Extracting a cellular hierarchy from high-dimensional single-cell data

Peng Qiu

Department of Bioinformatics and Computational Biology
University of Texas MD Anderson Cancer Center
Flow / mass cytometry data
Biology questions

• How many cell types are there?

• How are different cell types related to each other?

• Does the cellular composition of a sample correlate with its overall phenotype?
Introduction - gating

- Example data
  - 8-parameter flow cytometry
  - Mouse bone marrow
  - Parameters: c-kit, Sca-1, CD11b, B220, TCR-β, CD4, CD8

- Traditional analysis: Gating
Basic idea

- Consider the data as a point cloud
- Extract the shape of the cloud

Method:
Spanning-tree Progression Analysis of Density-normalized Events (SPADE)
SPADE

(a) Cytometry data
Density-dependent downsampling
Downsampled data
Agglomerative clustering
Clustering result
Minimum spanning tree construction
SPADE tree
Upsampling
Colored tree showing cellular heterogeneity

(b) Cytometry data
Marker 1
Marker 2
Density-dependent downsampling

(d) Agglomerative clustering

(e) Minimum spanning tree construction

(f) Marker 1 intensity
Marker 2 intensity
Graph representation of the underlying cellular heterogeneity

Qiu et al, *Nature Biotechnology*, in press
SPADE applied to mouse bone marrow data.
SPADE vs. gating
SPADE applied to human bone marrow data

Bendall et al, Science, 2011
SPADE applied to CyTOF data of human BM

Qiu et al, *Nature Biotechnology*, in press
Challenge 2: Normal vs AML

- 359 subjects
  - 316 normal subjects
  - 43 AML samples

- 8 Tubes per subject

- Channels per tube: FSC+SSC+5 colors
Challenge 2: Normal vs AML

Since the overlap among the 8 different staining panels/tubes is minimal, we consider them separately.

Therefore, we have 359 fcs files to compare.
Challenge 2: Normal vs AML

Since the overlap among the 8 different staining panels/tubes is minimal, we consider them separately.

Therefore, we have 359 fcs files to compare.

Apply SPADE to the union of the two clouds
SPADE tree for Tube 2
SPADE tree for Tube 2
SPADE tree for Tube 2
SPADE tree for Tube 2
RELIEF classifier & Earth Mover’s Distance

Earth Mover’s Distance:

a metric to compare two probability distributions over a structured domain.

RELIEF classifier

for each testing sample, find its nearest normal (N_N) and its nearest AML (N_AML)

compute the following score: dist-to-N_N – dist-to-N_AML
RELIEF classifier & Earth Mover’s Distance

Training samples

Testing samples
Challenge 3A

Use 48*2 samples to derive a SPADE tree

Compute cell freq distribution for each sample

For each sample, compute its distribution – the distribution of its paired sample.

PCA
Summary

• Using SPADE, we can:
  – Identify cell types
  – Compare multiple samples
Acknowledgement

• Sylvia Plevritis

• Garry Nolan
  – Erin Simonds, Sean Bendall,
  – Kenny Gibbs
  – Karen Sachs, Michael Linderman, Rob Bruggner
  – Matt Clutter, Tiffany Chen