Cytometry
PART A
Journal of the International Society for Advancement of Cytometry

Automation in High-Content Flow Cytometry Screening

U. Naumann,[1] M. P. Wand[2]*

[1]School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia

[2]School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia

Received 15 December 2008; Revision Received 26 March 2009; Accepted 15 May 2009

Grant sponsor: Australian Research Council Discovery Project; Grant number: DP0556518.

*Correspondence to: M.P. Wand, School of Mathematics and Applied Statistics, University of Wollongong, Northfields Avenue, Wollongong 2522, Australia.

Email: mwand@uow.edu.au

Published online 22 June 2009 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/cyto.a.20754

Abstract

High-content flow cytometric screening (FC-HCS) is a 21st Century technology that combines robotic fluid handling, flow cytometric instrumentation, and bioinformatics software, so that relatively large numbers of flow cytometric samples can be processed and analysed in a short period of time. We revisit a recent application of FC-HCS to the problem of cellular signature definition for acute graft-versus-host-disease. Our focus is on automation of the data processing steps using recent advances in statistical methodology. We demonstrate that effective results, on par with those obtained via manual processing, can be achieved using our automatic techniques. Such automation of FC-HCS has the potential to drastically improve diagnosis and biomarker identification. © 2009 International Society for Advancement of Cytometry

Key terms

density curvature; high throughput flow cytometry; highest density region; penalised splines; subject specific curves

THE last few years have seen a major change in flow cytometry technology, toward what has become known as high-throughput flow cytometry or high-content flow cytometric screening (FC-HCS) (e.g., 1,2). This 21st Century technology combines robotic fluid handling, flow cytometric instrumentation, and bioinformatics software so that relatively large numbers of flow cytometric samples can be processed and analyzed in a short period of time. A pioneering article on FC-HCS by (1) closes with: "Further improvements that completely automate the FC-HCS procedures and incorporate newly developed advanced data analysis and management features will further improve the efficiency and power of this technique."

The thrust of the present article is to demonstrate how advanced statistical techniques can help automate the processing of FC-HCS data. We draw upon recent research on feature significance for multivariate density estimation (3,4) which has led to an automatic gating procedure that we call `curvHDR` (Naumann, Luta, Wand, "The curvHDR method for gating flow cytometry samples", unpublished). We also make use of recent developments in semiparametric regression for grouped data (5). The end result is an algorithm that processes FC-HCS data with only a small amount of human involvement.

Although `curvHDR` is in its infancy as an alternative automatic gating method, we believe that it has great potential for aiding the processing of FC-HCS data. It is an intuitive method that aims to mimic human perception in terms of possibly relevant gates. In this article, its use is limited to univariate and bivariate data. However, there is no firm upper limit on the dimensionality and sample size for which `curvHDR` can be used. In (Naumann et al., unpublished), we describe implementational details for gating trivariate flow cytometric data. The regression models of (5) allow for nonlinear time effects to be handled, while also taking account of within-patient correlation.

Recently, (6) conducted a sophisticated analysis of FC-HCS data in an effort to identify cellular signatures for acute graft-versus-host-disease (GvHD). The data were

ISAC
International Society for Advancement of Cytometry

obtained from blood samples of 31 patients undergoing blood and marrow transplant. The samples were screened using 10 different four-color antibody combinations over several days. Their data analytic methods involved two main phases: (a) a multistep manual gating strategy involving both univariate and bivariate flow cytometry data, and (b) spline-based regression analysis applied to the proportions of cells in the final gates. Through, for example, their Figure 3, they demonstrate differences between GvHD and control patients in terms of the longitudinal paths of proportion of $CD3^{+}CD4^{+}CD8\beta^{+}$ cells (i.e., cells with high fluorescence corresponding to the binding activity of antibodies CD3, CD4, and CD8β.).

Although the results of (6) are very encouraging, their production requires a significant amount of human labour and judgement. Working with the same data, we develop a statistically based processing algorithm that aims to mimic the manual analyses of (6). Except for the choice of a small number of tuning parameters, the algorithm takes the raw FC-HCS data and automatically produces cellular signatures similar to those given in their Figure 3. Expert knowledge can be used to tune the parameters. The post-tuning automation aspect means that FC-HCS data can be obtained very quickly, and possibly more objectively. Our algorithm has great promise for future FC-HCS data analyses and has the potential to drastically improve diagnosis and biomarker identification.

Automatic statistical methods for high-content flow cytometry data is an emerging area of research. A recent contribution of which we are aware is (7). These authors develop a mixture model approach to automatic gating and also apply it to the data of (6).

The section entitled Data describes the (6) data. Our methods for their automatic processing are described in the Methods section. Results are given in the Results section. We close with some discussion in the section entitled Discussion.

Data

The data correspond to 31 patients who were undergoing blood and marrow transplantation at the Moffitt Cancer Center, Vancouver, Canada. Full details regarding patient demographics and the methods used to obtain the data are given in (6). In this section, we describe the most essential aspects of the data.

For each patient, blood samples were taken over a period of several weeks. The first sample was usually taken a few days before the transplantation, the second one on the day of the transplant and the rest of the samples were taken after the transplant with $\sim$7 days between each new measurement. Each of the blood samples was aliquotted into 96-well plates. The 96-well plates were stained with 10 different four-color antibody combinations.

Following transplant, 21 patients were diagnosed with acute GvHD while three patients did not develop GvHD and are used as controls. Seven of the GvHD patients were not included in the (6) analysis due to either being lost to follow-up, insufficient clinical samples, the development of de novo chronic GvHD, which may confound the analysis for acute GvHD, or because they died within 100 days of their transplant and it was not possible to determine whether or not they would have subsequently developed GvHD.

Table 1. Antibody combinations from (6) used in current study

ANTIBODY COMB. NO.	ANTIBODY 1 (AB1)	ANTIBODY 2 (AB2)	ANTIBODY 3 (AB3)
1	CD15	CD45	CD14
2	CD4	CD8β	CD3
3	CD16	CD2	CD3
4	CD10	CD20	CD19
5	TCR$\alpha\beta$	TCR$\gamma\delta$	CD3
6	CD44	CD25	CD3
7	CD4	CD134	CD3
8	CD4	CD122	CD3
9	CD45RA	CD45RO	CD3

This is an abbreviated version of Table 2 of (6).

For each of the 24 patients included in the analysis (6), the data takes the form of 10 time series of matrices — one for each antibody combination. Each matrix has six columns and, on average, about 16,000 rows. Each row represents a different cell. The columns correspond to: forward scatter, side scatter, and fluorescence measurements for each of the four antibodies. Also, we note that a quality assessment has been performed on these data and is summarized in (2).

The (6) analysis mainly restricted attention to the first three antibodies in each combination. This article restricts attention in this way. Because antibody combinations 9 and 10 differ only in their fourth antibody, we dropped the latter of these. Table 1 lists the 9 antibody combinations that we analyze. Additional details on these antibodies are given in (6).

Figure 1 illustrates the nature of the data by plotting a subset. The upper two rows are cellular fluorescence measurements, corresponding to antibodies CD4 and CD8β, after gating on CD3-positive cells, on patients that develop GvHD, whereas the lower two are controls.

Methods

Our methodology for obtaining cellular signatures of GvHD involves three phases:

- *Cleaning and Transformation Phase*
  This involves: (a) removing recordings that accumulate on the boundaries (usually upper boundaries) of the flow cytometry samples; (b) possibly transforming samples to reduce their skewness.
- *Longitudinal Data Creation Phase*
  This phase involves setting gates using data collected from patients near the start of their involvement in the study; and then collecting the proportion of cells falling into those gates longitudinally.
- *Penalized Spline Regression Phase*
  In this phase, flexible regression models are fitted to the longitudinal data. The fits provide a cellular signature for GvHD based on the particular antibody combination used to create the data.

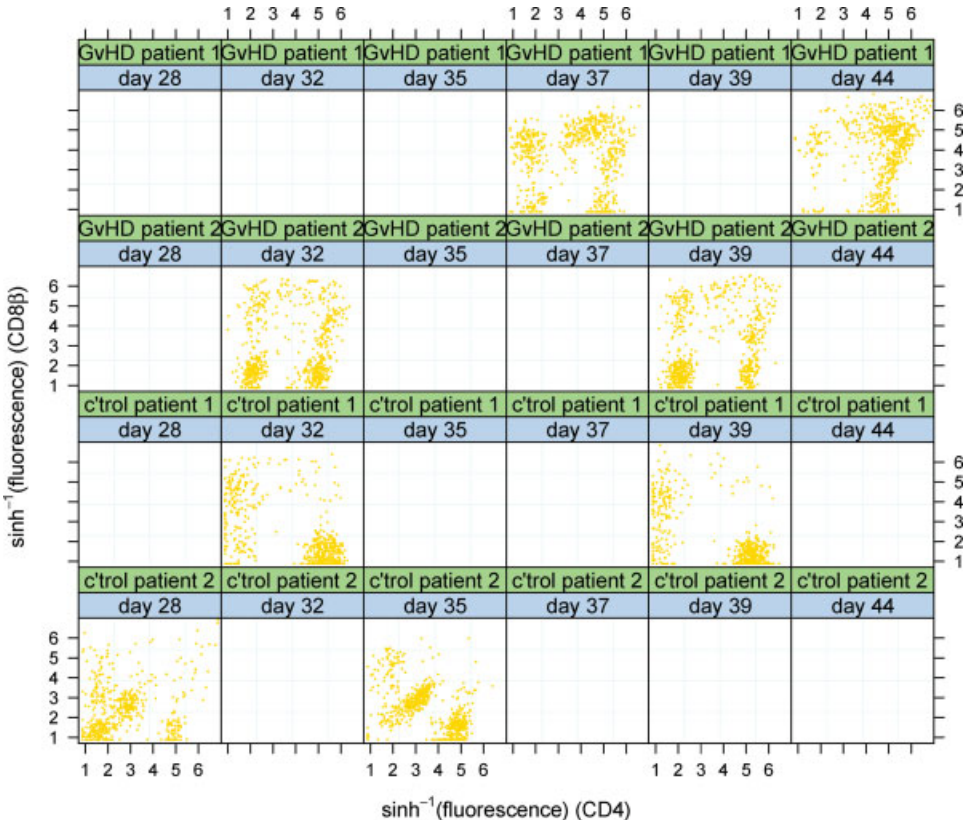

Figure 1. Portion of the GvHD data of (6). The upper two rows are cellular fluorescence measurements, corresponding to antibodies CD4 and CD8β, after gating on CD3-positive cells, on patients that develop GvHD, while the lower two are controls. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

The longitudinal data creation phase is by far the most intensive in terms of human effort. For the present analysis, 69 gates need to be determined. The section entitled Automatic Gating via curvHDR describes automatic gating strategies that rectify this situation and allow the entire analysis to be run on a standard computer in less than one hour. Details on the full automatic spline regression phase are given in the section Penalized Spline Regression Analysis. The Algorithm section summarizes the entire process.

Automatic Gating via `curvHDR`

An integral component of flow cytometric data analysis is gating, where the cells are subsetted according to physical and fluorescence measurements. The most important components are (i) bivariate cell-type gating (e.g., identification of lymphocytes from scatterplots of forward-scatter versus side-scatter measurements) and (ii) univariate fluorescence-channel gating (e.g., identification of cells that recognise a particular antibody). The analysis in (6) involved manual gating of hundreds of flow cytometric samples.

We automate the gating process via our recently developed `curvHDR` method (Naumann et al., unpublished). This method combines the ideas of significant density curvature and highest density region estimation in an attempt to mimic human-based gating, with a minimal number of tuning parameters. In principle, the `curvHDR` method applies to data of arbitrary dimension, although in (Naumann et al., unpublished), the full details of the method are worked out for univariate, bivariate, and trivariate flow cytometry data sets. The method has no problems with very large sample sizes that typically arise in flow cytometric data analysis. This paper uses only the univariate and bivariate versions. Bivariate `curvHDR` is the easiest to explain graphically, so we will first give some details on this case. An appendix provides full details on `curvHDR` for arbitrary dimension.

Bivariate `curvHDR`. Figure 2 is from (Naumann et al., unpublished) and graphically describes the bivariate version of `curvHDR`. Full details are given in that reference.

There are several tweaking factors in `curvHDR` gating. However, the main ones are

τ = number between 0 and 1 that specifies the particular highest density region,

G_2 = growth factor for bivariate high curvature polygons.

For a density function f on the real number plane, the τ highest density region is the region nearest the peak for which the volume under f is $1 - \tau$. For example, the highest density region in Figure 2(f) with $\tau = 0.1$ is such that 90% of the

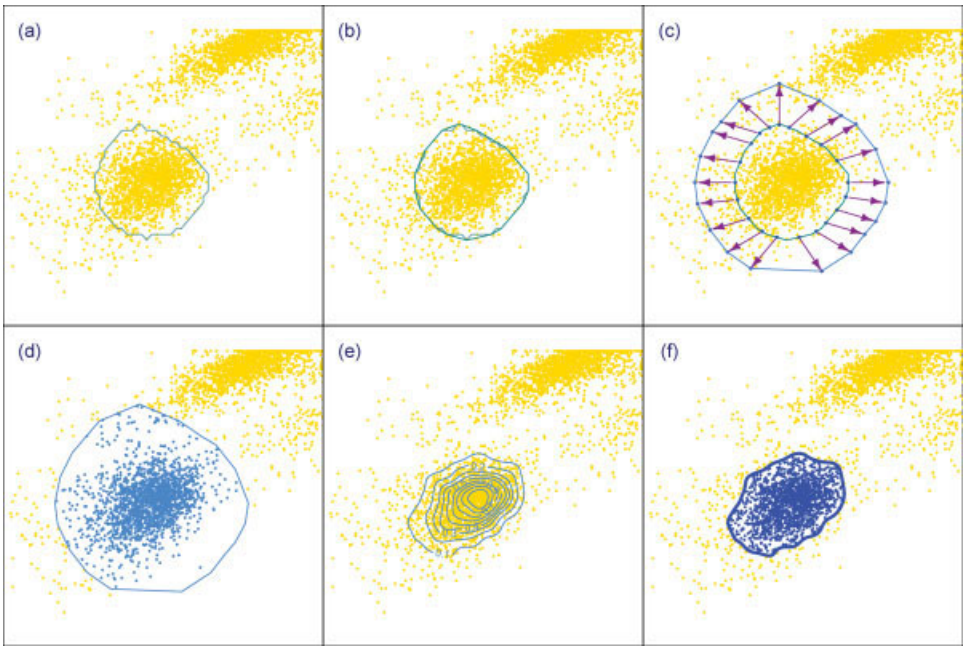

Figure 2. Graphical illustration of `curvHDR` gating for bivariate data. Panel (**a**): Polygon corresponding to a region of statistically significant high negative curvature. Panel (**b**): The convex hull of the polygon from (a). Panel (**c**): A new, larger, polygonal region obtained by growing the region from (b) using the notion of ''circle rolling'' around the inner polygon. Approximate circle rolling is achieved by taking normal vectors of equal length from the centre of each edge of the inner polygon. The size of the outer polygon is chosen so that the ratio of its area to the inner polygon is a prespecified growth factor G_2. Panel (**d**): The bivariate measurements are subsetted according to inclusion inside the polygon from (c). Panel (**e**): A kernel density estimate is obtained using only the subsetted data from (d). Panel (**f**): The final gate corresponds to a high density region contour of the kernel density estimate from (e), in this case the $\tau = 0.1$ highest density region. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

probability mass (for data in the grown polygon of Figure 2(d)) is within this highest density region.

The growth factor is simply

$$G_2 = \frac{\text{Area of grown polygon}}{\text{Area of high curvature polygon}}.$$

In our algorithm for processing the GvHD data, we allow for τ and G_2 to be specified by the user. All other `curvHDR` parameters are set to defaults.

Often `curvHDR` gating will lead to several separate regions within the bivariate region covered by the data. However, only one or two of these will be of practical interest and further gating is usually required. An effective means of doing this is via intersection with a rectangle. We call the resulting region a `rectangle-curvHDR` gate. Figure 3 provides graphical description of `rectangle-curvHDR` gating.

Obviously, `rectangle-curvHDR` gating can be extended to the situation where there is more than one rectangle involved. However, for the analysis in this paper, we only use single rectangles. This has the advantage of requiring only four additional parameters to go from a `curvHDR` to `rectangle-curvHDR`; namely, those defining the rectangle limits. Effective choice of the rectangle parameters often can be made via visual inspection. In the analyses presented here, we choose the rectangular parameters to (a) to give good separation between bright and dull fluorescence, and (b) to obtain results similar to those of Brinkman et al. (2007). For other analyses of this type, we believe that the cytometrists would have a good sense of how the rectangular parameters should be set.

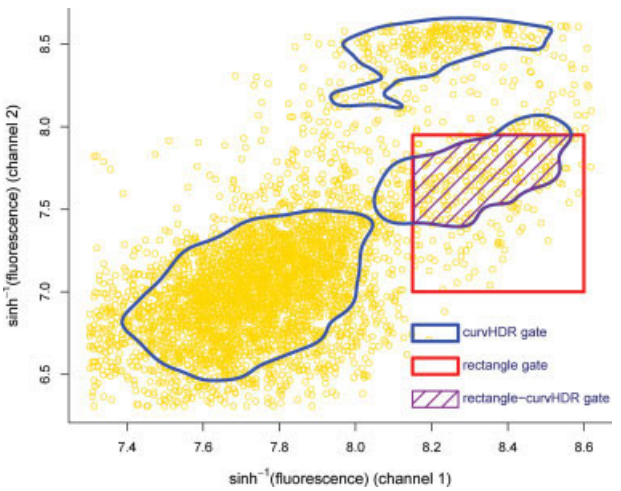

Figure 3. Graphical description of `rectangle-curvHDR` gating for bivariate flow cytometry data. The irregular shapes correspond to a `curvHDR` gate. The `rectangle-curvHDR` is the intersection between these shapes and the rectangle, and is shown using cross-hatching. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

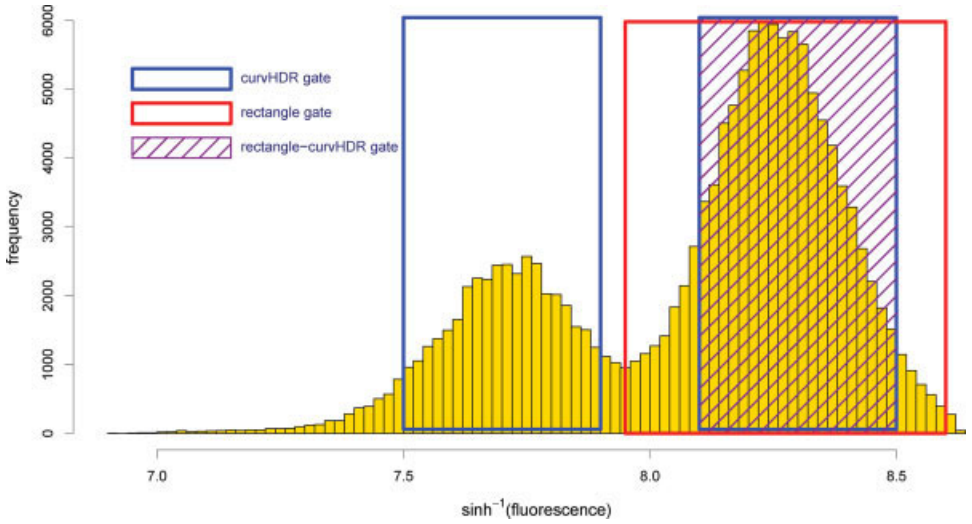

Figure 4. Graphical description of `rectangle-curvHDR` gating for univariate flow cytometry data. Intervals correspond to the base of each box. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Univariate `curvHDR`. In principle, both `curvHDR` and `rectangle-curvHDR` gating apply to data of any dimension. For the latter, note that we are using the word "rectangle" to mean linear bounds parallel to the coordinate axes. In the univariate case, rectangular gates correspond to intervals on the real line. Figure 4 illustrates `rectangle-curvHDR` gating for univariate data. Here the data are shown in histogram form, and intervals are shown using boxes to aid visualization. The `curvHDR` gate corresponds to the two intervals on which the blue boxes are based. The "rectangle" gate corresponds to the base of the red box. The `rectangle-curvHDR` is the intersection of these two regions and is shown by the purple cross-hatched box.

To distinguish the growth factor for univariate `curvHDR` gating from its bivariate counterpart, we define

$$G_1 = \text{growth factor for univariate high curvature intervals} = \frac{\text{Length of each growth interval}}{\text{Length of each high curvature interval}}.$$

Penalized Spline Regression Analysis

Figure 5 shows a typical data set arising from our `rectangle-curvHDR`-based gating strategy. The Algorithm section provides full detail on how these data are obtained. In this section, we describe how penalized spline regression analysis

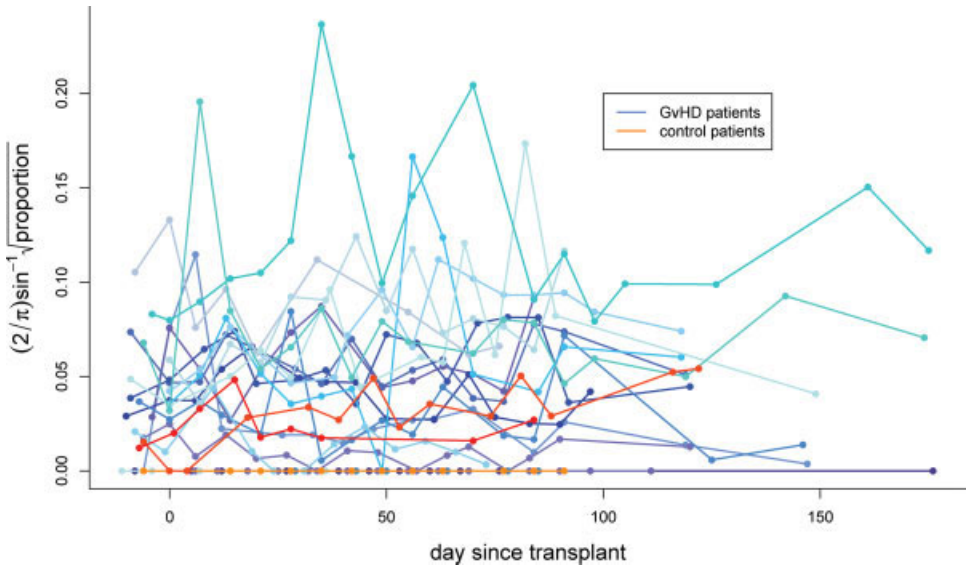

Figure 5. Typical data set arising from the longitudinal data creation phase. These data are for Antibody Combination 2. The response variable is the variance stabilised transformation of the proportion of gated cells. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

can be used to provide a statistically valid signature for GvHD.

Let proportion_{ij} denote the proportion of gated cells for patient i ($1 \leq i \leq 24$) based on data for day j ($1 \leq j \leq n_i$, where n_i is the number of days for which data are recorded for patient i.). The transformation

$$y_{\text{new}} = (2/\pi)\sin^{-1}(\sqrt{y})$$

has variance stabilizing properties for proportion data (e.g., 8), so we work with the proportions after undergoing this transformation.

An appropriate model for the data exhibited in Figure 5 is the subject specific curves model mean

$$\left\{(2/\pi)\sin^{-1}\left(\sqrt{\text{proportion}_{ij}}\right)\right\} = f_{\text{control}}(\text{day}_{ij}) + G_i \times c(\text{day}_{ij}) + g_i(\text{day}_{ij}) \text{ (Model A)}$$

where $G_i \in \{0,1\}$ is an indicator of the ith patient being diagnosed with GvHD. Here $f_{\text{control}}(\text{day})$ is the mean transformed proportion for control patients, $f_{\text{control}}(\text{day}) + c(\text{day})$ is the mean transformed proportion for GvHD patients and g_i is a subject specific deviation. For identifiability, the g_i are subject to constraints such as an average value of zero.

Reference (5) describe, automatic fitting of Model A via penalized splines and linear mixed model software. The smoothing parameters are chosen via restricted maximum likelihood. The cellular signature is the estimated contrast curve

$$\hat{c}(\text{day}) \text{ versus day}$$

along with ± 2 estimated standard error curves; which represent pointwise $\sim$95% confidence regions.

As explained in (5), several earlier papers [e.g., (9)] develop models that are similar to Model A. An advantage of the (5) approach is that it is readily implemented in standard software. The Software Issues section provides some further details.

Algorithm

We are now in a position to describe our full algorithm for processing the FC-HCS data of (6). It requires manual choice of three rectangles for each antibody combination, as well as values for `curvHDR` tuning parameters τ, G_1, and G_2. Effective outputs were obtained with global specification of τ, G_1, and G_2. In keeping with (6), we used data for the day of diagnosis to set the rectangular gates. Apart from these few choices, the algorithm is automatic and can be used to process FC-HCS data with relatively little human involvement.

The first phase involves cleaning and transformation of the flow cytometry data. The cleaning phase involves removal of the significant proportion of "debris cell" observations that pile up at the boundaries of the data. Leaving these in could have an adverse affect on our automatic gating strategy. In a similar vein, because the antibody fluorescence measurements are often highly skewed, the log or $\sinh^{-1}$ transformation leads to better automatic gating via `curvHDR`. We have a slight preference for $\sinh^{-1}$ because it handles values near zero better. Away from zero the two transformations behave similarly. Although such transformation is necessary for the FC-HCS data of (6), it may not always be required; particularly when high skewness is not present. In general, preliminary data analytic checks should be used to decide whether transformations are desirable.

Cleaning and Transformation Phase

1. Determine the minima and maxima of each flow cytometry sample. Remove all values in the sample that equal the minima or maxima.
2. Transform the antibody fluorescence measurements using $x_{\text{new}} = \sinh^{-1}(x) = \log(x + \sqrt{x^2+1})$.

Longitudinal Data Creation Phase

Set trial values of the τ, G_1, and G_2 parameters.

For each `Antibody Combination`:

1. Set trial values for the rectangular parameters via visual inspection for a few GvHD patients at approximate day of diagnosis.
2. For each patient:
   a. If patient developed GvHD then determine the approximate day of diagnosis. If patient did not develop GvHD then use day closest to 35 (average diagnosis day for GvHD patients). Using data for this day:
      i. Set a `rectangle-curvHDR` gate on `(FSC, SSC)`; denoted by `GATE(FSC,SSC)`.
      ii. Set `rectangle-curvHDR` gate on `Ab3`; denoted by `GATE(Ab3)`.
      iii. Set a `rectangle-curvHDR` gate on `(Ab1, Ab2)`; denoted by `GATE(Ab1,Ab2)`.
      iv. Note that (i) and (ii) involve `curvHDR` applied to the full data. (iii) usually involves `curvHDR` applied to data after passing through `GATE(FSC,SSC)` followed by `GATE(Ab3)`. If there were no cells that passed through `GATE(FSC,SSC)` followed by `GATE(Ab3)` then (iii) was applied to the cells that passed through `GATE(FSC,SSC)`.
   b. For each day number:
      i. Determine the `proportion` of cells that pass through `GATE(FSC,SSC)`, followed by `GATE(Ab3)`, followed by `GATE(Ab1,Ab2)`. The proportion is defined as

$$\text{Proportion} = \frac{\text{Number of cells after all gating}}{\text{Total number of cells}}$$

      ii. Obtain the longitudinal responses to be the variance stabilizing transformed responses:

$$(2/\pi)\sin^{-1}(\sqrt{\text{proportion}}).$$

   c. Adjust the rectangular parameters via visual inspection of the gates across several patients days and then repeat Step 2.

Adjust τ, G_1, and G_2 via visual inspection of the longitudinal data. For this study, these were chosen to be $\tau = 0.01$, $G_1 = 5$, and $G_2 = 10$; which gave longitudinal data similar to that obtained by (6) for `Antibody Combination 2`.

Penalized Spline Regression Phase

The data collection phase produces a set (one for each `Antibody Combination`) of longitudinal paths of proportion values for each patient, as well as labels denoting the patient as GvHD or control. Fit the (5) contrast model – Model A.

Software Issues

The entire analysis was performed within the R computing environment (10), and packages from `Bioconductor` (11). The `curvHDR` procedure is about to be incorporated into the `Bioconductor` package `flowCore` (12). The spline regression phase uses the linear mixed models function `lme()` in the R package `nlme` (13). (5) provides illustratory code for models of this type.

Results

Application of the longitudinal data creation phase of the algorithm described in the Algorithm section led to Figure 6. The panels correspond to antibody combinations. The GvHD patients are shown in shades of blue; shades of orange are used for the control patients.

Figure 7 shows the estimated contrast functions, c_i(day), after feeding the data from Figure 6 into the spline regression model, Model A. The shaded regions about the curves correspond to approximate pointwise 95% confidence intervals. For all antibody combinations, except for number 6, the contrast curves deviate significantly from zero for most of the longitudinal range. Antibody Combinations 1–5 and 9 have positive differences in mean transformed proportion values. The differences are negative for Antibody Combinations 7 and 8.

The properties of the estimated contrast curves provide some insight into how the difference in proportions changes as the disease progresses. For Antibody Combinations 7 and 8, the change in the difference is quite pronounced. The time effect is fairly flat for Antibody Combinations 4 and 9. Some interesting nonlinear contrast effects are apparent for Antibody Combinations 3 and 5. For Antibody Combination 3, there is is a pronounced peak in the proportions for GvHD patients, when compared with those of the control patients, about 1 months after transplant.

Discussion

We have demonstrated that recent developments in methodological statistics, multivariate density feature significance, and mixed model-based spline regression for longitudinal data can aid automation in high-content flow cytometry data analy-

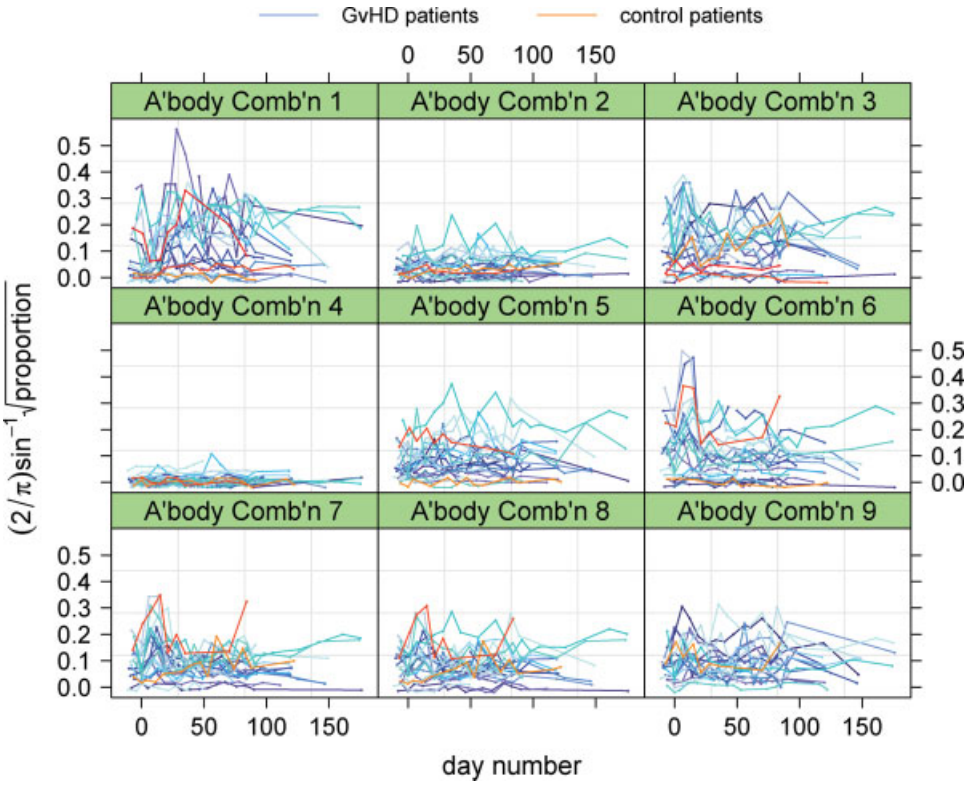

Figure 6. The data obtained from application of the longitudinal data creation phase of the algorithm given in the Algorithm section. Some jittering has been applied to the response values to aid visualisation. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

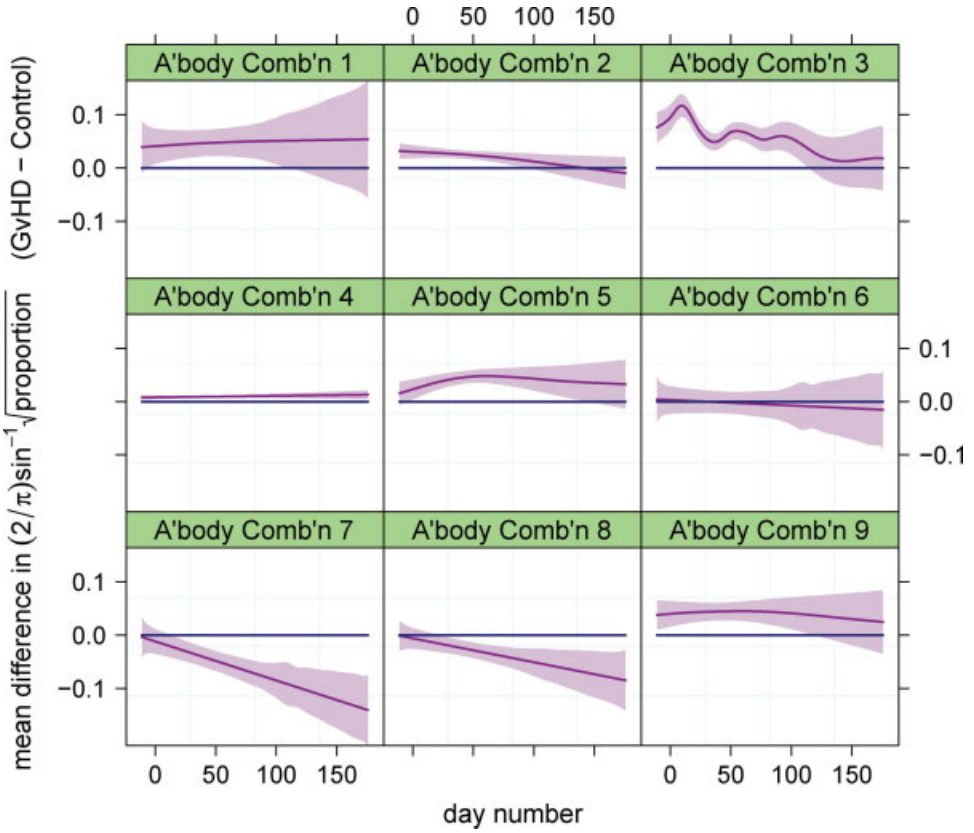

Figure 7. Estimated contrast curves (cellular signatures) arising from fitting Model A to the longitudinal data displayed in Figure 6. The shading around each curve corresponds to approximate pointwise 95% confidence intervals. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

sis. Viable cellular signatures can be obtained with a minimal amount of human labor, with run times in the tens of minutes on standard 2008 computers. Our gating routines are in the process of being made available within Bioconductor, and soon the full algorithm will be implementable in public domain packages in the `R` environment. This will facilitate application, assessment and honing of the algorithm for future analyses.

Acknowledgments

This research has benefited from discussions with Robert Gentleman. We are grateful to Nolwenn Le Meur for assistance with acquisition of the GvHD data. The lattice graphics package of Deepayan Sarkar (14) assisted this research.

Literature Cited

1. Gasparetto M, Gentry T, Sebti S, O'Bryan E, Nimmanapalli R., Blaskovich MA, Bhalla K, Rizzieri D, Haaland P, Dunne J, Smith C. Identification of compounds that enhance the anti-lymphoma activity of rituximab using flow cytometric high-content screening. J Immunol Methods 2004;292:59–71.
2. Le Meur L, Rossini A, Gasparetto M, Smith C, Brinkman RR, Gentleman R. Data quality assessment of ungated flow cytometry data in high throughput experiments. Cytometry A 2007;71A:393–403.
3. Godtliebsen F, Marron JS, Chaudhuri P. Significance in scale space for bivariate density estimation. J Comput Graph Stat 2002;11:1–22.
4. Duong T, Cowling A, Koch I, Wand MP. Feature significance for multivariate kernel density estimation. Comput Stat Data Anal 2008;52:4225–4242.
5. Durban M, Harezlak J, Wand MP, Carroll RJ. Simple fitting of subject-specific curves for longitudinal data. Stat Med 2005;24:1153–1167.
6. Brinkman RR, Gasparetto M, Lee S-JJ, Ribickas AJ, Perkins J, Janssen W, Smiley R, Smith C. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. Biol Blood Marrow Transplant 2007;13:691–700.
7. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry A 2008;73A:321–332.
8. Draper NR, Smith H. Applied Regression Analysis, (3rd ed). New York: Wiley; 1998.
9. Donnelly CA, Laird NM, Ware JH. Prediction and creation of smooth curves for temporally correlated longitudinal data. J Amer Stat Assoc 1995;90:984–989.
10. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN3-900051-07-0. 2008 www.R-project.org.
11. Gentleman R, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang YH, Zhang J. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.
12. Ellis B, Haaland P, Hahne F, Le Meur N, Gopalakrishnan N. flowCore: Basic structures for flow cytometry data. R package version 1.8.0. 2008.
13. Pinheiro J, Bates D, DebRoy S, Sarkar D. the R Core team. nlme: linear and nonlinear mixed effects models. R package version 3.1–89. 2008.
14. Sarkar D. Lattice: Lattice Graphics. R package version 0.17–15. 2008.
15. Duong T, Hazelton ML. Plug-in bandwidth matrices for bivariate kernel density estimation. J Nonpar Stat 2003;15:17–30.
16. Wand MP, Jones MC. Multivariate plug-in bandwidth selection. Comput Statist 1994;9;97–116.
17. Duong T.ks: Kernel Smoothing. R package version 1.5.10. 2009.
18. Hyndman RJ. Computing and graphing highest density regions. Am Stat 1996;50: 120–126.

Appendix

Details of the `curvHDR` gating method

Let d be the dimension of data in which a gate is sought and let $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ be the sample in $\mathbb{R}^d$ to be gated. The `curvHDR` methods assumes that gates of interest correspond

to modal regions in the sample. This first entails assuming that the $\boldsymbol{x}_i$s are a sample from a d-variate density function f. Modal regions then correspond to local maxima in f and their surrounds. The first phase of the `curvHDR` method uses recently developed feature significance technology (4) to find regions where f has statistically significant high negative curvature. This phase can be thought of as filtering process where aberrant regions of high relative density are ignored and only those regions having statistical evidence of modality are retained. The second phase aims to improve upon the regions obtained in the first phase by modifying them to suit the local density of the data around each high curvature region.

The specific steps of the `curvHDR` gating method are:

1. Remove excessive boundary points and other debris from the data. If the data exhibits heavy skewness then transform the data to reduce skewness. A good "all-purpose" transformation is the inverse hyperbolic sine transformation $x_{\text{new}} = \sinh^{-1}(x) = \log(x + \sqrt{x^2 + 1})$.
2. Standardize all variables to have zero mean and unit standard deviation.
3. Obtain significant (negative) curvature regions using the test described in Section 3.2 of (4) over a d-dimensional mesh.
4. Replace each significant curvature region by its convex hull.
5. Grow each convex hull so that its volume is G times larger (for some prespecified growth factor $G > 1$). This is achieved by "rolling" a d-dimensional sphere around the perimeter of the region.
6. For each grown region, obtain a kernel density estimate, based on a multistage plug-in bandwidth selector (15), and using only the data within that region.
7. The `curvHDR` gates are the level-τ highest density regions (see definition below) based on the kernel density estimates from Step 5.
8. Transform the gates back to the original units.

Step 3 requires estimates of the Hessian matrix of f, the $d \times d$ matrix with (i,j) entry equal to $\frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x})$, with x_i denoting the ith entry of $\boldsymbol{x}$. Each derivative estimate is obtained via appropriate differentiation of the d-variate kernel density estimator

$$\hat{f}(\boldsymbol{x}; \boldsymbol{H}) = n^{-1} |\boldsymbol{H}|^{-1/2} \sum_{i=1}^{n} K\{\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{x}_i)\},$$

where K is a d-variate kernel function and $\boldsymbol{H}$ is a $d \times d$ bandwidth matrix. Details are given in (4). In `curvHDR`, we use a single parameter bandwidth matrix $\boldsymbol{H} = h_{\text{curv}}^2 \boldsymbol{I}$ for some $h_{\text{curv}} > 0$. This is partially justified by the fact that input data for kernel density estimation is such that each variable has unit standard deviation. The Penalized Spline Regression Analysis section of (4) describes how the estimated Hessian matrix can be used to determine regions in $\mathbb{R}^d$ where f has significant negative curvature. These correspond to local maxima in the underlying density and identify candidate locations for which gating might be appropriate. The R package `feature` provides implementation of the significant curvature determination.

The convex hull of a closed polygon in $\mathbb{R}^2$ is a well-known geometrical construct. A useful physical interpretation involves imagining the vertices of the polygon as nails on a board and placing an elastic band around the nails. Software for convex hull computation in Step 4 is available in the R computing environment via the function `chull()`.

We will describe Step 5 in the case $d = 2$. The case $d = 1$ is analogous, but simpler. Step 5 involves growing a convex polygon to be G times larger in area via the notion of "circle-rolling". We first note that the area of a polygon with vertices

$$\mathcal{P} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

and ordered clockwise and such that $(x_1,y_1) = (x_N,y_N)$ is

$$A(\mathcal{P}) = \frac{1}{2} \sum_{i=1}^{N-1} (x_i y_{i+1} - x_{i+1} y_i).$$

Now suppose that we roll a circle of radius r around the perimeter of $\mathcal{P}$. A polygonal approximation to the resulting region is obtained by forming normal vectors to each edge of $\mathcal{P}$ that start from the centre of the edge and radiate outwards a distance of $2r$. This approach is illustrated in Panel (c) of Figure 2. Let $\mathcal{P}_r$ denote the polygon obtained by joining each of the normal vectors. Step 5 is completed by solving for the r that satisfies $A(\mathcal{P}_r)/A(\mathcal{P}) = G$.

Step 6 involves application of the kernel density estimator to each grown region and the data that it contains. The kernel K is taken to be the d-variate standard normal density function $K(\boldsymbol{x}) = (2\pi)^{-d/2} \exp(-\boldsymbol{x}^T\boldsymbol{x}/2)$. The bandwidth matrix is chosen using multistage plug-in strategies (15,16) courtesy of the R package `ks` (17). In most cases, the Step 6 density estimates are concerned with unimodal structure where plug-in bandwidths perform quite well.

For a d-variate density function f and $\tau \in [0,1]$ the τ highest density region (HDR) is $R_\tau \equiv \{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) \geq f_\tau\}$ where f_τ is the greatest number for which $\int_{R_\tau} f(\boldsymbol{x}) d\boldsymbol{x} \geq 1 - \tau$ (e.g., (18)). We can think of the R_τ as corresponding "meaningful" contours of the density function f. For example, $R_{0.9}$ is the region inside that contour of f for which the probability is 0.1, a relatively small region near the peak of f. The HDR $R_{0.1}$ encompasses to 90% of the probability mass of f. In practice, where f is unknown, estimated HDRs can be obtained by substituting f with a density estimate. In Step 7, we apply the HDR paradigm to each of the density estimates from Step 6.