Pattern recognition of multidimensional PBMC flow cytometry histograms for prostate cancer identification

Dong L. Tong[1] and Graham R. Ball[1]

[1] The John van Geest Cancer Research Centre, School of Science and Technology,
.Nottingham Trent University, Nottingham, United Kingdom, NG11 8NS.
{dong.tong, graham.balls}@ntu.ac.uk

Abstract. Flow cytometry is a technique that is used to count cells and to characterize property of the cells. In spite of enormous information content on the cells provided by flow cytometry, cytometry data is still analyzed based on step-by-step gating, either manually or automatically via bioinformatics. This paper presents a new strategy of interpreting cytometry data in a different manner. The proposed strategy utilizes clustering approach to identify cell population of interest and supervised approach to identify statistical significant cell regions in the population that can differentiate prostate cancer patients from the benign patients.

1 Introduction

Recent advancement in flow cytometry allowing up to 20 fluorochrome dyes to be generated by cytometer instrument [1-2]. As a result, this technique generates very complex data and requires bioinformatic techniques to simplify the data while preserving information on the data. Several studies have been made to reduce data complexity, emphasizing on improving the gating and visualization techniques on the cytometry data, using both supervised and unsupervised methods, on single or multidimensional fluorochrome dyes. These visualization/gating techniques are normally programmed as an add-on package for statistic programs such as BioConductor [1, 3-7] and Matlab software [8].

The common drawback of these techniques is that it requires programming knowledge in R or Matlab to perform. Furthermore, these techniques involve complex analysis steps to derive its conclusions, which could lead to high accumulated error rate, albeit, these errors are insignificant in each step. For example, Zare et al. [6] developed the SamSPECTRAL R package which requires pre-requisite knowledge in BioConductor software to operate the package. Jeffries et al. [8] designed an APT package based on Matlab software to visualize cytometry data which involves complex analysis steps, such as density detection, outlier removal, pixel removal, cluster smoothing and cluster breakpoint detection.

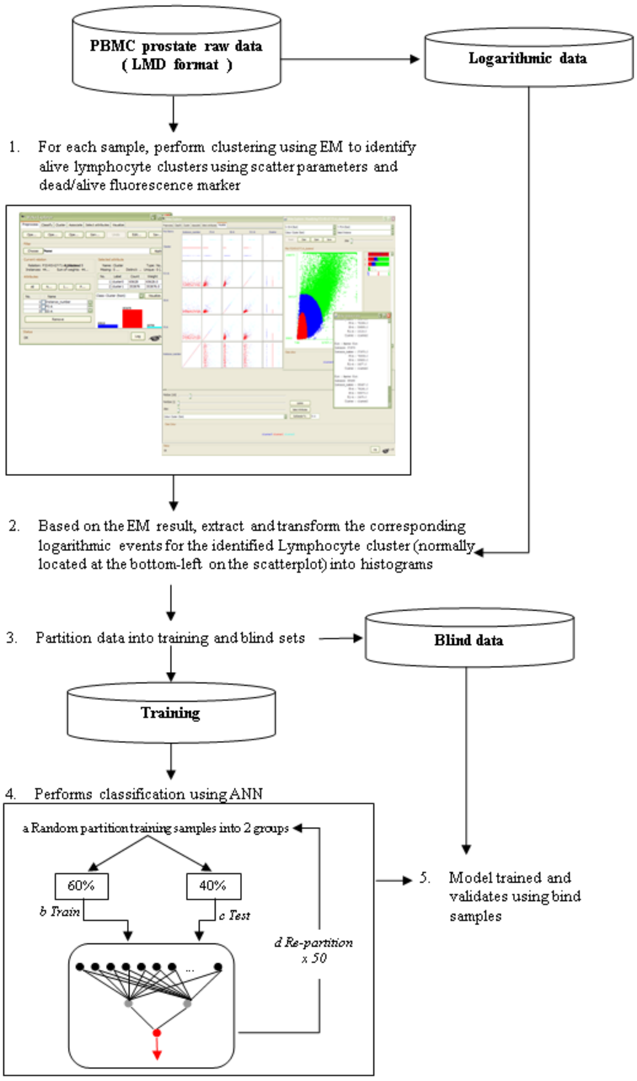
Unlike these conventional techniques, this paper presents a new strategy for interpreting cytometry data in a different manner. In the proposed strategy, the cell population of interest (i.e. lymphocyte cells, in this study) in each raw cytometry sample is first extracted using clustering method and the cell regions in the population are analyzed using supervised method. The strength of the proposed strategy is that it does not involve complex analysis steps to derive its result which is comparable to the conventional approach. This can reduce error rate incurred in complex analysis steps. In addition, no pre-requisite programming knowledge is required to operate the proposed strategy as we utilize clustering algorithm provided by Weka data mining platform [9]. The possible drawback of the proposed strategy is that the identified cell regions may not be easily related to biology explanation due to our present knowledge in flow cytometry is still limited on visualizing fluorescence wavelengths into 2 populations, i.e. positive and negative; than narrowing the focus into the cell regions in these populations.

The proposed strategy is examined using 10-color cytometry data that repeated in 3 different tubes, yielding 18 unique fluorescence markers, i.e. 6 common markers + (3 tubes x 4 unique markers). These data were clinically diagnosed as prostate cancer and prostate benign.

The rest of the paper is organized as follows. Section 2 describes our approach. Section 3 presents the experiment data. The results are presented in Section 4. Finally in Section 5, the conclusion is drawn with discussions.

2 Implementation

The aim of this paper is to identify a group of events (i.e. cell regions) that can differentiate 2 types of prostate disease. The proposed strategy involves 3 main components, which are cluster extraction, data transformation and class prediction. Fig. 1 below presents the schematic work of our approach.



PBMC prostate raw data
(LMD format)
Logarithmic data
1. For each sample, perform clustering using EM to identify alive lymphocyte clusters using scatter parameters and dead/alive fluorescence marker
2. Based on the EM result, extract and transform the corresponding logarithmic events for the identified Lymphocyte cluster (normally located at the bottom-left on the scatterplot) into histograms
3. Partition data into training and blind sets
Blind data
Training
4. Performs classification using ANN
a Random partition training samples into 2 groups
60%
40%
b Train
c Test
d Re-partition
x 50
5. Model trained and validates using bind samples

Fig. 1. Schematic work for cytometry analysis.

The cytometry data in the LMD format were first retrieved from cytometer instrument. Using the scatter parameters (i.e. forward and side scatters), the WEKA expectation maximization (EM) clustering algorithm was then applied to cluster lymphocyte cells for each cytometry sample using the raw values (i.e. first dataset in the LMD file). Based on the identified lymphocyte clusters from the first dataset, the corresponding log values for each fluorescence parameter for the lymphocyte cells (i.e. second dataset in the LMD file which is similar to FCS data) were extracted and the dead lymphocyte cells were then removed using side-scatter and dead/alive fluorescence parameters. Finally, the log cytometry data were then analyzed using artificial neural network (ANN).

2.1 Expectation Maximization

Expectation maximization (EM) is a probabilistic clustering approach that uses maximum likelihood estimation, coupled with the probability computation for each possible solution (i.e. cluster) for the observation (i.e. cytometry event), to predict the likelihood of the possible solutions where an observation should belongs. The EM algorithm iterates between the steps of estimating a probability distribution for the clusters (known as E-step) on the current model and re-estimates the model parameters using new probability distribution (known as M-step). The EM algorithm has been adopted in several cytometry studies [3-4] as the preliminary step to discover cell populations in each cytometry sample.

The reason for choosing EM from the Weka data mining suite in this study is due to it can be implemented directly on the LMD data without any prior preprocessing step and programming knowledge required. In addition, the EM algorithm in Weka provides flexibility on auto-clustering (i.e. detecting possible number of clusters in each sample) and manual-clustering (i.e. user can pre-define maximum number of clusters in each sample).

In this paper, the default setting in the EM algorithm was applied, except that the random seeding parameter was increased from 100 to 1000 and the maximum number of clusters was pre-set to 3.

2.2 Logarithmic Data

There are several studies looking into the deficiency of logarithmic transformation could lead to serious misinterpretation of cytometry data [10-11] and proposed a new statistics procedure to overcome this deficiency [12-13]. However, there is no standard procedure on the efficient transformation technique should be used in cytometry data analysis and is subject to individual research groups.

The logarithmic data provided in the LMD file was used in this study due to the EM algorithm has eliminated non-lymphocyte cells from being extracted for log transformation. The new statistic procedure aiming to alleviate logarithmic deficiency in visualizing negative and zero valued data (which will appeared at the edges of the log scatterplot). Since the EM algorithm has eliminated most of these “debris” from lymphocyte cells, there is no necessary for using new transformation techniques in this study. In addition, majority of the flow cytometry users still prefer to visualize cytometry data with logarithmic display and the logarithmic data in the LMD files fulfilled FCS standard.

2.3 Artificial Neural Network

Artificial neural networks (ANNs) have been used for classification in flow cytometry [14-15]. In this paper, a 3-layered backpropagation neural network, coupled with stepwise search procedure to identify the significant cell regions in discriminate cytometry samples. The model was trained with training set (i.e. 60% training samples) and tested using the testing set (i.e. 40%). To avoid any bias on the reported results, samples were re-shuffled 50 times in each training, testing and validation sets. The trained model was then further validated using another new set of blind samples.

For the ANN architecture, as the stepwise search procedure was applied, an increment of 1 node in the input layer each time a new network model was created and each input node represent a unique cell region in a specific florescence histogram; 2 hidden nodes in the hidden layer and an output node indicated the class of the samples.

3 Prostate Cancer Dataset

This study used the peripheral blood mononuclear cell (PBMC) collected from 20 patients which were clinically confirmed with prostate disease as either cancerous or benign. The data were collected on the Beckman Coulter Gallios instrument that able to collect data in 10 fluorescence wavelengths as well as forward and side scatters. Each sample was prepared in 3 different tubes with 10 fluorescence markers in each tube, yielding 16 unique markers (6 common markers and 4 unique markers in each tube). All data storage was at a resolution of 1024 x 1024 with 1024 channels.

Among 20 clinically diagnosed prostate samples, 14 were clinically diagnosed as benign prostate, 6 were cancerous with different Gleason scores, ranging from Gleason scores 7 to 9. Ten samples (i.e. 6 benign and 4 prostate cancer) were randomly selected to train/test the model. The remaining 10 samples were kept separately and used as blind validation set.

The 18 unique fluorescence markers used in this study were CD3, CD25, FoxP3, CD4, Dead/Alive, CD49d, CD127, CD39, BTLA, PD1, Lag3, CD62L, CCR7, CD45RA, GITR, CCR3 and ICOS. The average log values of the common markers CD3, CD25, FoxP3 and CD4 were used for analysis in this study.

4 Results

Amongst 17 unique fluorescence markers (excluding the dead/alive marker), the marker CCR7 has been predominantly selected by the ANN model in this study. By looking into the top 100 regions selected by the ANN model, 72 regions were selected from CCR7+, followed by 15 regions from CD4+, 2 from FoxP3-, 8 from FoxP3+, 2 from BTLA+ and 1 from CD127+. Table 1 shows the summary results of the identified markers based on the top 100-ranked regions. The overall test predictive error for the top 100 selected region is < 0.07 and the p-value for the selected markers is $p < 1 \times 10^{-6}$. The classification result is depicted in Table 2.

Table 1. The summary result for the identified marker based on the top-100 ranked regions.

Marker	Total number of regions in the dataset	Total number of selected regions
CCR7	330	72
CD4	368	15
CD127	454	1
FoxP3	550	10
BTLA	*422*	2

Table 2. The classification results for 20 blind samples.

Samples	Target	CCR7	CD4	CD127	FoxP3	BTLA
P35	**Prostate**	Prostate	Benign*	Benign*	Benign*	Benign*
P53	**Prostate**	Prostate	Prostate	Benign*	Prostate	Benign*
P113	**Benign**	Benign	Benign	Benign	Benign	Benign
P114	**Benign**	Benign	Benign	Benign	Prostate*	Benign
P116	**Benign**	Benign	Benign	Benign	Benign	Benign
P139	**Benign**	Benign	Benign	Benign	Prostate*	Benign
P144	**Benign**	Benign	Benign	Benign	Prostate*	Benign
P42	**Benign**	Benign	Benign	Benign	Prostate*	Benign
P44	**Benign**	Benign	Benign	Prostate*	Benign	Benign
P45	**Benign**	Prostate*	Benign	Benign	Prostate*	Prostate*

* indicates misclassification.

All the selected regions, except 3 regions in FoxP3, were belong to positive population on the fluorescence histogram displays. The distribution of cell positive in FoxP3 is sample-dependent, as these regions seemed to be the cutoff areas for 2 populations (i.e. negative and positive) in the FoxP3. On the histogram display for FoxP3, clear populations cutoff on benign patients are likely to be happened at log 380-390, while at log >400 on cancer patients.

Among the selected cell regions, 6 were found in markers CD8+, 3 in FoxP3 and 1 in CD4+. This indicates that CD8+ is a strong predictor marker, statistically speaking, followed by FoxP3, to differentiate prostate patients from the benign one, than CD4+.

Among the selected cell regions in CD8+, all regions show clear separation between cancer and benign groups with the presence of lymphocyte cells (not detectable or low) in either group. Meanwhile, 2 out of 3 regions selected in FoxP3 show a moderate number of cells presented in benign patients but is not detected in cancer patients, indicates that these 2 regions may be a better prognosis features for benign patients than cancer patients, however, experiment on these region on healthy patients will be performed in near future to confirm whether they are group-dependent.

These selected regions were further validated using 10 blind samples and results showed that these regions have high prediction power with 2 misclassifications, i.e. prostatitis and healthy samples. It is not surprised that the ANN model misclassified prostatitis and healthy blind samples as these disease patterns were never presented in the training process.

It is important to note that due to the high dimension of channels and parameters used in this study (i.e. 1024 channels x 9 fluorochromes), there is a potential of different sets of fluorochrome/region combinations may deliver better, worse or equivalent performance to the reported region/fluorochrome in this study. Due to the low supply on cancerous samples, the network was trained with 12 samples, consequently, may yield optimistic results, from statistic perspective, albeit, the identified regions were further validated using blind samples.

5 Discussions

This study demonstrated the feasible use of this new strategy in interpreting cytometry data from the conventional manner. The use of EM and ANN are able to distinguish prostate patients, cancerous and benign, from PBMC flow cytometry histograms characterized with multi-parameters (up to 11 parameters including scatters parameters) and multi-channels (up to 1024 channels), with better, at least comparable, performance as the conventional approach. The strength of our method is the removal of user prejudice on certain markers and thus, all the parameters in the data are equally treated in the analysis. Furthermore, this method can be operated by users with little knowledge in programming and can be easily implemented with any classifiers, customized classifier or publicly available classifier.

The proposed strategy adopted similarity concept as presented by Ravdin at al. [14] on the study of S-phase for breast cancer patients. The differences between the proposed strategy and Ravdin et al. are the latter analyzed single-dimension histogram based on prior manual gating technique with only 32 channels and emphasized on the regression process of S-phase. The proposed strategy, on the other hand, looking into multi-parameters and multi-channels. Rather than using manual gating approach which is user subjective, this study applied clustering approach to isolate lymphocyte cells from ‘debris’ and identified cell regions that are, statistically, significant to differentiate benign and prostate cancer patients based on 9 fluorochromes.

The drawback of our method is lay on the supervised classifier, which was not able to identify unknown sample or the sample that was never been presented in the learning process. The other possible drawback of our method is the identified regions/parameters may not contain much biological information as compared to the identified parameters using gating approach; however, it is very much user subjective. We believe that our method has fulfilled, to some extent, requirement as a preliminary diagnostic tool for confirming the status of the patients as either benign or cancerous, so that the correct treatment can be designed for different disease groups.

Acknowledgment

The authors wish to thank Dr. S. McArdle from the John van Geest Cancer Research Centre (JvGCRC) for preparing FCM data. This work was financial supported by The John and Lucille van Geest Foundation.

References

1. Frelinger, J., Kepler, T.B., Chan, C.: Flow: Statistics, visualization and informatics for flow cytometry. Source Code Biol Med. 3 (2008) 10
2. Lugli, E., Roederer, M., Cossarizza, A.: Data analysis in flow cytomerty: The future just started. Cytometry A 77 (2010) 705-713
3. Lo, K., Brinkman, R.R., Gottardo, R.: Automated gating of flow cytometry data via robust model-based clustering. Cytometry A 73 (2008) 321-332
4. Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T-I., Maier, L.M., Baecher-Allan, C., McLachlan, G.J., Tamayo, P., Hafler, D.A., De Jager, P.L., Mesirov, J.P.: Automated high-dimensional flow cytometric data analysis. PNAS 106 (2009) 8519-8524
5. Frelinger, J., Ottinger, J., Gouttefangeas, C., Chan, C.: Modeling flow cytometry data for cancer vaccine immune monitoring. Cancer Immunol Immunother. 59 (2010) 1435-1441
6. Zare, H., Shooshtari, P., Gupta, A., Brinkman, R.R.: Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC Bioinformatics 11 (2010) 403
7. Naumann, U., Luta, G., Wand, M.P.: The curvHDR method for gating flow cytometry samples. BMC Bioinformatics 11 (2010) 44
8. Jeffries, D., Zaidi, I., de Jong, B., Holland, M.J., Miles, D.J.: Analysis of flow cytometry data using an automatic processing tool. Cytometry A 73 (2008) 857-867
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: WEKA data mining software: An update. ACM SIGKDD Explorations 11 (2009) 1-18
10. Herzenberg, L.A., Tung, J., Moore, W.A., Herzenberg, L.A., Parks, D.R.: Interpreting flow cytometry data: A guide for the perplexed. Nature Immunology 7 (2006) 681-685
11. Novo, D., Wood, J.: Flow cytometry histograms: Transformations, resolution, and display. Cytometry Part A 73A (2008) 685-692
12. Bagwell, C.B.: Hyperlog – A flexible log-like transform for negative, zero, and positive valued data. Cytometry Part A 64A (2005) 34-42
13. Parks, D.R., Roederer, M., Moore, W.A.: A new “logical” display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. Cytometry Part A 69A (2006) 541-551
14. Ravdin, P.M., Clark, G.M., Hough, J.J., Owens, M.A., McGuire, W.L.: Neural network analysis of DNA flow cytometry histograms. Cytometry 14 (1993) 74-80
15. Boddy, L., Wilkins, M.F., Morris, C.W.: Pattern recognition in flow cytometry. Cytometry 44 (2001) 195-209