Analysis and Classification of Lymphoma Sub-types

The Terry Fox Research Institute
L'Institut de recherche Terry Fox

Adrin Jalali, Nima Aghaeepour, Kieran O'Neill, Andrew P. Weng, Ryan R. Brinkman

University of British Columbia and British Columbia Cancer Agency - Terry Fox Laboratory

{ajalali,naghaeep,koneill,aweng,rbrinkman}@bccrc.ca

Abstract

*Analysis of high-dimensional flow cytometry datasets can reveal novel cell populations that were not previously envisioned and with poorly understood biology. Following discovery, characterization of these populations in terms of the critical markers involved is an important step, as this can help to both better understand the biology of these populations and aid in designing simpler marker panels to identify them on simpler instruments and with fewer reagents (*i.e., *in resource poor or highly regulated clinical settings). We developed a computational tool that constructs cellular hierarchies by combining automated gating with dynamic programming and graph theory to provide the best gating strategies to identify a target population to a desired level of purity or correlation with a clinical outcome, using the simplest possible marker panels. The pipeline was executed on lymphoma 8-color dataset to find immunophenotypes correlating with differentiation of diffused large B-cell lymphoma and follicular lymphoma.*

1. Problem Definition and Pipeline

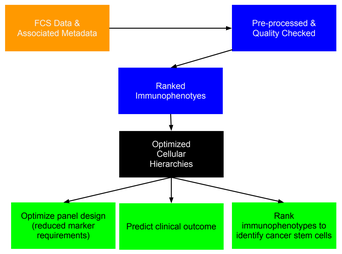


Figure 1: *Automated data analysis pipeline*

- To optimize cellular hierarchies, the problem is modelled as a minimization problem, and then multiple minimum weight paths finding problem.

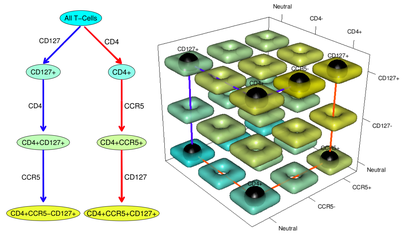


Figure 2: *Dynamic space and paths*

- Dynamic programming recursive function:

$$T^*(P^k) = \begin{cases} -S(P^k) & \text{if } k = 1 \\ \min\{T^*(P^k \setminus P_i^k) - S(P^k) | i = 1, \ldots, k\} & \text{otherwise} \end{cases}, \quad (1)$$

- Dynamic programming complexity: $O(m \times 2^{m-1})$
- Multiple path search in graph complexity:

$$\begin{aligned} O(e + v + l) &= O(m \times 2^{m-1} + 2^m + l) \\ &= O(m \times 2^{m-1} + 2 \times 2^{m-1} + l) \\ &= O((m + 2) \times 2^{m-1} + l). \end{aligned} \quad (2)$$

2. Lymphoma Cell Hierarchy Analysis

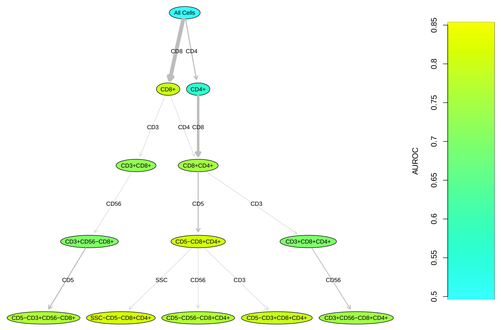


Figure 3: *Important cell populations and their detected hierarchies using the tube including $CD8$ of the dataset. This tube includes these markers: FSC, SSC, $CD5$, $CD3$, $CD57$, $CD7$, $CD2$, $CD56$, $CD8$, $CD4$*

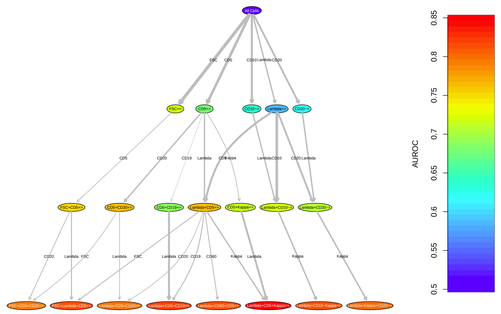
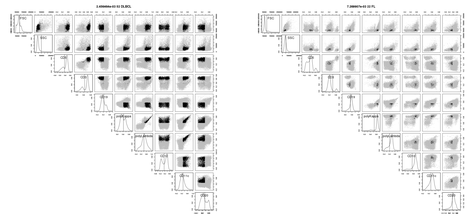


Figure 4: *Running the analysis on the tube including $kappa$ and $lampda$ results this structure. As illustrated, it seems that there are some $kappa^+lambda^+$ cells that are powerful in discriminating DBLCL from FL. DLBCL samples tend to have more of those cells. This tube includes FSC, SSC, $Lambda$, $CD80$, $CD5$, $CD10$, $Kappa$, $CD20$, $CD3$, $CD19$*

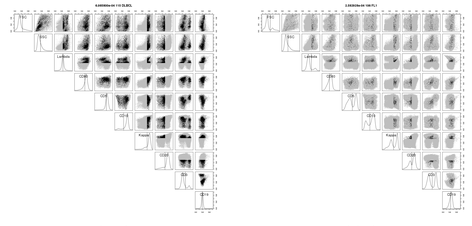
3. Sample Detected Cell Populations

- In the following plots, the left plot is a diffused large B-cell lymphoma sample, and the right plot is a follicular lymphoma sample. Black dots represent the immunophenotype mentioned in the caption of each plot.



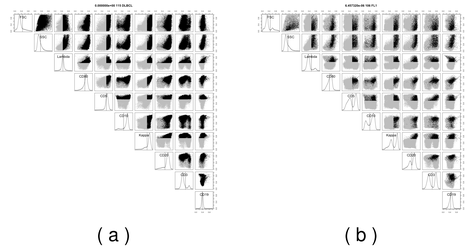
(a) (b)

Figure 5: $CD5^+CD19^-polyKappa^+polyLambda^+CD20^-$



(a) (b)

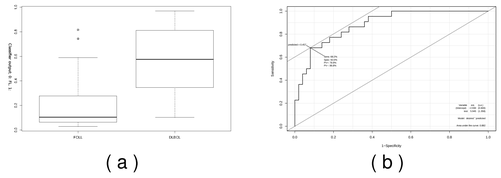
Figure 6: $Kambda^+Kappa^+CD20^-$



(a) (b)

Figure 7: $Lambda^+Kappa^+CD5^+$

4. DLBCL/FL Classification Results



(a) (b)

Figure 8: *Classification Performance - (a) plot illustrates the boxplot of the classification output; (b) shows the Receiver Operating Characteristic curve.*

5. Discussion

The dataset of study came from British Columbia Cancer Agency. The dataset includes 76 samples, all from 2009, for each of which, 3 8-color flow-cytometry experiments are done.

It is possible to run the same pipeline on other problems, for example, finding important cell populations in differentiating normal samples from Hodgkin's lymphoma. It is also possible to use other datasets, for instance, leukemia, and run the pipeline to classify different leukemia subtypes.

Misclassified patients are currently being clinically reviewed to gain further insight to the possible causes for the misclassification and potentially improve the classifier. The method is being expanded to all subtypes so that it can be used as part of routine standard of care to identify individual cases that are candidates for review (i.e., those cases that the algorithm predicted differently that the diagnosis). We hypothesize these individuals would be of higher value for review than a randomly selected patients.

All software packages are available for R on the web and accessible from Bioconductor repository. Immunophenotype ranking is done using flowType package [1,2], and optimization of cell hierarchies is done using RchyOptimyx package [3].

References:

1. Nima Aghaeepour and Pratip K. Chattopadhyay and Anuradha Ganesan and Kieran O'Neill and Habil Zare and Adrin Jalali and Holger H. Hoos and Mario Roederer and Ryan R. Brinkman. Early Immunologic Correlates of HIV Protection can be Identified from Computational Analysis of Complex Multivariate T-cell Flow Cytometry Assays. Bioinformatics, accepted for publication., 2012:28(7):1009-1016.

2. http://www.bioconductor.org/packages/release/bioc/html/flowType.html

3. http://www.bioconductor.org/packages/2.10/bioc/html/RchyOptimyx.html

Acknowledgment: This project was supported by Michael Smith Foundation for Health Research Scholar Award to RRB and AW, a Terry Fox new Investigator award to RRB, NIH/NIBIB grant R01 EB008400, the Terry Fox Foundation and the Terry Fox Research Institute.