31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A Pipeline for Automated Analysis of Flow Cytometry Data: Preliminary Results on Lymphoma Sub-Type Diagnosis

Ali Bashashati, Kenneth Lo, Raphael Gottardo, Randy D. Gascoyne, Andrew Weng, Ryan Brinkman

Abstract—**Flow cytometry (FCM) is widely used in health research and is a technique to measure cell properties such as phenotype, cytokine expression, etc., for up to millions of cells from a sample. FCM data analysis is a highly tedious, subjective and manually time-consuming (to the level of impracticality for some data) process that is based on intuition rather than standardized statistical inference. This study proposes a pipeline for automatic analysis of FCM data. The proposed pipeline identifies biomarkers that correlate with physiological/pathological conditions and classifies the samples to specific pathological/physiological entities. The pipeline utilizes a model-based clustering approach to identify cell populations that share similar biological functions. Support vector machine (SVM) and random forest (RF) classifiers were then used to classify the samples and identify biomarkers associated with disease status. The performance of the proposed data analysis pipeline has been evaluated on lymphoma patients. Preliminary results show more than 90% accuracy in differentiating between some sub-types of lymphoma. The proposed pipeline also finds biologically meaningful biomarkers that differ between lymphoma sub-types.**

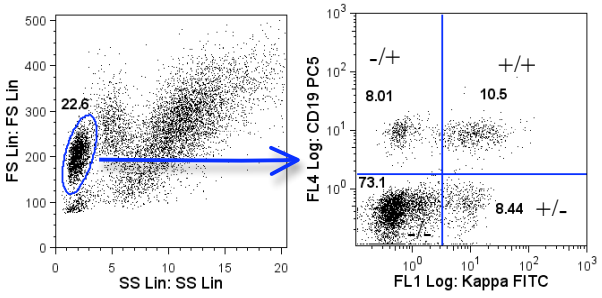
I. Introduction

FLOW cytometry (FCM) is widely used in health research and in treatment for a variety of tasks, such as in the diagnosis and monitoring of leukemia and lymphoma patients, providing the counts of helper-T lymphocytes needed to monitor the course and treatment of HIV infection, the evaluation of peripheral blood hematopoietic stem cell grafts, and many other diseases [1-5].

In FCM, intact cells and their constituent components are tagged with fluorescently conjugated monoclonal antibodies and/or stained with fluorescent reagents and then analyzed individually by a flow cytometer. In the instrument, hydrodynamic forces align the cells and the fluorescent molecules in/on each cell are excited by passing through the laser light at speeds exceeding 70,000 cells per second. Each cell passing through the beam scatters the light providing an indication of cell shape and size, and fluorescent chemicals found in the cell or attached to the cell may be excited into emitting fluorescent light to provide information on the physical and chemical characteristics of each individual cell. Typically, each data file generated by a flow cytometer contains measurements of cell properties (including phenotype, cytokine expression, and cell-cycle status) in up to 20 dimensions for each cell for up to millions of individual cells [1].

It is widely recognized that one of the limiting aspects of FCM technology is the analysis of the data [2, 5]. FCM data analysis involves two major components: (1) identifying homogeneous cell populations (traditionally known as gating) that share a particular biological function and (2) finding correlations between identified cell populations and clinical diagnosis or survival rate.

Typically, finding homogenous cell populations among FCM data involves selection of groups of cells based on the graphical representations of one or two characteristics of cells. Conditional on the selection of cell populations, further gating may be done using the other characteristics of cells. Figure 1 shows a two-stage gating example of an FCM data. At first, the cells that share common morphological properties (i.e., size and shape) are selected (e.g., cells in ellipsoidal region in Figure 1(a)). Cell populations are usually defined using a '+' or a '–' symbol to indicate whether a certain cell fraction expresses or lacks a specific molecule. Therefore, in the next step, the selected cells are examined to identify the ones that express certain markers by dividing the space into four quadrants representing '–/–', '+/–', '–/+', and '+/+' expressed cells (Figure 1(b), for example '+/+' cells in upper right quadrant of Figure 1(b) represent the cells that are '+' for both Kappa and CD19 markers). Properties of '+' and '–' expressed cells



(a) (b)

Figure. 1. Manual FCM data analysis procedure

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and Michael Smith Foundation for Health Research.

A. B. is with the British Columbia Cancer Research Center, Vancouver, Canada (email: abashash@bccrc.ca)
K. L. is with the statistics department of the University of British Columbia (email: c.lo@stat.ubc.ca)
R. G. is with the Institut des Recherches Cliniques de Montreal and the Biochemistry department of the university of Montreal (Raphael.gottardo@ircm.qc.ca)
R. D. G. and A. W. are with the British Columbia Cancer Agency, Vancouver, Canada and also Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada (email: rgascoyne@bcca.ca, aweng@bccrc.ca)
R. B. is with the British Columbia Cancer Research Center, Vancouver, Canada and also the Department of Medical Genetics, University of British Columbia, Vancouver, Canada (email: rbrinkman@bccrc.ca)

such as their percentages are then used to identify different pathological/physiological states.

Subsetting (gating) is typically accomplished "manually" by using proprietary software provided by instrument manufacturers. This approach is a highly tedious, subjective, and time-consuming (to the level of impracticality for some datasets) process that is based on intuition rather than standardized statistical inference [5]. In addition, finding correlations between identified cell populations and clinical diagnosis or survival rate is also performed manually and suffers from the same problems mentioned above. To date, only rudimentary statistical and bioinformatics tools exist to manage, analyze, present, and disseminate FCM data; yet, there is considerable demand for development of appropriate tools.

Finding cell populations in data in an automated fashion (automated gating) can be utilized by employing clustering algorithms. Various approaches such as k-means [6], neural networks [7], multidimensional binary trees [8], and model-based clustering [9] have been used in the context of gating FCM data. However, these approaches only focus on the first stage of FCM data analysis that identifies cell populations.

This paper proposes the first completely automatic FCM data analysis pipeline (i.e., one without manual intervention) that identifies FCM cell populations, facilitates disease diagnosis and identifies biomarkers that correlate with disease. The proposed data analysis pipeline uses a previously developed model-based clustering approach [9] for identification of cell sub-populations in FCM data and utilizes feature selection and classification approaches to identify biomarker changes and classify the samples.

As FCM has assumed an important role in the diagnosis of lymphoma [10, 11], we show preliminary results of applying this pipeline to differentiate between selected sub-types of lymphoma and identify biomarkers that contribute to differential classification of lymphoma sub-types.

Lymphoma is a cancer that originates from lymphocytes. According to the Revised European American Lymphoma (REAL) classification, i.e., the most recent classification of lymphoma, lymphoma can be divided into many sub-types based on morphology and cell lineage, each with differing prognosis. The REAL classification has received worldwide acceptance and is used by most haematopathologists and haemato-oncologists today. Folicular lymphoma (FOLL), mantle cell lymphoma (MCL), marginal zone lymphoma (MZL), small lymphocytic lymphoma (SLL), and diffuse large B-cell lymphoma (DLBCL) are examples of sub-types of lymphoma according to the REAL classification. Table 1 shows immunophenotypic features helpful in the differential diagnosis of the above-mentioned sub-types. For example, SLL differs from MZL in the expression of CD5 and CD23 markers. While SLL samples are CD5+ and CD23+, MZL samples are CD5- and CD23- [12].

TABLE 1. COMPARATIVE IMMUNOPHENOTYPIC SIGNATURE OF SELECTED SUB-TYPES OF LYMPHOMA

Marker	Selected Lymphoma Sub-Types for This Study				
	FOLL	MCL	MZL	SLL	DLBCL
CD20	+	+	+	+	+
CD5	−	+	−	+	−
CD43	−	+	−	+	+/−
CD10	+	−	−	−	−
CD23	−/+	−	−	+	+/−

II. MATERIALS AND METHODS

A. Data Description

FCM data generated from biopsies of lymph nodes of 438 lymphoma patients were available for analysis. These data were generated at the British Columbia Cancer Agency, Vancouver, Canada between 2002 and 2007.

Samples were divided into seven tubes and stained with different monoclonal antibodies conjugated with three fluorescent markers, namely, fluorescein isothiocyanate (FITC), phycoerythrin (PE) and phycoerythrin-Cy5 (PE-Cy5). Tube 1 contained CD45-FITC, CD14-PE and CD19-PE-Cy5 markers. Tube 2 contained isotype controls IgG1-FITC, IgG1/IgG2a-PE and IgG1-PE-Cy5 markers. Tube 3 contained CD10-FITC, CD11c-PE, CD20-PE-Cy5 markers. Tube 4 contained CD5-FITC, CD19-PE and CD3 PE-Cy5 markers. Tube 5 contained CD7-FITC, CD4-PE, CD8-PE-Cy5 markers. Tube 6 contained FMC7-FITC, CD23-PE and CD19-PE-Cy5. Tube 7 contained kappa-FITC, lambda-PE and CD19-PE-Cy5 markers. Each of the seven tubes of each sample were then run through a Beckman Coulter Cytomics FC500 flow cytometer to quantify the amount of antibodies on the cells.

B. Proposed Data Analysis Methodology

The main objective of this paper is to propose a fully automatic pipeline for FCM data analysis; a task that is currently performed manually. More specifically, this paper focuses on introducing a data analysis pipeline that is useful for differentiating between disease sub-types based on FCM.

Figure 2 shows the overall diagram of the proposed automatic data analysis pipeline. The details of each component are explained in the following.

Step 1: the model-based clustering approach adapted to identify cell populations in FCM data [9] is used to identify morphologically similar cell populations in the 2-dimensional (2-D) plot of forward light scatter (FSC) against sideward light scatter (SSC) (shown in Figure 3(a)). FSC and SSC parameters measure the size and granularity of each cell. The model-based clustering approach in [9] provides a unified framework to answer central questions such as: How many cell populations are there? How should we deal with outliers? These questions are fundamental to FCM analysis where one does not usually know the number of cell populations and where outliers are frequent. The method in [9] is based on *t*-mixture models with Box-Cox transformation to handle the issues of transformation selection and outlier identification, and uses the Bayesian

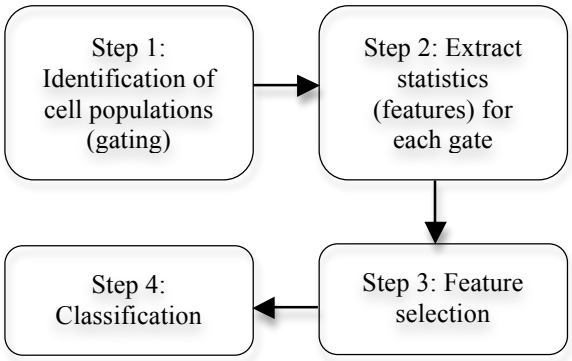


Figure 2. Overall diagram of the proposed automatic data analysis pipeline

Information Criteria (BIC) to determine the number of cell populations (i.e., clusters) present in the data.

Step 2: Once the cell populations (clusters) are identified, a specific cluster is picked and the cells in that cluster are observed in other dimensions (shown in Figure 3.b). The density distribution of data for each dimension is calculated by kernel density estimation approach (using Gaussian kernel). After smoothing, the location of the minimum of the density distribution is calculated. The locations of the minima are used as thresholds to divide the 2-D space into four quadrants representing '–/–', '+/–', '–/+', and '+/+' cells. Different features representing the percentage of cells in each quadrant, the mean of each quadrant, standard deviation across each dimension, etc are derived as features.

As we do not know which cluster in *"Step 1"* carries information about the label of the sample, e.g. disease status, the above procedure is repeated for each cluster and cluster combination. Furthermore, each of the seven tubes are analyzed following the above procedure resulting in generation of many features which represent the different characteristics of cell populations in each tube.

Step 3: Since the generated features in *"step 2"* may consist of some features that do not have discriminatory information, a feature selection scheme is used to discard the uninformative and redundant features. The output of this stage can either be used directly to identify biomarkers associated with disease diagnosis or used to label the samples (e.g., healthy vs. disease). For feature selection, we use maximum relevance minimum redundancy feature selection technique (MRMR) [13]. A common practice for feature selection is to select top rank features based on their relevance to their label (i.e., class). A deficiency of this simple ranking approach is that the selected features could be correlated among themselves. MRMR technique aims at not only selecting features that are relevant to the labels but also aims at reducing the redundancy of the selected features. In other words, this technique expands the representative power of the selected feature set by selecting features that are maximally dissimilar to each other and at the same time have high mutual information with the classes.

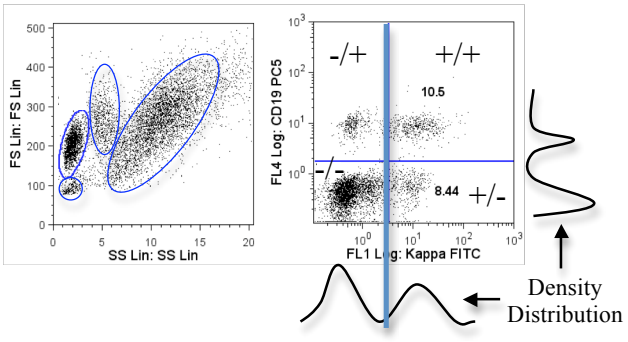


Figure 3. (a) Step 1 of data analysis, (b) Step 2 of data analysis

Step 4: Using the selected features in the previous step, a classifier is developed to label (classify) the samples. This stage, in fact, identifies any correlation between feature changes and label of the samples. In classification, a prediction model is built based on the known samples (referred to as the training set), which is used to make future predictions about unclassified samples. Two classification schemes based on Random Forest (RF) [14] and Support Vector Machine (SVM) [15] are implemented in this stage.

SVMs map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximizes the distance between the two parallel hyperplanes. Special properties of the decision surface ensure high generalization ability of the learning machine [15].

RFs are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [14].

III. Results

The proposed FCM data analysis pipeline is used to differentiate between sub-types of lymphoma. Using the model-based clustering approach in [9] and according to the Bayesian Information Criteria (BIC), four cell populations (clusters) which have similar size and shape (based on FSC and SSC parameters) were found. Using the identified cell populations, the features (as explained in *"Step 2"* of data analysis) are extracted. This procedure is repeated for each of the seven tubes and for different cluster combinations.

To evaluate the performance of the proposed pipeline, the patients are split into two randomly selected groups namely training- and test-sets. The data of training-set are used to select the informative features (by using MRMR feature selection technique) and train the classifier, while the data of the test set are used for performance evaluation. This procedure is repeated 100 times to reduce the bias of the results over a specific training-set.

Table 2 shows the mean performance of the proposed data analysis pipeline in differentiating between some lymphoma sub-types using SVM and RF classifiers. The performances of SVM and RF classifiers were not significantly different according to the student's t-test. The last column of Table 2 shows the top biomarkers identified by the proposed pipeline that contribute to the differentiation between lymphoma sub-types. For example, for differentiating between SLL and FOLL sub-types, from the 21 candidate markers analyzed, the developed pipeline automatically identifies features related to CD10, CD5 and CD23 as the markers with discriminatory information. These results are in line with established biological knowledge (see Table 1) and show strong evidence that the proposed FCM data analysis platform can extract biologically meaningful features from the data without manual intervention.

IV. Conclusions

In practice, gating and classification of FCM samples are performed manually. In this paper, preliminary results on the application of a completely automatic pipeline for FCM data analysis of lymphoma samples have been presented. The proposed data analysis platform automatically selects cell populations, extracts various features representing the identified cell populations, identifies informative features and eventually classifies the samples. Results of the evaluations of the proposed analysis platform show classification accuracies in the range between 80% and 95.3% in differentiating between some sub-types of lymphoma. More importantly, the proposed system identifies biologically meaningful biomarkers that differ between lymphoma sub-types.

The designed pipeline is, nevertheless, only capable of performing binary classification, i.e., it can differentiate between two sub-types of lymphoma. Our future direction would be expanding this pipeline to perform a multi-class classification task so that we can determine any sub-type of lymphoma from a FCM data. Moreover, further improvements are still needed to increase discrimination accuracy of the classifier. For example, results for differentiating between some sub-types is still around 80% that is not enough to be used as a diagnosis tool in practice. Meanwhile, the proposed data analysis platform can still be useful in the sense that it can guide the manual analysis procedure by suggesting the possible diagnosis of the sample.

TABLE 2. A TABLE DESCRIBING THE FEATURES THAT ARE EXTRACTED FROM THE FIRST STAGE

Discrimination task	Mean accuracy		Top selected markers
	RF	SVM	
SLL vs. Other Sub-types	89%	88%	CD23
SLL vs. FOLL	92.6%	92.0%	CD10, CD5, CD23
SLL vs. DLBC	90%	90%	CD23, CD10, CD5
SLL vs. MZL	88%	88%	CD23, CD5
SLL vs. MCL	95.3%	92.6%	CD23, CD19
MCL vs. DLBC	81%	80%	CD5, CD11c

In the proposed data analysis platform, finding and selecting cell populations among FCM data rely on automatic analysis of the 2-D representations of the multi-dimensional data. This process ignores the high-dimensional information of the FCM data, which can lead to missing biologically important cell populations. Future directions would be utilizing the multi-dimensional characteristics of the FCM data. Moreover, it is possible to incorporate a-priori biological information to guide the automated data analysis and hence improve the performance. For example, if the focus is only on lymphocyte cell population in a specific dataset, using biological information of the place of lymphocytes we can guide the algorithm to pick these cells.

Development of an automated FCM data analysis platform will greatly facilitate both basic research and clinical applications in medical/agricultural areas that depend upon this technique. This study is an initial step to demonstrate that a completely automatic FCM data analysis is possible. However, testing the performance of this platform on different FCM data is necessary in future studies.

References

[1] H. M. Shapiro, "The evolution of cytometers," *Cytometry,* vol. 58A, pp. 13-20, 2004.

[2] R. C. Braylan, "Impact of flow cytometry on the diagnosis and characterization of lymphomas, chronic lymphoproliferative disorders and plasma cell neoplasias," *Cytometry,* vol. 58A, pp. 57-61, 2004.

[3] R. L. Hengel and J. K. Nicholson, "An update on the use of flow cytometry in HIV infection and AIDS," *Clin. Lab. Med.,* vol. 21, pp. 841-856, 2001.

[4] F. L. Kiechle and C. A. Holland-Staley, "Genomics, transcriptomics, proteomics, and numbers," *Arch. Pathol. Lab. Med.,* vol. 127, pp. 1089-1097, 2003.

[5] C. B. Bagwell, "DNA histogram analysis for node-negative breast cancer," *Cytometry,* vol. 58A, pp. 76-78, 2004.

[6] T. C. Bakker Schut, B. G. De Grooth and J. Greve, "Cluster analysis of flow cytometric list mode data on a personal computer," *Cytometry,* vol. 14, pp. 649-659, 1993.

[7] D. S. Frankel, S. L. Frankel, B. J. Binder and R. F. Vogt, "Application of neural networks to flow cytometry data analysis and real-time cell classification," *Cytometry,* vol. 23, pp. 290-302, 1996.

[8] M. Bigos, D. R. Parks, W. A. Moore, L. A. Herzenberg and L. A. Herzenberg, "Pattern sorting: a computer-controlled multidimensional sorting method using k-d trees," *Cytometry,* vol. 16, pp. 357-363, 1994.

[9] K. Lo, R. R. Brinkman and R. Gottardo, "Automated gating of flow cytometry data via robust model-based clustering." *Cytometry A.,* 2008.

[10] J. B. Cousar, "Surgical pathology examination of lymph nodes. Practice survey by American Society of Clinical Pathologists," *Am. J. Clin. Pathol.,* vol. 104, pp. 126-132, 1995.

[11] A. Tbakhi, M. Edinger, J. Myles, B. Pohlman and R. R. Tubbs, "Flow cytometric immunophenotyping of non-Hodgkin's lymphomas and related disorders," *Cytometry,* vol. 25, pp. 113-124, 1996.

[12] A. Dogan, "Modern histological classification of low grade B-cell lymphomas," *Best Pract. Res. Clin. Haematol.,* vol. 18, pp. 11-26, 2005.

[13] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 27, pp. 1226-1238, 2005.

[14] L. Breiman, Random forests. *Mach. Learning*, vol. 45, pp. 5-32, 2001.

[15] C. J. C. Burges, A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.