**Automated Gating of Flow Cytometry Data via Adaptive Markov Random Fields Clustering[[1]](#footnote-1)**

**Running Headline:** Markov Random Fields Clustering for FCM Data

**Corresponding Author:**

Kevin Christopher Brosnan

Department of Mathematics and Statistics,

University of Limerick,

Limerick,

Ireland.

Email: [kevin.c.brosnan@ul.ie](mailto:kevin.c.brosnan@ul.ie)

Tel: +353 87 644 3584

Fax: +353 61 334927

**Co-Authors:**

Dr. Norma Bargary & Dr. Kevin Hayes

Department of Mathematics and Statistics,

University of Limerick,

Limerick,

Ireland.

Email: [norma.bargary@ul.ie](mailto:norma.bargary@ul.ie)

[kevin.hayes@ul.ie](mailto:kevin.hayes@ul.ie)

**Abstract**

***Keywords:*** statistics; gating; clustering; flow cytometry

**Introduction**

Flow cytometry (FCM) is now the leading state-of-the-art sensor technology used in food quality control and medical diagnostics. In the past ten years major advances have occurred in the instruments used to record FCM data, allowing fine cell analysis of up to twenty parameters (Rosa 2003). However, the analysis of collected data relies heavily on intuition rather than on a unified statistical framework (Eudey 1996). The increased volume and complexity of flow cytometry data boosts the demand for reliable statistical methods and accompanying software implementations to complete the analysis and draw meaningful conclusions from the data (Braylan 2004; Lizard 2007).

The analysis of FCM data is generally comprised of two key components (a) gating, involving the identification of homogeneous cell populations, and (b) tagging, identification of correlations between characteristics of the identified cell populations. The gating stage of analysis is traditionally a manual process whereby a computer mouse is used to draw gates around a region of interest on a 2D representation of a pair of FCM variables. This manual gating process based primarily on expert knowledge (Bagwell 2004; Parks 1997; Suni 2003) is time-consuming, highly subjective and limits the reproducibility of later analysis. Furthermore, the restriction of a 2D graphical representation as the gating strategy for high-dimensional FCM data is highly error-prone, as unexplored variables may provide additional information not evident in the chosen pair of variables (Lo 2008).

The identification of sub-populations in a large population, where each sub-populations members share a particular function is known as clustering in the statistical literature. Gating in FCM hopes to find homogenous cell populations within a large cell population analyses by the technology, as such the statistical approaches to automating gating follow a clustering methodology. Several of the methods utilised frequently in statistical clustering have been applied to the area of FCM data analysis. The k-means algorithm and its extensions has been attempted extensively (Bakker Schut 1993; Wilkins 2001; Lugli 2010), however classical k-means methods only allow cells to belong to one cluster. This hard clustering approach is restrictive in nature, as such a fuzzy k-means approach allowing cells to belong to different clusters with an associated probability was proposed (Rousseeuw 1996). Of more importance is the clustering criterion used in these approaches, which can impact the shape, size and orientation of identified sub-populations, thus failing to identify true cluster shapes in FCM. Many authors have attempted to use supervised learning algorithms for the automated gating procedure, such as neural networks (Kothari 1996; Boddy 2000) and support vector machines (Morris 2001), however these approaches require training data which reverts the basis of the methodology to expert opinion. More recently, a model-based clustering (Fraley 2002) approach utilising t-mixtures has been developed (Lo 2008). While model-based clustering has become a state-of-art in clustering high-dimensional data its application to FCM is inappropriate regardless of the distributional properties utilised. It is restrictive in its nature as it models each sub-population using the same distribution with varying location and scale parameters. The shape, size and orientation of sub-populations in FCM analysis appear not to follow a distribution and as such tackling FCM gating with a non-distributional methodology seems appropriate.

**Our approach…**

**Materials and Methods**

***Data Description***

Two publicly available datasets will be used to demonstrate the application of the proposed methodology to FCM. The use of these two datasets follows the work

To demonstrate our proposed methodology for the automated gating of FCM data we have utilised two FCM datasets, one publicly available dataset (Gasparetto 2004) and one from the University of Limerick Life Sciences Department.

**The Rituximab dataset.** Flow cytometry analysis was used in a drug-screening experiment to identify compounds that would augment the activity of Rituximab. The NCI diversity set compound library was used to select 1600 compounds distributed into duplicate 96 well plates. Both plates were incubated overnight using Daudi lymphoma cell line. Rituximab was then added to one duplicate plate and both plates were incubated for an additional number of hours. The samples contained cells treated with the compounds, untreated control cells and cells treated only with Rituximab. The cells were incubated with BrdU to identify newly formed DNA during the culture period, and were stained with anti-BrdU and 7-AAD following the culture period. FCM analysis was then used to identify the proportion of cells in various cell cycles.

**The GvHD dataset.**

***Markov Random Fields***

The statistical modelling and analysis of image data Markov Random Fields

***Cluster Identification***

***Clustering on preselected subsets***

The gating of FCM data on the entire sample of cells measured is rarely done in practice. Instead an initial selection of cells of interest is identified by cell size and shape, allowing the removal of dead cells or the removal of measurement artefacts found within the data. A projection of forward light scatter (FSC) against sideward light scatter (SSC) can identify these basic cell characteristics, shape and size. As such, we follow the expert knowledge in this paper and initially gate the population of interest using FSC and SSC. Utilising the identified sub-populations we apply our gating methodology to the variables corresponding to fluorescent markers used in the experiment. While this is the standard approach of cytometry experts, our methodology is not restricted to this approach and can be applied to all measured variables.

**Results**

**Discussion**

**Literature Cited**

1. De Rosa SC, Brenchley JM, Roederer M. Beyond Six Colors: A New Era in Flow Cytometry. Nat Med 2003;9:112-117.
2. Eudey TL. Statistical Considerations in DNA Flow Cytometry. Statistical Science 1996;11:320-334.
3. Braylan RC. Impact of Flow Cytometry on the Diagnosis and Characterization of Lymphomas, Chronic Lymphoproliferative Disorders and Plasma Cell Neoplasias. Cytometry A 2004;58A:57–61.
4. Lizard, G. Flow Cytometry Analyses and Bioinformatics: Interest in New Softwares to Optimize Novel Technologies and to Favor the Emergence of Innovative Concepts in Cell. Cytometry A 2007;71A:646-647.
5. Bagwell CB. DNA histogram analysis for node-negative breast cancer. Cytometry A 2004;58A:76-78.
6. Parks DR. Data processing and analysis: Data Management. In: Current Protocols in Cytometry. Robinson JP, editor. New York: Wiley; 1997. Chapters 10:10.1.1-10.1.6.
7. Suni MA, Dunn HS, Orr PL, De Laat R, Sinclair E, Ghanekar SA, Bredt BM, Dunne JF, Maino VC, Maecker HT. Performance of plate-based cytokine flow cytometry with automated data analysis. BMC Immunol 2003;4:9.
8. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry A 2008;73A:321–332.
9. Bakker Schut TC, Grooth BGD, Greve J. Cluster analysis of flow cytometric list mode data on a personal computer. Cytometry 1993;14:649-659.
10. Wilkins MF, Hardy SA, Boddy L, Morris CW. Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data. Cytometry 2001;44:210-217.
11. Lugli E, Roederer M, Cossarizza A. Data analysis in flow cytometry: The future just started. Cytometry A 2010;77:705-713.
12. Rousseeuw PJ, Kaufman L, Trauwaert E. Fuzzy clustering using scatter matrices. Comput Statist Data Anal 1996;100:2349-2356.
13. Kothari R, Cualing H, Balachander T. Neural network analysis of flow cytometry immunophenotype data. IEEE Trans Biomed Eng 1996;43:803-810.
14. Boddy L, Morris CW, Wilkins MF, Al-Haddad L, Tarran GA, Jonker RR, Burkill PH. Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. Mar Ecol Prog Ser 2000;195:47-59.
15. Morris CW, Autret A, Boddy L. Support vector machines for identifying organisms - A comparison with strongly partitioned radial basis function networks. Ecol Model 2001;146:57-67.
16. Fraley C, Raferty AE. Model-based clustering, discriminant analysis, and density estimation. J Am Statist Assoc 2002;97:611-631.
17. Gasparetto M, Gentry T, Sebti S, O'Bryan E, Nimmanapalli R, Blaskovich MA, Bhalla K, Rizziere D, Haaland P, Dunne J, Smitch C. Identification of compounds that inhance the anti-lymphoma activity of Rituximab using flow cytometric high-content screening. J Immunol Methods 2004;292:59-71.

1. Kevin C. Brosnan was supported by an Irish Research Council Government of Ireland Scholarship [grant number GOIPG/2014/19]. Kevin Hayes was supported by Science Foundation Ireland Research Investigator’s Award [grant number SFI-12/IA/1683]. [↑](#footnote-ref-1)