A Framework for High-Throughput Flow Cytometry Data Analysis and its Application to Diagnosis and Discovery

Ryan Brinkman, Medical Genetics, UBC

Terry Fox Laboratory
BC Cancer Agency, Vancouver, BC, Canada

10-Sep-2009
Why Flow Cytometry?

- Technique for counting and examining microscopic particles suspended in a stream of fluid
- Widely used in clinical medicine and basic research for diagnosis and discovery
  - cancer, immunity, stem cells
- Ideally suited for high throughput assays of cells in solution (e.g., blood)
  - 50,000 cells/second, for a million cells per sample
Flow Cytometer Schematic

Flow Cytometry
QA and Normalization
Clustering and Labeling Cells
Analysis Example
Conclusions
Acknowledgements
Fluorescence Intensity Per Cell Corresponds to Proteins Detected by Conjugated Antibodies
Data Analysis Can Be Complicated

High Throughput Flow Experiments

• 96-well plates
• Automated
  • 1,000+ samples/day
  • 300,000+ cells/sample
  • 6+ parameters (dimensions)/event
• The ability to produce data is outstripping the ability to analyze both the amount and complexity of data generated

• Bioinformatics to the rescue!
  • But first we need to develop some tools ...
Data Analysis Pipeline

Raw Data → Quality Assessment → Normalized Data


Population Statistics → Sample Classification → Diagnosis & Discovery

Tools:
- flowQ & plateCore
- flowUtils & flowCore
- fdaNorm
- gaussNorm
- flowClust/Merge
- SamSPECTRAL
- mclust
- randomForest
- e1071
- flowViz
- flowQ & plateCore
- flowUtils & flowCore
- fdaNorm
- gaussNorm
- flowClust/Merge
- SamSPECTRAL
- mclust
- randomForest
- e1071
- flowViz
Automated Analysis of Flow Cytometry Data

- R is an open source (free as in beer & free as in speech) robust statistical programming environment for Windows, Mac & Linux which offers a wide range of statistical and visualization methods
- Bioconductor provides R software modules for biological and clinical data analysis
- A **scripted** approach to high throughput data analysis

www.r-project.org
www.bioconductor.org
Getting Started: Coercing Data & Assessing Quality

Hahne et al. *BMC Bioinformatics* 2009
Strain et al. *Advances in Bioinformatics* 2009
Manual Analysis
- Import raw FCS files
- Per well compensation
- Select population(s) of interest
- Per plate quality control
- Set control gates
- Score test samples
- Annotation and analysis

plateCore Workflow
- \texttt{pbmcFP <- flowPlate(pbmFP,wellAnnotation,plateName="PBMC.001")}
- \texttt{pbmcFP <- compensate(pbmFP,compensation.matrix)}
- \texttt{pbmcFP <- Subset(pbmFP, rectangleGate("FSC-H"=c(300,700),"SSC-H"=c(50,400)))}
- \texttt{ecdfplot(~`FSC-H`|as.factor(Row.Id),plateSet(pbmFP))}
- \texttt{pbmcFP <- setControlGates(pbmFP, gateType="Negative.Control")}
- \texttt{pbmcFP <- applyControlGates(pbmFP)}
- \texttt{pbmcFP <- summaryStats(pbmFP)}
Why Assess Quality?

- Detect systematic and stochastic effects that are not likely to be biologically motivated
- Systematic errors often indicate the need for adjustments in sample handling or processing
- Aberrant samples should be identified & potentially removed from downstream analyses to avoid spurious results

We developed a variety of Exploratory Data Analytic (EDA) tools (graphical methods) for exploring ungated FCM data in a time and cost effective manner
Flow Cytometry
QA and Normalization
Clustering and Labeling Cells
Analysis Example
Conclusions
Acknowledgements

flowQ: Quality Checking
## flowQ: Summary web page

<table>
<thead>
<tr>
<th>flow set details</th>
<th>time line</th>
<th>time flow</th>
<th>cell number</th>
</tr>
</thead>
<tbody>
<tr>
<td>s5a01</td>
<td>FSC-H</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5a02</td>
<td>FSC-H</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5a03</td>
<td>FSC-H</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5a04</td>
<td>FSC-H</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5a05</td>
<td>FSC-H</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5a06</td>
<td>FSC-H</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s5a07</td>
<td>FSC-H</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Frames**: 1-14, 15-28, 29-30
Data Normalization

- Quality Assessment
- fdaNorm
- guassNorm
- Normalized Data

Hahne et al. *Cytometry A* 2009
Data Normalization

- Between-sample variation poses a significant challenge for analysis of large scale data sets
  - Hard to match biologically relevant cell populations across samples (technical variation in sample acquisition, instrumentation differences)
- Therefore, remove technical between-sample variation by aligning prominent features (landmarks) in the raw data on a per-channel basis.
Data Normalization Schematic

- Identification
- Classification
- Alignment

Sample 1
Sample 2
Sample 3
Sample 4
Data Normalization

- raw data
- gaussNorm
- fdaNorm
Data Normalization

![Manual Gates](manual_gates.png)

![Static Gates](static_gates.png)

![Static Gates gaussNorm Normalized](static_gates_gaussNorm_normalized.png)

![Static Gates fdaNorm Normalized](static_gates_fdaNorm_normalized.png)
Data Normalization

<table>
<thead>
<tr>
<th>CD5−SS</th>
<th>CD5+CD3+</th>
<th>CD5+SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>p = 0.95</td>
<td>p = 0.96</td>
<td>p = 0.97</td>
</tr>
<tr>
<td>raw data</td>
<td>gaussNorm</td>
<td>fdaNorm</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CD3−SS</th>
<th>CD3+SS</th>
<th>CD5−CD3−</th>
</tr>
</thead>
<tbody>
<tr>
<td>p = 0.99</td>
<td>p = 0.99</td>
<td>p = 0.99</td>
</tr>
<tr>
<td>raw data</td>
<td>gaussNorm</td>
<td>fdaNorm</td>
</tr>
</tbody>
</table>
Automated Gating

Outlier Removal → flowClust/Merge → SamSPECTRAL → Population Identification [Gating]

Lo et al. *Cytometry A* 2008
Finak et al. *Advances in Bioinformatics* 2009
Shooshtari et al.  ? 2009
Data Analysis: Different Analysis Problems Require Different Solutions

Two fundamentally different problems in automated flow analysis:

Diagnosis vs. Discovery

How we develop automated approaches to analyze data depends on the what the underlying goal is:

Find the same subset of populations in every sample

vs.

Find “every” population in every sample
Automated Gating (Take I): flowClust

Finding cohesive groups (cell sub-populations) in data in an automated fashion can be addressed through a methodology termed clustering.

Several problems need to be overcome:

- Multiple dimensions are hard to think in
- Difficult to visualize
- Computationally challenging
Automated Gating (Take I): flowClust

We developed a model based clustering approach that automates the process of:

- Identifying how many sub-populations
  - Use the Bayesian Information Criterium (BIC)
- Dealing with outliers
  - Gaussian distributions for cell populations fail due to outliers
  - Therefore, replace Gaussian distribution with a \( t \)-mixture model using BoxCox transformation
Manual Analysis vs flowClust
Automated Gating (Take 2): flowMerge

- Estimating the number of clusters is hard
- BIC and AIC tend to overestimate the number of clusters
- Integrated Complete Likelihood (ICL) is an entropy penalized BIC criterion (models with overlapping components are penalized by a larger entropy due to the overlap)
- But - ICL tends to underestimate number of clusters

Combine BIC and ICL!
flowMerge vs flowClust

A

ICL solution

B

BIC solution

D

Merged solution

CD4 PE

CD7 FITC

CD8 PC5

CD4 PE

CD7 FITC

CD8 PC5

CD4 PE

CD7 FITC

CD8 PC5

CD4 PE

CD7 FITC
Automated Gating (Take 3): SamSPECTRAL

• Stop modeling populations by their supposed(!) shape
• Instead, model connections between cells based on closeness (connectivity)

• Stop trying to guestimate the number of clusters
• Instead, cut populations where there are few connections

• But, spectral clustering computationally impractical for flow cytometry data (>3,000 points)

Combine spectral clustering with density based sampling approach to preserve rare populations
SamSPECTRAL Algorithm

1. Set neighborhood threshold $h$;
2. Label all the data points as unregistered;
3. Pick a random unregistered point $p$ and find all unregistered data points within distance $h$ from $p$;
4. Put all of these points in a set called community $p$, and label them as registered;
5. Repeat above 2 steps until no unregistered points;
6. Compute similarity between communities;
7. Build graph with communities as vertices, edges weighted by similarity;
8. Run classical spectral clustering;
9. Combine clusters if $\frac{\text{connectivity between clusters}}{\text{connectivity within cluster}} \geq \text{separation factor}$.
flowMerge vs. SamSPECTRAL
Estimating Clusters: Intersecting Lines the Best Way?

BIC and ICL

Eigenvalues and $y=1$
Estimating Clusters: Recursive Kernal Density Estimates
Population Labeling

Population Identification [Gating] \[\xrightarrow{\text{mclust}}\] Populations Labelled
Population Labeling

Finding Similar Cell Populations

Marker 2

Marker 1

Marker 1

Marker 1

Marker 1

Patient 1

Patient 2

Patient 3

Patient 4
Population Labeling

Finding Similar Cell Populations

Marker 2

Marker 1

Marker 2

Marker 1

Marker 2

Marker 1

Marker 2

Marker 1

Patient 1

Patient 2

Patient 3

Patient 4
Population Labeling

Finding Similar Cell Populations

Marker 2

Marker 1

Patient 1

Patient 2

Patient 3

Patient 4
Population Labeling

Marker 2

P1 P3 P2

P4

Marker 1

P1 P3 P2

P4 P3 P2
Data Analysis Pipeline

Raw Data
- flowQ & plateCore
- flowUtils & flowCore

Quality Assessment
- fdaNorm
- gaussNorm

Normalized Data

Outlier Removal
- flowClust/Merge
- SamPECTRAL

Population Identification [Gating]
- mclust

Populations Labelled

Population Statistics
- randomForest
- e1071

Sample Classification
- flowViz

Diagnosis & Discovery
All available in High Throughput in GenePattern*
Diffuse Large B-Cell Lymphoma

Biology:
- Most common type of lymphoma
- Outcome (survival) is highly variable
  - Half of the patients survive less than 5 years (why?)

Hypothesis:
- Automated clustering of flow data will identify novel cell populations that correlate with outcomes (e.g., survival)
- This will lead to better understanding of DLBCL, its diagnosis and treatment

Data 252 patients assessed over a 6 year period
Successful Identification of DLBCL Patient Subgroups
Accurate Retrospective Prediction of Date of Biopsy

Histogram of date of biopsy - Groups 1 to 4

Histogram of data of biopsy - Groups 5 to 8

2001

Month
We Need a Data Normalization Method!
Heat Map of Flow Data

Cell Populations

Patients

Percentage of Cells

Cell Population 8
Cell Population 7
Cell Population 6
Cell Population 5
Cell Population 4
Cell Population 3
Cell Population 2
Cell Population 1
Significant Difference in Survival

Overall Survival Curves

Median Survival (good) = 3.4
Median Survival (poor) = 2.3
p-value = 3.2e-06
n = 95 out of 252 (38%)
Cut-off Threshold = 30%
Significance Not Threshold Specific
P-values vs. Threshold vs. Number of Patients
Population of Interest
 Clinically important and significant biomarker identified and verified by manual analysis, not correlated with treatment (CHOP/CHOP+R), or other prognostic indicators (e.g., IPI)

- Compare microarray data across DLBCL patient groups
  - Microarrays already done, this analysis points to new way to segregate groups
  - Identify pathways up or down regulated in each group

Apply framework to HIV, Leukemia and cancer stem cell datasets
Apply framework to build automated diagnosis tool
Conclusions

- Automated, unsupervised flow cytometry analysis can be:
  - As accurate as manual analysis
  - Potentially more informative when datasets get too large for manual analysis
Acknowledgements

- Brinkman lab
  Josef Spidlen, Ali Bashashati, Alireza Khodabakhshi, Melanie Courtot, Parisa Shooshtari, Adrian Cortes, Aaron Barsky, Kieran O’Neill, Habil Zare, Maik Hassel
- flow* Software
  IRCM Kenneth Lo, Greg Finak, Raphael Gottardo
  FCCC Florian Hahne, Nolwenn Le Meur, Nishant Gopalakrishnan, Byron Ellis, Deepayan Sarkar, Robert Gentleman
  BD Errol Strain Jim Walking, Michael Goldberg, Christopher Wolf, Perry Haaland, John Dunne
- Lymphoma
- BCCA Andrew Weng, Randy Gascoyne, Nathalie Johnson

NIH/NIBIB, CIHR, HeRRO, MSFHR, Genome [BC/Canada]

rbrinkman@bccrc.ca