A framework for analytical characterization of monoclonal antibodies based on reactivity profiles in different tissues

Elizabeth Rossin1,6,†, Tsung-I Lin2,3,†, Hsiu J. Ho4, Steven J. Mentzer5 and Saumyadipta Pyne6,7,*

1 Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA, 2 Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, 3 Department of Public Health, China Medical University, 4 Laboratory of Adaptive and Regenerative Biology, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA 02115, 5 Broad Institute of MIT and Harvard University, Cambridge, MA 02142 and 6 Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA

ABSTRACT

Motivation: Monoclonal antibodies (mAbs) are among the most powerful and important tools in biology and medicine. MAb development is of great significance to many research and clinical applications. Therefore, objective mAb classification is essential for categorizing and comparing mAb panels based on their reactivity patterns in different cellular species. However, typical flow cytometric mAb profiles present unique modeling challenges with their non-Gaussian features and intersample variations. It makes accurate mAb classification difficult to do with the currently used kernel-based or hierarchical clustering techniques.

Results: To address these challenges, in the present study we developed a formal two-step framework called mAbprofiler for systematic, parametric characterization of mAb profiles. Further, we measured the reactivity of hundreds of new antibodies in diverse tissues using flow cytometry, which we successfully classified using mAbprofiler.

First, mAbprofiler fits a mAb’s flow cytometric histogram with a finite mixture model of skew t distributions that is robust against non-Gaussian features, and constructs a precise, smoothly varying and mathematically rigorous profile. Then it performs novel curve clustering of the fitted mAb profiles using a skew t mixture of non-linear regression model that can handle intersample variation. Thus, mAbprofiler provides a new framework for identifying robust MAb classes, all well defined by distinct parametric templates, which can be used for classifying new mAb samples. We validated our classification results both computationally and empirically using mAb profiles of known classification.

Availability and Implementation: A demonstration code in R is available at the journal website. The R code implementing the full framework is available from the author website – http://amath.nchu.edu.tw/teacher/tlin/software

Received on March 22, 2011; revised on July 21, 2011; accepted on August 7, 2011

1 INTRODUCTION

Monoclonal antibodies (mAbs) are among the most powerful, popular and important tools in a biomedical laboratory for probing different cellular types, states and functions. Research in the past decades has led to the development of large collections of mAb for specific binding to cell surface antigens, which facilitated purification and functional characterization of a variety of cell populations. It also unlocked the great potential of using mAb for therapy in many serious diseases such as cancer. Using platforms such as flow cytometry, one can measure quantitatively the binding of a mAb, in single cell resolution, to the corresponding antigen whose expression may serve as a marker of cellular characteristics for a given specimen, see Herzenberg et al. (2001). Therefore, it is important to characterize mAb reactivity patterns in different cell types and tissues with analytical precision and rigor so that both known and new mAb can be categorized and compared accurately and objectively.

mAb classification is of great practical importance to many fields in biomedicine such as immunology, hematology, pathology and clinical immunotherapy. Large-scale attempts at analyzing mAb to identify new molecules were pioneered in the human leukocyte differentiation antigens (HLDA) workshops [see review in Zola and Swart (2005)] where the reactivities of large panels of mAbs were measured against widely available cell lines. The reactivity was given a binary assignment compared with a negative control—either the antibody bound to its antigen on a given cell or it did not—as measured by fluorescence intensity. The frequency with which this occurred over a cell population was then recorded, and hierarchical clustering was employed to group similar reactivity—thus was born the ‘Clusters of Differentiation’ (CD) classes, widely used today to identify various cell populations (Bernard and Boumsell, 1984).

In recent years, the workshop approach for identifying new molecules to define cell types has become less applicable due to the current capabilities of molecular identification at gene level (Zola and Swart, 2005). An alternative approach for mAb characterization involves the use of primary cell populations that
are derived systematically from different tissues in selected species (e.g. Pratt et al., 2009). Typically, mAb reactivity patterns, as measured with cytometric density histograms, can present jagged non-smooth curves with features in the form of peaks and shapes that are difficult to characterize analytically. Intersample variation in cytometric data makes the modeling problem even more challenging. Not only do these make accurate binary percent positive/negative calls harder but also render ineffective the current clustering approaches that are poorly suited to model or classify such noisy curve profiles.

In general, analytical characterization of mAb reactivity patterns has received limited attention in statistics and computer science (Gilks and Shaw, 1995; Kim et al., 2002; Pratt et al., 2009; Salganik et al., 2005; Spiegelhalter and Gilks, 1987; Zeng et al., 2002, 2007 and references therein). As shown in Pratt et al. (2009), mAb classification faces technical challenges at multiple levels. Single parameter flow cytometric histograms used for measuring mAb reactivity often have multiple peaks with non-Gaussian features and irregular shapes. Few of the known algorithms can model the underlying distributions and their key features precisely and robustly. In addition, due to cytometric platform noise, the measurements of peak features tend to vary in terms of both significance and location, making direct comparison of samples challenging. Moreover, standard clustering approaches meant for multivariate points, such as hierarchical clustering, are not well suited for grouping curves, which in this case represent histogram profiles. Histogram profiles, when viewed as points, can vary considerably with different choices of binning parameters, producing jagged patterns. Hence, a new clustering approach is necessary that can robustly detect the characteristic features lying within every mAb’s noisy curve profile, which is not merely a multivariate point. Simultaneously, the approach must also account for the cytometric intersample variation among the curve features across mAb profiles to achieve accurate classification.

To address these challenges, in the present study we generated (i) new data for a large collection of mAb; and (ii) developed mAbProfiler, a new general framework to characterize and cluster mAb profiles systematically and rigorously. More than 1000 subcloned murine hybridomas, made against sheep cell membrane antigens, were considered for analysis. A subset of mAbs were selected for inclusion in this study based on their distinctive staining profile and surface expression in six diverse tissues (splenocytes, lymph node cells, alveolar macrophages, effenter lymphocytes, fetal thymus and thymocytes). Further, in mAbProfiler, we present a two-step framework based on new parametric modeling algorithms.

In the first step, it profiles every mAb defined by its flow cytometric histogram with a finite mixture model of skew t distributions that is robust against both outliers and asymmetry, which are often responsible for producing non-Gaussian features. An Expectation–Maximization (EM) algorithm is used for fitting every pattern with a smooth and mathematically rigorous profile that specifies all key features precisely, with the help of a probability density function. In the second step, for each tissue, mAbProfiler performs curve clustering of the fitted mAb profiles with a novelty of skew t mixture of non-linear regression model that is robust against intersample variation. We used an effective criterion, the Jump Statistic, for model selection with the optimal number of clusters (or mAb classes). In addition to these robust tissue-specific mAb classes, our framework uncovered new group structures among profiles undetected by traditional approaches like hierarchical clustering. Importantly, mAbProfiler also generates class-specific parametric signatures that can be used for (i) comparing and categorizing mAb classes and (ii) classifying new mAb panels.

Finally, we validated our classification results for different tissues both computationally and empirically using mAb profiles of previously known classification.

2 MATERIALS AND METHODS

MAb production and sample generation: we followed the system of mAb cloning and harvesting using protocols developed in our lab and described in Li et al. (1995), Pratt et al. (2009) and references therein. The panel of antisheep mAbs were tested for reactivity against six different sheep tissues: splenocytes, lymph node cells, alveolar macrophages, effenter lymphocytes, fetal thymus and thymocytes. The cells were washed and fixed before they were analyzed with an Epics XL flow cytometer (Beckman Coulter, Miami, FL, USA). After quality control by human expert inspection, we generated 561 mAb reactivity patterns categorized by tissue: spleen, lymph node, lung lavage, bone marrow, fetal thymus and thymocytes. The fitted mAb profiles are available from the authors upon request.

Cytometric data preparation: each flow cytometric sample was represented as a three-column matrix, where columns contained forward-scatter, side-scatter and the fluorescence intensity for a particular fluorophore tagged antibody, and each row represented a single cell (or event). The data were preprocessed to remove debris and dead cells. In all, 98% of the data consisted of experiments where 10 000 events were captured. After log10 transformation of the data, we performed multistep cleanup and filtering: first, we removed points from samples whose maximum intensity value was populated by >25% of cells. Such spikes are signs of poor calibration during data acquisition. Second, we filtered zero fluorescent values which might also represent possible calibration problems. Finally, we removed extreme outliers in data (points more than 3 SDs away from the sample mean) that are most likely due to platform noise. After filtering, the median number of events for each sample was ~9390.

Step 1 of mAbProfiler (histogram profiling): after filtering, the resulting histogram of each antibody’s fluorescence intensity was fitted with a finite mixture model of univariate skew t distributions using an EM algorithm described in Supplementary Material. Since Bayesian Information Criterion (BIC) is known to select restrictive models which may be inadequate for feature detection in our non-smooth data, we instead used the well-known Integrated Completed Likelihood (ICL) criterion for our model selection (McLachlan and Krishnan, 2008). ICL scores for optimal models showed no further improvement for most samples beyond 10 components, which was the maximum number of components fit by the mixture model. Figure 1 shows a sample histogram and the fitted profile with a gray curve. Clearly, the optimal model produced a smooth and accurate profile, and all the significant features and their locations are captured and specified by the model parameters. Since the model is a univariate version of (Pyne et al., 2009) approach, we have described it along with its EM algorithm in the Supplementary Material for completeness.

Step 2 of mAbProfiler (profile clustering): here we present a new and robust model-based curve clustering approach using skew t mixture of non-linear regression model. The model along with the EM algorithm for clustering of the mAb profiles (which were fitted in Step 1) are described below.

Following Azzalini and Dalla-Valle (1996), a random vector Z is said to follow the multivariate skew normal (MSN) distribution, denoted by 

\[ Z \sim SN_{p}(\mu, \Sigma, \lambda) \]

if its density takes the form

\[ f(z) = 2\phi(z|\mu, \Sigma) \phi(\lambda)^{1/2}(1-\phi(\lambda)) \]

where \( \phi(z|\mu, \Sigma) \) denotes the pdf of p-variate normal distribution with mean vector \( \mu \) and covariance matrix \( \Sigma \). \( \phi(.) \) represents the cdf (short for cumulative distribution function) of the standard normal distribution
Fig. 1. Profiling of a mAb reactivity against the surface of sheep cells: a cytomeric histogram measuring the reactivity of a particular mAb tested against the surface of sheep effluent lymphocytes is shown as log of fluorescence intensity against surface expression. In Step 1 of the profile, the surface expression pattern is profiled with a skew t mixture model as depicted by the smooth and precisely fit gray curve. It captures non-Gaussian features such as skewness and outliers common in cytomeric distributions. For the original non-skew, symmetric, see Supplementary Figure S1.

and $\Sigma^{-1/2}$ is the square root matrix of $\Sigma^{-1}$ satisfying $\Sigma^{-1/2}\Sigma^{-1/2} = \Sigma^{-1}$. If $\lambda = 0$, then the distribution of $Z$ reduces to $\mathcal{N}(\mu, \Sigma)$.

For the ease of theoretical and computational developments, Arellano-Valle et al. (2005) gave the following stochastic representation for the MSN distribution:

$$Z = \mu + \Sigma^{1/2}[\Delta(U_0 + iU_1 + i\nu)^{-1/2}U_1], \quad U_0 \perp U_1,$$

(1)

where $\Delta = \lambda/\nu + i1/\nu$, $U_0 \sim \mathcal{N}(0, 1)$, $U_1 \sim \mathcal{N}(0, \lambda^2)$ and the symbol $\perp$ indicates independence.

The multivariate skew t (MST) distribution was proposed by Azzalini and Capitanio (2003), which is related to the MSN distribution as follows:

$$Y = \mu + \tau^{-1/2}Z, \quad Z \sim \tau, \quad \tau = \Gamma^2\gamma,$$

(2)

where $Y \sim \mathcal{SN}(\mu, \Sigma, \lambda, \nu)$ and $\gamma = \Gamma\gamma$ given as follows:

$$\gamma |y \sim \Phi(\nu, \tau) \prod_{i=1}^n \Phi(\nu, \tau),$$

(3)

where $\Phi(\nu, \tau)$ is the pdf of the Student’s $t$ distribution with $\nu$ degrees of freedom with mean $\mu$ and variance $\tau$. We shall denote $Y \sim \mathcal{SN}(\mu, \Sigma, \lambda, \nu)$ if $\tau$ has density given in (5).

Suppose we have a set of $m$ input profiles $y_j^{(j)}$, and each response vector $y_j$ consists of $n_j$ observations. We assume the response vector $y_j \in \mathbb{R}^{n_j}$ is generated from

$$y_j = \mu(\beta, x_j) + \delta_j, \quad (j = 1, \ldots, m),$$

where $\beta$ is a $p \times 1$ vector of regression coefficients related to design matrix $X = [x_1, \ldots, x_m]^T$ with $x_0 = [1, x_1, \ldots, x_p]^T$, $\mu(\beta, x)$ is a vector-valued non-linear (differentiable) function of $\beta$ governing within-profile behavior, and $\delta_j$ is the resulting error vector equal to the discrepancy between $y_j$ and $\mu_j$.

A skew-t based non-linear regression model is defined as follows:

$$\psi_j(y_j, \theta) = \tau_j \mathcal{N}(y_j^{(j)}; \mu_j^{(j)}, \sigma_j^{(j)}),$$

(7)

where $\tau_j$ is the scale parameter, $\mu_j^{(j)}$ and $\sigma_j^{(j)}$ are the mean and standard deviation of $y_j$, respectively. The density of a cluster-specific skew-t non-linear regression model that relates $(y_j, x_j)$ to $\theta = (\phi, \tau_j^{(j)}, \mu_j^{(j)}, \sigma_j^{(j)})$

The mixture model for profile clustering is written as:

$$y_j = \sum_{c=1}^{K} w_c y_c + \epsilon_j, \quad (j = 1, \ldots, m),$$

(8)

where $w_c$ is the mixing proportion which is constrained to be non-negative and $\sum_{c=1}^{K} w_c = 1$. The parameter $\theta = (\psi_{1}, \ldots, \psi_{c}, \phi_{1}, \ldots, \phi_{c})$ represents all unknown parameters.
parameters. The observed data log-likelihood function of \( \Theta \) is
\[
\ell (\Theta ; y) = \sum_{i=1}^{N} \log f(y_i | \Theta).
\] (9)

In general, there are no explicit analytical solutions for computing the ML estimator of \( \Theta \). The EM algorithm (Dempster et al., 1977) is considered as a standard tool when applied for mixture models. In the EM framework for supporting the interpretation of incomplete data, it is convenient to introduce a set of allocation variables \( Z_i = (Z_{ij})_{j=1, \ldots, m}, \) \( j = 1, \ldots, m \). The element \( Z_{ij} \) is taken to be one or zero to indicate if \( y_i \) does or does not come from the \( i \)-th component. This implies that \( Z_i \) follows a multinomial distribution with 1 trial and cell probabilities \( w_1, \ldots, w_m \), denoted by \( Z_i \sim \text{Mult}(1, w_1, \ldots, w_m) \). Then, a hierarchical formulation of (8) obtained in conjunction with (4) is
\[
y_j(y_j, \tau_j, Z_j) = N_0 \left( \mu_0 + \frac{\lambda_j}{1+\mu_0^2} \right) \sigma_j^2 \left( 1+\mu_0^2 \right)^{-1/2}.
\] (10)

Let \( y = (y_1, \ldots, y_n), y_j = (y_j, \ldots, y_{j+1}), y = (y_1, \ldots, y_n) \) and \( Z = (Z_1, \ldots, Z_n) \). It follows from (10) that the complete data log-likelihood function of \( \Theta \) given \( y, z, Z, \tau, \gamma \) is
\[
\ell_c (\Theta ; y, z, Z, \tau, \gamma) = \sum_{j=1}^{N} \log \left( \frac{1}{\sigma_j^2} \right) - \frac{1}{2} \sigma_j^{-2} \sum_{j=1}^{N} \left( y_j - \mu_j \right)^2,
\] (11)
where \( \sigma_j^2 = \lambda_j \sigma_j \left( 1+\mu_0^2 \right)^{-1/2} \).

The EM algorithm proceeds by alternately repeating the E- and M-steps where, at the \( k \)-th iteration, the E-step involves the calculation of the Q-function, which is the expected value of the complete data log-likelihood (11) conditional on \( y \) and the current estimate \( \hat{\Theta} \) for \( \Theta \), is given by
\[
Q(\Theta ; \hat{\Theta}) = E(\ell_c (\Theta ; y, z, Z, \tau, \gamma) | y) = \sum_{j=1}^{N} \int \left[ \sum_{i=1}^{N} \log f(y_i | \Theta) \right] \frac{1}{Z_i} = \sum_{j=1}^{N} \int \log f(y_j | \Theta) \frac{1}{Z_j}.
\] (12)

To evaluate (12), the necessary conditional expectations include
\[
\gamma^{(k+1)} = E (y_i | \Theta^{(k+1)}), \quad \gamma^{(k+1)} = E (y_i | \Theta^{(k+1)})
\] (13)
which are the posterior probability that the \( i \)-th curve belongs to the \( i \)-th component evaluated at the \( k+1 \)-st iteration. Therefore, the Q-function (12) can be written as
\[
Q(\Theta ; \hat{\Theta}) = \sum_{j=1}^{N} \int \left[ \sum_{i=1}^{N} \log f(y_i | \Theta) \right] \frac{1}{Z_i} = \sum_{j=1}^{N} \left[ \gamma^{(k+1)} \log \left( \frac{1}{\sigma_j^2} \right) \right] + \frac{\gamma^{(k+1)} \log \left( \frac{1}{\sigma_j^2} \right)}{2} - \frac{1}{2} \sigma_j^{-2} \sum_{j=1}^{N} \left( y_j - \mu_j \right)^2
\] (14)

We can easily done with the help of the R routine "uniroot." The E- and M-steps are alternately repeated until a suitable convergence rule is satisfied, e.g. the Aitken acceleration-based stopping criterion \( |\gamma^{(k+1)} - \gamma^{(k)}| < \epsilon \), where \( \epsilon \) is the observed log-likelihood evaluated at \( \gamma = 1 \). \( \ell(\gamma) \) is the asymptotic estimate of the log-likelihood at iteration \( k+1 \). (McLachlan and Krishnan 2008, Chapter 4.9) and \( \epsilon \) is the desired tolerance.

Model selection for Step 2: let \( X \) be a p-dimensional random sample drawn from a mixture distribution of \( g \) components, each with homogeneous covariance matrix \( \Sigma_{\gamma} \). Let \( \gamma_{1}, \ldots, \gamma_{g} \) be a set of candidate cluster centers with \( \gamma_i \) being the one closest to \( X \). Sugar and James (2003) developed an alternative simple approach to identify the optimal number of clusters based on the ‘distortion function’, defined as
\[
d_{\mathcal{C}} = \frac{1}{p} \sum_{i=1}^{g} \min_{\gamma \in \mathcal{C}} E(X - \gamma)^2 \Gamma^{-1}(X - \gamma),
\] (15)
where \( \mathcal{C} \) is a quantity that measures the average Mahalanobis distance between \( X \) and its closest cluster center \( \gamma \). The Jump function due to Sugar and James (2003) is defined as
\[
J_{\mathcal{C}} = d_{\mathcal{C}}^{(k+1)} - d_{\mathcal{C}}^{(k)}
\] (16)
where \( C \) is an appropriate positive constant that makes a sharp jump at the true number of clusters and \( d_{\mathcal{C}}^{(k)} \) is the minimum distortion obtained by the clustering algorithms. They have proven that an appropriate number of clusters can be identified at the peak of jump based on information-theoretic ideas. Their simulation studies have also empirically shown that the jump plot has good performance in finding the true number of clusters.

We applied the Jump function approach to the problem of curve clustering analyses. Each curve is assigned to the component with the largest posterior probability obtained by fitting model (8) for \( g=1, \ldots, g_{\text{max}} \), a pre-specified maximum number of components. We chose \( g_{\text{max}} \) to be 12 for all tissues, except for two (spleen and bone marrow) where the model did not converge for \( g > 10 \). Let \( \beta_k \) be the fitted vector of \( y_j \) if \( \beta_j \) has been assigned outright to \( i \)-th cluster, say \( y_j, x_j \). This gives
\[
\beta_k = \tilde{E} \left( \mu_j + \frac{\lambda_j}{1+\mu_0^2} \right) \left( Z_{ij} = 1, \Theta \right) \Theta \Theta.
\]
KL (the maximal value of the Jump statistic over a range of clusters, log-likelihood maxima, BIC values, the distances $D_n$ specified as density curves, with our new algorithm for non-linear modeling with our skew $\hat{g}$ step. The optimal choice of $\hat{g}$ is to look for the point at which the magnitude of change in $d_g$'s becomes negligible, especially when the subclasses are well separated. However, using the raw distortion curve could fail in certain cases. As suggested by Sugar and James (2003), the Jump plot method performs extremely well, provided that some suitable values for $\gamma$ are chosen. The optimal number of clusters in data can be visually determined from the peak patterns on the Jump plot. Empirical studies show that the point with largest or secondary largest jump is often the best choice.

Quality of curve clustering: to determine the quality of our clustering of mAb profiles, modeled as probability density functions, we measured mean intra- and intercluster distances using a symmetric form of Kullback-Leibler distance, denoted by $\text{sKL}(p,q)$, between a pair of profiles $(p,q)$, defined as follows:

$$\text{sKL}(p,q) = KL(p,q) + KL(q,p)/2,$$

where $KL(p,q) = \sum_j p_j \log(p_j/q_j)$ at each observation point $r$.

To determine the quality of hierarchical clustering of mAb data, we used the R functions $\text{hclust}$ (with Euclidean distance metric) and $\text{average}$ (average silhouette width). The R package $\text{sil}c$ is used for $\text{SiZer}$ plot.

3 RESULTS

Following data preparation and preprocessing, in Step 1, mAbprofiler modeled cytometric histograms for 561 mAb samples from six tissues using skew $t$ mixture models. Figure 1 illustrates how a profile (shown as gray curve) constructed by mAbprofiler offers a smooth and precise representation of mAb density histograms. This can be contrasted with the original cytometric histograms. For empirical validation, for four tissue types (spleen, thymocytes, bone marrow and lymph node) and two classes of antibodies [class I mAb T2/39 and anti-LFA-1 mAb F10-150 as described in Pratt et al. (2009)]; we generated data for two pairs of mAb such that the mAb within each pair were known to target molecules of the same class, but across pairs, they targeted molecules from distinct classes. As shown in Supplementary Figure S4, indeed all profile pairs (black thin curves) cluster together within each class (Class I in left panel or anti-LFA-1 in right panel), but separately across two distinct classes, providing experimental evidence for precise and objective classification by our framework. As in Figure 3, the mean profiles are

---

Table 1. Clustering statistics for Step 2

<table>
<thead>
<tr>
<th>Tissue type</th>
<th>No. of samples</th>
<th>No. of classes</th>
<th>IIR</th>
<th>CT</th>
<th>$f_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spleen</td>
<td>59</td>
<td>6</td>
<td>0.078</td>
<td>27.09</td>
<td>10</td>
</tr>
<tr>
<td>Thymocyte</td>
<td>111</td>
<td>12</td>
<td>0.068</td>
<td>118.04</td>
<td>12</td>
</tr>
<tr>
<td>Lung</td>
<td>123</td>
<td>11</td>
<td>0.017</td>
<td>95.06</td>
<td>12</td>
</tr>
<tr>
<td>Bone marrow</td>
<td>48</td>
<td>8</td>
<td>0.014</td>
<td>23.94</td>
<td>10</td>
</tr>
<tr>
<td>Fetal thymus</td>
<td>89</td>
<td>12</td>
<td>0.038</td>
<td>81.24</td>
<td>12</td>
</tr>
<tr>
<td>Lymph node</td>
<td>131</td>
<td>8</td>
<td>0.072</td>
<td>111.21</td>
<td>12</td>
</tr>
</tbody>
</table>

Fig. 2. Model selection for curve clustering: maximum value of Jump for spleen indicates that 6 is the optimal number of mAb classes in that tissue.
MAb classification

Fig. 3. Classification of mAb profiles for spleen: in Step 2 of mAbprofiler, the profiles of all 59 mAb samples in spleen were clustered with non-linear regression of skew t mixture models. The profiles belonging to each of the six clusters (a)–(f) are shown in distinct colors. The joint plot (g) of all six mean profiles in cluster-specific colors allows visual comparison of the cluster templates.

3.1 Comparative analysis with other methods

Besides internal validation, we also compared the performance of mAbprofiler with other established methods. We began with hierarchical clustering, which is the most commonly used approach for mAb classification (Bernard and Boumsell, 1984). When we used hierarchical clustering on our mAb profiles, then the method clearly failed to capture the complex class structure and detected few clusters. Based on Average Silhouette Width (ASW), a common measure for determining the quality of hierarchical clustering, we noted that the optimal number of mAb classes according to hierarchical clustering of our data was typically restricted to four or even fewer for all tissues other than thymocytes. Moreover, little difference among the ASW scores for different number of clusters indicated that the hierarchical clusters had low separation (Supplementary Figure S5).

Thereafter, we adopted the established protocol of Pratt et al. (2009) in which mAb histograms were first smoothed with SiZer, and then hierarchical clustering was performed with those smoothed profiles. We show the results of that approach on our data using SiZer plots for the different tissues in Supplementary Figure S6a–f. As depicted with the dendrograms, while the larger clustering structures were detected with smoothing, the finer structures were often ignored, thus resulting in highly heterogeneous classes. This can be seen clearly in the largest clusters in spleen, lung lavage and bone marrow.

Finally, we also studied a combination of our approach with that of Pratt et al. (2009) in which we clustered the SiZer-smoothed profiles (i.e. we replaced Step 1 of mAbprofiler) using our NLRST algorithm (i.e. we retained Step 2 of mAbprofiler). We observed that while NLRST could identify more classes in the same SiZer profiles for some tissues, the overall gain was not significant. In other words, the fact that mAbprofiler identified a much richer class structure could be attributed to the dual contributions of both Steps 1 as well as 2 of the new framework. While NLRST tackles the intersample variation along x-direction (feature location), the skew t mixture pdf provides a precise and continuous representation of the y-direction (feature significance). The resulting effectiveness of mAbprofiler’s two-step approach is illustrated, for example, in the eight class-templates for the lymph node which are distinctive along both x- and y-directions (Supplementary Figure S2, left panel bottom plot). In contrast, the other methods failed to capture that dual complexity and identified only few dominant clusters. The full comparison of classes detected by all four methods is shown in Supplementary Table S2.

4 DISCUSSION

Monoclonal antibodies play an immensely important role in molecular biology, biochemistry and medicine. Their utility for probing, stimulating or inhibiting specific target molecules supports numerous diagnostic and immunotherapeutic applications (Zola, 2006). Further, design and development of new mAbs are also of great industrial significance. Therefore, objective mAb classification is essential for categorizing and comparing the known as well as the newly developed mAb panels. Besides biochemical methods like immunoprecipitation, this is achieved by clustering flow cytometric reactivity patterns of mAbs in different cell types. Unlike traditional HLDA workshops which classified leukocyte surface (CD) molecules (Zola and Swart, 2005) using cell lines, Pratt et al. (2009) recently described a system to facilitate practical mAb characterization in animal tissues. This approach is consistent with the new human cell differentiation molecules (HCDM) focus on various non-hematopoietic cell types (Zola, 2006). In the present study, we enhanced that approach further by (i) generating a new, larger and more varied collection of mAb patterns in six different tissues, and importantly, by (ii) constructing a new analytical framework, mAbprofiler, to formally address...
the technical challenges of mAb characterization. Our two-step framework provides precise profiling of cytometric histograms (Step 1) followed by novel clustering of these curve profiles (Step 2). In addition to characterizing mAb for the present study, mAbprofiler can also provide a general framework to allow users to search for archived class signatures or to construct and classify new mAb profiles in a systematic way.

Previous mAb classification studies (e.g. Pratt et al., 2009; Salganik et al., 2005) have used non-parametric kernel density estimation techniques for detection of significant features in cytometric histograms, typically followed by hierarchical clustering based on Euclidean distances between the features. While being practical, such approaches may not always be precise or robust. For instance, the accuracy of density estimates by kernel-based methods are known to be strongly influenced by bandwidth selection (Jones et al., 1996). As observed in Supplementary Figure S1, significance of the peak features in a given sample, as detected by the program SiZer, is clearly dependent on the choice of bandwidth. This poses a key practical problem, especially since we seek to do unsupervised classification of new mAb profiles. While recent advances in kernel-based techniques have addressed different aspects of cytometric analysis (e.g. Duong et al., 2009; Naumann et al., 2010), we followed the parametric approach developed by Pyne et al. (2009) and (Frühwirth-Schnatter and Pyne, 2010), which uses finite mixtures of skewed t distributions, for our purposes. Observations of non-Gaussian features in cytometric data made by these and other recent studies (Ho et al., 2011; Lo et al., 2008; Pyne et al., 2011) led us to use this more general parametric family of distributions, which also includes Gaussian distribution as a special case.

Finite mixture models have been extensively used in biology and medicine (Frühwirth-Schnatter, 2006; McLachlan and Peel, 2000). In Step 1 of mAbprofiler, we presented a univariate version of the Pyne et al.’s (2009) approach for profiling asymmetric and noisy mAb patterns with finite mixture model of skewed t distributions fit via our own EM algorithm (see Supplementary Material). The EM algorithm converges fast in practice, and supports multiple well-known model selection criteria such as AIC, BIC and ICL. In the resulting smooth and precise profiles (see illustrative sample in Fig. 1), every component is specified by rigorous model parameters such as location, size, shape, variance and degrees of freedom. Further, the parametric design enables mAbprofiler to specify the significance of every mAb feature with a smooth and continuous probability density function, which can be represented as a curve that is well defined at any resolution. Importantly, in Step 2, mAbprofiler’s non-linear regression of skew t mixture models can cluster these curve profiles accurately for every cell type. While Step 1 follows the approach of Pyne et al. (2009), Step 2 introduces novel methodology and the EM algorithm implementing it.

A key challenge for cytometric data analysis is intersample variation. Similar mAb profiles can vary considerably in both their significance and location, which must be addressed by any algorithm designed for classifying cytometric data. While it is possible to transform or shift and align the data (e.g. Hahne et al., 2010; Lo et al., 2008), we want to cluster the mAb profiles precisely in terms of the distinctive features that they present as curves with a robust approach. To systematically model that intersample variation, in Step 2, mAbprofiler presented a new non-linear regression algorithm. It is also a solution for the more generic problem of curve clustering, an important topic in the field of pattern recognition which has not received much attention in the past (e.g. Gaffney, 2004; Gaffney et al., 2007; Liu and Yang, 2009). Here we extended the work of Gaffney (2004) and Jones and McLachlan (1992) to introduce non-linear regression of skew t mixture models for robust curve clustering with asymmetric variation among the curve features. In our comparative analysis with other methods, we observed that hierarchical clustering is not as well suited for such clustering probably because it critically relies on precise pairwise distances between points. Trying to reduce a curve profile to a point—albeit a multidimensional point—can lead to loss of information about features due to binning of the data as specified by a cytometric histogram. That leads to fewer and less well-separated hierarchical clusters, as illustrated in Supplementary Figure S5.

The problem of using hierarchical clustering for mAb classification gets further compounded with the issue of bandwidth selection in smoothing of cytometric histograms such as in the protocol of Pratt et al. (2009). For our data, the SiZer-smoothed features for a predetermined bandwidth led to mAb classes with high heterogeneity. While our NLRST (Step 2) clustering could improve detection of the classes with the same SiZer profiles, the net gain was not significant. Therefore, the identification of a much richer group structure by mAbprofiler, as shown in Supplementary Table S2 (and the class templates in Fig. 3g and Supplementary Figure S2, left panel) may be attributed to the dual advantage of both Steps 1 and 2 of the new framework. Hierarchical clustering fails to capture the complexity of data when presented in the form of noisy curve profiles in which the true significance of features is not apparent. This is even more difficult if there are few significant features, which, in turn, might suffer from intersample variation. By addressing these issues, mAbprofiler produced robust mAb classes—specified as curves of probability density functions—even in the presence of non-Gaussian variation. It achieves this without any need for transforming the profiles or reducing them to points as required for hierarchical clustering.

The new framework has several additional advantages. Its use of Jump statistics provides a suitable criterion for optimal model selection in profile clustering. Each step of mAbprofiler can be performed independently with its own EM algorithm, which offers the flexibility of pipelining the framework with external algorithms. As output, not only does mAbprofiler produce a smooth profile for a mAb histogram, it also generates a mean template for the ‘signature’ pattern of every mAb class, along with parametric description of significant features therein. As a result, class templates can be archived, and later searched for information on overall or specific characteristics of such known mAb classes. Thus, it facilitates pattern matching with newly constructed mAb profiles, which can be grouped with classes having the most similar templates. Our computational and empirical validation of mAbprofiler classification shows how this is achieved. Another feature of our approach is that it does not require a clonal population. Expression can be analyzed both on individual cells and within a complex cell population. Moreover, our non-Gaussian model can be easily extended to temporal mAb profiling (Pyne et al., 2011), e.g. for measurements over the course of dampening of an inflammation in a certain tissue. The strength of mAbprofiler lies in providing a much-needed robust and objective framework for mAb characterization in different cell types and tissues.
Funding: TL thanks the National Science Council of Taiwan (NSC99-2118-M-005-001-MY2) for financial support.

Conflict of Interest: none declared.

REFERENCES

Pycr,S. et al. (2011) Parametric modeling of cellular state transitions as measured with flow cytometry. In Proceedings of IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCARS), Institute of Electrical and Electronics Engineers (IEEE), pp. 147–152.