FLOCK: A Density-Based Clustering Method for Automated Identification and Comparison of Cell Populations in High-Dimensional Flow Cytometry Data

Max Yu Qian, Ph.D.
Division of Biomedical Informatics and Department of Pathology
University of Texas Southwestern Medical Center, Dallas, TX
September 21, 2010
Why Computation Is Necessary

- Segregating overlapping cell populations
Solution: Clustering

• Assumption: Cells of the same *population* express ALL biological markers similarly
Related Work in Clustering

• Density-based (such as DBSCAN)
• Partitioning approaches (such as K-means)
• Hierarchical approaches (such as HAC)
• Grid-based approaches (such as STING)

J. Han, M. Kamber, A. K. H. Tung, “Spatial Clustering Methods in Data Mining: A Survey”

There is another category called Model-based Clustering, such as the EM method.
Previous Methods not Directly Applicable

FCM data requires the clustering method to be:
1) Efficient
2) Able to handle high-dimensionality
3) Easy setting parameters
Four populations on 2D display
Let $K=4$;
Select random seeds
Space partitioning based on centroids
Recalculate centroids
Repartition based on new centroids
Repeat the procedure many times
Final centroids
Let $K=3$
Space partitioning based on centroids
Recalculate centroids
Repartition based on new centroids
Repeat the procedure

... ...
Final Centroids
Final clustering results
Seeds trapped in local optimum even if K is correct
Non-spherical populations
K-means Applied to High-Dimensional Data

Three different ways used to generating random seeds

Number of Iterations = 1000, K=2
“For high dimensional data clustering, standard algorithms such as EM and K-means are often trapped in local minimum”


When number of dimension increases, there are more and more local optimum traps. This is also called *Curse of Dimensionality*. 
Therefore

Dimensions need to be reduced

However, the relationship between dimension selection and clustering is *chicken-egg*:

- to cluster high-dimensional data, dimensionality must be reduced (due to curse of dimensionality)
- it is more effective to select dimensions within individual data clusters than for whole dataset
flock
Flow cytometry clustering
without K
The Procedure of

1) Generate initial clusters (yes, chicken first!)
   - Parameter selection
2) Normalize dimensions within clusters
3) Select dimensions for initial clusters
4) Partition and merge the initial clusters in their selected subspaces
5) Output the final clusters

*Details of each step in following slides*
Generation of Initial Clusters
2D example
Divide with hyper-grids
Find dense hyper-regions
Merge neighboring dense hyper-regions
Clustering based on region centers
Bin selection methods

Goal is to minimize the Mean Squared Error

\[ L(h(x), f(x)) = \int (h(x) - f(x))^2. \]

- **Scott’s method**
  \[ v_{\text{scott}} = 3.49sN^{-1/3}. \]

- **Stone’s method**
  \[ K(v, M) = \frac{1}{v} \left( \frac{2}{N - 1} - \frac{N + 1}{N - 1} \sum_{m=1}^{M} \pi_i^2 \right) \]

- **Knuth’s method**, to maximize
  \[ N \log M + \log \Gamma \left( \frac{M}{2} \right) - M \log \Gamma \left( \frac{1}{2} \right) - \log \Gamma \left( N + \frac{M}{2} \right) + \sum_{k=1}^{M} \log \Gamma \left( n_k + \frac{1}{2} \right) + K \]
Density threshold selection

• Minimum description length

\[ \mu_s(i) = \left( \frac{\sum_{1 \leq j \leq i} x_j}{i} \right) \]

\[ \mu_d(i) = \left( \frac{\sum_{i+1 \leq j \leq \sigma} x_j}{(\sigma - i)} \right) \]

\[ L(i) = \log_2(\mu_s(i)) + \sum_{1 \leq j \leq i} \log_2(|x_j - \mu_s(i)|) + \log_2(\mu_d(i)) + \sum_{i+1 \leq j \leq \sigma} \log_2(|x_j - \mu_d(i)|) \]
Simulation Study

Birch dataset (Zhang et al, SIGMOD 1996)
Two assumptions with the above model

1) The center area is denser than the surrounding area in a population
2) There is only one group of adjacent hyper-regions in one population

When number of dimensions increases:
1) Assumption 1 may not hold for a sparse population; further partitioning to identify the sparse population may be necessary
2) There could be multiple adjacent hyper-regions within one population; they need to be merged.

Merging and partitioning will be done in the reduced-dimensional space
Density Variability in High-Dimensional Data Space

Fix the number of bins and density threshold, and use a Gaussian simulator to simulate 2-d, ..., 10-d data with 2 Gaussian clusters.

- **X-axis:** Number of dimensions
- **Y-axis:** Number of groups of adjacent hyper-regions

![Graph 1](chart1.png)

![Graph 2](chart2.png)

- **X-axis:** Number of dimensions
- **Y-axis:** Number of bins selected by Stone’s Method
Dimension Selection and Cluster Merging

1) 0-1 column-wise normalize each cluster

2) Select 3 dimensions for each cluster based on standard deviations (if number of dimensions < 3, all dimensions are used)

3) Partition a cluster into two, *if necessary* (this step can be optionally repeated)

4) 0-1 column-wise normalize each pair of partitions

5) Select 3 dimensions for each pair of partitions

6) Starting from the pair that are closest in the 3-dimensional space, merge a pair of partitions, *if necessary*

7) Repeat Steps 4) to 6) until there is no pair to merge
Merging/Partitioning Criteria

The most common approach is nearest/mutual neighbor graph, but it is very slow ($O(N^2)$).

Two partitions should not be merged

Two partitions should be merged
Results
FlowCAP Challenges

• Challenge 1 (fully automated)
• Challenge 2 (tuned parameters allowed)
• Challenge 3 (number of clusters known)
• Challenge 4 (manual gating results of a couple of files known)

Evaluation criteria: manual gating

Data: diffuse large B-cell lymphoma, graft versus host disease, normal donors, symptomatic west nile virus, and hematopoietic stem cell transplant
## FlowCAP Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Samples</th>
<th>#Events</th>
<th>Analyte</th>
<th>Detector</th>
<th>Reporter</th>
<th>Provided By</th>
</tr>
</thead>
<tbody>
<tr>
<td>GvHD</td>
<td>12</td>
<td>14,000</td>
<td>CD4, CD8b, CD3, CD8</td>
<td>Anti-CD4, Anti-CD8b, Anti-CD3, Anti-CD8</td>
<td>FITC, PE, PerCP, APC</td>
<td>BCCRC &amp; TreeStar</td>
</tr>
<tr>
<td>DLBCL</td>
<td>30</td>
<td>5,000</td>
<td>CD3, CD5, CD19</td>
<td>Anti-CD3, Anti-CD5, Anti-CD19</td>
<td>CY5, FITC, PE</td>
<td>BCCRC</td>
</tr>
<tr>
<td>ND</td>
<td>30</td>
<td>17,000</td>
<td>CD56, CD8, CD45, CD3/CD14</td>
<td>Proprietary, Anti-CD56, Proprietary, Anti-CD8, Proprietary, Anti-CD45, Anti-CD3/CD14</td>
<td>FITC, PerCPCy5, PacificBlue, PacificOrange, Qdot605, APC, Alexa700, PE, PECy5, PECy7</td>
<td>Amgen</td>
</tr>
<tr>
<td>WNV</td>
<td>13</td>
<td>100,000</td>
<td>IFN-γ, CD3, CD4, IL17, CD8</td>
<td>Anti-IFN-γ, Anti-CD3, Anti-CD4, Anti-IL17, Anti-CD8, NA</td>
<td>PEA, PECy5, PECy7, APC, AlexaFluor700, CFSE</td>
<td>McMaster</td>
</tr>
<tr>
<td>HSC</td>
<td>30</td>
<td>10,000</td>
<td>CD45.1, Ly65/Mac1, Dead Cells, CD45.2</td>
<td>Anti-CD45.1, Anti-Ly65/Mac1, NA</td>
<td>FITC, PE, PI, APC</td>
<td>BCCRC</td>
</tr>
<tr>
<td>Method</td>
<td>Rank</td>
<td>Score</td>
<td>Total Runtime</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>------</td>
<td>-------</td>
<td>-------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FlowVB</td>
<td>1</td>
<td>30.5</td>
<td>03:02:23:09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDP</td>
<td>2</td>
<td>17</td>
<td>00:01:48:06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEK</td>
<td>3</td>
<td>21</td>
<td>00:15:25:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLOCK</td>
<td>4</td>
<td>42.3</td>
<td>00:00:37:38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flowMeans</td>
<td>5</td>
<td>45.3</td>
<td>00:04:23:27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flowClust/Merge</td>
<td>6</td>
<td>23</td>
<td>10:13:00:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAME</td>
<td>7</td>
<td>41.8</td>
<td>00:05:31:12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MM&amp;PCA</td>
<td>8</td>
<td>26.5</td>
<td>00:00:04:35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MM</td>
<td>9</td>
<td>25.5</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SamSPECTRAL</td>
<td>10</td>
<td>41.5</td>
<td>00:07:21:44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SWIFT</td>
<td>11</td>
<td>15.5</td>
<td>05:23:24:30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Click on any thumbnail image to view and adjust populations of taskID=1 File=001

<table>
<thead>
<tr>
<th></th>
<th>FS</th>
<th>SS</th>
<th>FL1LOG</th>
<th>FL2LOG</th>
<th>FL4LOG</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
</tr>
<tr>
<td>SS</td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
</tr>
<tr>
<td>FL1LOG</td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
</tr>
<tr>
<td>FL2LOG</td>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
</tr>
<tr>
<td>FL4LOG</td>
<td><img src="image21" alt="Image" /></td>
<td><img src="image22" alt="Image" /></td>
<td><img src="image23" alt="Image" /></td>
<td><img src="image24" alt="Image" /></td>
<td><img src="image25" alt="Image" /></td>
</tr>
</tbody>
</table>
X: FL2; Y: FL4

DLBCL_001

DLBCL_006
<table>
<thead>
<tr>
<th></th>
<th>FSCH</th>
<th>SSCH</th>
<th>FL1H</th>
<th>FL2H</th>
<th>FL3H</th>
<th>FL4H</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSCH</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td>SSCH</td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td>FL1H</td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
</tr>
<tr>
<td>FL2H</td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
</tr>
<tr>
<td>FL3H</td>
<td><img src="image25.png" alt="Image" /></td>
<td><img src="image26.png" alt="Image" /></td>
<td><img src="image27.png" alt="Image" /></td>
<td><img src="image28.png" alt="Image" /></td>
<td><img src="image29.png" alt="Image" /></td>
<td><img src="image30.png" alt="Image" /></td>
</tr>
<tr>
<td>FL4H</td>
<td><img src="image31.png" alt="Image" /></td>
<td><img src="image32.png" alt="Image" /></td>
<td><img src="image33.png" alt="Image" /></td>
<td><img src="image34.png" alt="Image" /></td>
<td><img src="image35.png" alt="Image" /></td>
<td><img src="image36.png" alt="Image" /></td>
</tr>
</tbody>
</table>
High-dimensional Data
<table>
<thead>
<tr>
<th>ND_001</th>
<th>PDC-A</th>
<th>SSC-A</th>
<th>FSC-A</th>
<th>PerCP-Cy5.5-A</th>
<th>PacificBlue-A</th>
<th>PacificOrange-A</th>
<th>GoxN5-A</th>
<th>APC-A</th>
<th>Alexa700-A</th>
<th>PE-A</th>
<th>PE-Cy5-A</th>
<th>PE-Cy7-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD3/CD14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Challenge 2 (tuned)

<table>
<thead>
<tr>
<th>Method</th>
<th>WNV</th>
<th>ND</th>
<th>Rank Score</th>
<th>Total Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF-curvHDR</td>
<td>0.81 (0.77, 0.84)</td>
<td>0.83 (0.82, 0.84)</td>
<td>15</td>
<td>07:23:04:00</td>
</tr>
<tr>
<td>CDP</td>
<td>0.75 (0.71, 0.78)</td>
<td>0.86 (0.84, 0.88)</td>
<td>11</td>
<td>00:00:33:30</td>
</tr>
<tr>
<td>FLOCK</td>
<td><strong>0.84 (0.82, 0.86)</strong></td>
<td><strong>0.89 (0.87, 0.91)</strong></td>
<td>25.5</td>
<td>00:00:28:31</td>
</tr>
<tr>
<td>flowClust/Merge</td>
<td>0.77 (0.75, 0.79)</td>
<td>0.88 (0.81, 0.92)</td>
<td>21.5</td>
<td>10:13:00:00</td>
</tr>
<tr>
<td>FLAME</td>
<td>0.84 (0.82, 0.85)</td>
<td>0.87 (0.86, 0.87)</td>
<td>22.5</td>
<td>-</td>
</tr>
<tr>
<td>SamSPECTRAL</td>
<td>0.85 (0.83, 0.88)</td>
<td>0.91 (0.91, 0.92)</td>
<td>25.5</td>
<td>00:13:00:00</td>
</tr>
<tr>
<td>SamSPECTRAL-FixK</td>
<td>0.76 (0.71, 0.81)</td>
<td>0.92 (0.91, 0.93)</td>
<td>19</td>
<td>00:08:26:10</td>
</tr>
</tbody>
</table>

Compared with Challenge 1
FLOCK in ImmPort (www.immport.org)
## Automated Identification of Cell Populations

<table>
<thead>
<tr>
<th></th>
<th>FSC</th>
<th>SSC</th>
<th>FL1</th>
<th>FL2</th>
<th>FL3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSC</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td>SSC</td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td>FL1</td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
<tr>
<td>FL2</td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
<tr>
<td>FL3</td>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
<td><img src="image25.png" alt="Image" /></td>
</tr>
</tbody>
</table>

FCM data from Montgomery Lab, Yale Univ.
Cross-Sample Comparison with FLOCK

Proportion change of PlasmaBlasts at different days with Tetanus study

FCM data from Sanz Lab, Univ. of Rochester
Download FLOCK Results to Your Own Software

Casale FCM data from Immune Tolerance Network
Visualization Software: Tableau
Discussion

• Computational analysis most needed for high-dimensional dataset
• Preprocessing is also important
• FlowCAP2 can include cross-sample comparison, since the alignment and mapping is also challenging
• From cluster to population
Conclusions

FLOw Clustering without K - FLOCK
- Identifies cell populations within multi-dimensional space
- Automatically determines the number of unique populations present using a rapid binning approach
- Can handle non-spherical hyper-shapes
- Maps populations across independent samples
- Calculates useful summary statistics
- Reduces subjective factors in gating
- Implemented in ImmPort and freely available
Acknowledgment

UT Southwestern
Richard Scheuermann
Megan Kong
Paula Guidry
David Dougall
Eva Sadat
Jamie Lee
Jennifer Cai
Jie Huang
Nishanth Marthandan
Diane Xiang
Young Kim
Adam Seegmiller
Nitin Karandikar

Northrop Grumman
John Campbell
Yue Liu
Liz Thompson
Patrick Dunn
Jeff Wiser
Mike Atassi

Rochester
Iñaki Sanz
Chungwen Wei
Eun Hyung Lee
Tim Mosmann
Jessica Halliley
Chris Tipton

Immune Tolerance Network
Dave Parrish
Keith Boyce
Tom Casale
Jason Liu

FlowCAP Organization Committee

Supported by NIH N01 AI40076 (BISC)