence distributions composed of negative and weekly positive cells. *Cytometry* **32**, 241--254.

Mason, D. Y., Jones, M., Hardie, D. L., Schindel, G. V., Johnson, G. D., van Lier, R., and MacLennan, I. C. M. (1997). Blind panel report. *Leucocyte Typing VI* (eds. Kishimoto, T. *et al.).* Garland Publishing, Inc., New York, 206--229.

Mason D., Andre P., Benussian, A., Buckley, C., Civin, C., Clark, E., de Haas, M., Goyert, S., Hadam, M., Hart, D., Horesi, V., Jones, Y., Meuer, S., Morissey, J., Schwartz-Albiez, R., Shaw, S., Simmons, D., Turni, L., Uguccioni, M., van der Schoot, Vivier E., and Zola, H. (eds) (2002). *Leucocyte Typing VI*. Oxford University Press, Oxford.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**, 283--298.

Miyazaki, S., Sugawara, H., Tamuro, T., Okayama, T., Ishi, I., Shimura, J., Yoshida, K., Saijo, K., Ohno, T., Blanchard, D., Bouquet, N. V., Bochner, B., Buhring, H.-J., Xheng, Z., Goyert, S., Henniker, A., Hirano, T., Horesi, V., Itoh, K., Kanakura, Y., Matsuo, Y., Miyasaka, M., Muraguchi, A., Okumura, K., Sakaguchi, N., Saitoh, H., Springer, T. A., Sugamura, K. and Kikutani, H. (1997) Cross-lineage (blind panel) study and human leucocyte differentiation database. *Leucocyte Typing VI* (eds. Kishimoto, T. *et al).* Garland Publishing, Inc., New York, 3--20.

Paramigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L. (eds) (1990). *The Analysis of Gene Expression Data. Methods and Software*. Springer-Verlag New York Inc., New York.

Schlossman, S. F., Boumsell, L., Gilks, W., Harlan, J. M., Kishimoto, T., Morimoto, C., Ritz, J., Shaw, S., Silverstein, R., Springer, T., Tedder, T. F., and Todd, R. F. (eds) (1995). *Leucocyte Typing V.* Oxford University Press, Oxford.

Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization.* John Wiley, New York.

Shapiro, H. M. (2003). *Practical Flow Cytometry.* John Wiley, New Jersey.

Shaw, S., Luce, G. G., Gilks, W. R., Anderson, K., Ault, K., Bochner, B., Boumsell, L., Denning, S. M., Engleman, E. G., Fleisher, T., Freedman, A. S., Fox, D. A., Gailit, J., Guttieres-Ramos, J. C., Hurtubise, P. E., Lansdorp, P., Lotze, M. T., Mawhorter, S., Marti, G., Matsuo, Y., Minowada, J., Michelson, A., Picker, L., Ritz, J., Roos, E., van der Schoot, C. E., Springer, T. A., Tedder, T. F., Telen, M. J. Thompson, J. S., and Valent, P. (1995). Leucocyte differentiation antigen database. *Leucocyte Typing V* (eds. Schlossman, S. *et al.*) Oxford University Press, Oxford, 8--13.

Spiegelhalter, D. J., and Gilks, W. R. (1987). Statistical analysis. *Leucocyte Typing I* (eds. McMichael, A. J. *et al.*). Oxford University Press., Oxford, 14—24.

Stuart, C., and Nicholson, J. K. A. (eds) (1995). *Immunophenotyping*. John Wiley and Sons, Inc., New York.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hill, London,

Watson, J. V. (1998). Proof without prejudice revisited: immunofluorescence histogram analysis using cumulative frequency subtraction plus ratio analysis of means. *Cytometry* **43**, 55--68.

Zola, H. and Swart, B. (1998). Human leucocyte differentiation antigens. *Trends in Immunology* **24**, 353--354.

Zeng, Q., Wand, M., Young, A., Rawn, J., Milford, E. L., Mentzer, S. J., and Greenes, R. A. (2002). Matching flow-cytometry histograms using information theory in feature space. *Proceedings, American Medical Informatics Association, Annual Fall Symposium;* 929--933.