

# Misty Mountain – A Parallel Clustering Method. Application to Fast Unsupervised Flow Cytometry Gating

István P. Sugár and Stuart C. Sealfon

Department of Neurology and Center for Translational Systems Biology, Mount Sinai School of Medicine, New York

## Misty Mountain clustering/automated gating:

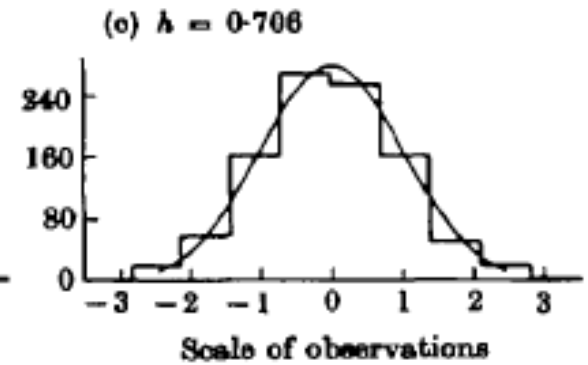
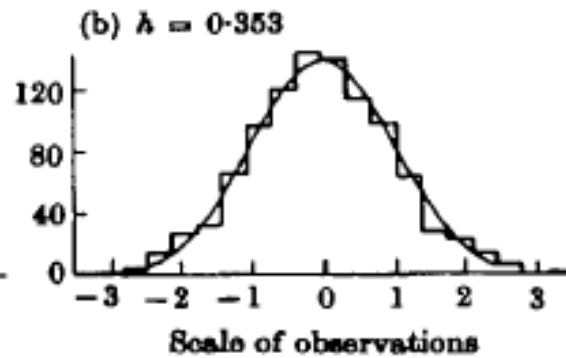
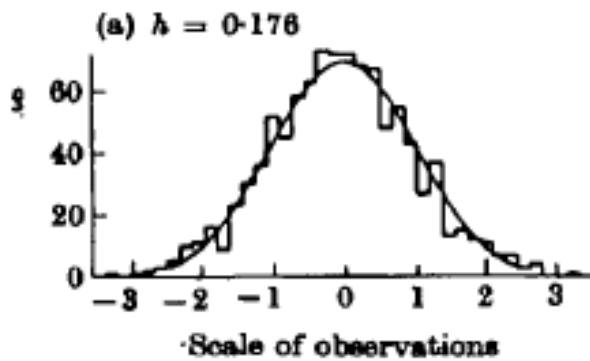
- unsupervised
- unbiased for cluster shape
- fast (run time increases linearly with the number of data points)
- high clustering accuracy in multiple “gold standard tests”

# Steps of Misty Mountain clustering

The multi-dimensional data is first processed to generate a histogram containing an optimal number of bins by using Knuth's data-based optimization criterion.

Then cross sections of the histogram are created. The algorithm finds the largest cross section of each statistically significant histogram peak.

The data points belonging to these largest cross sections define the clusters of the data set



## Knuth's data-based binning for histogram

The  $N$  that maximizes the following function is the optimal bin number along each coordinate axis:

$$\log p(N | \underline{d}) =$$

$$n \log N^D + \log \Gamma(0.5N^D) - N^D \log \Gamma(0.5) - \log \Gamma(n + 0.5N^D) + \sum_{k=1}^{N^D} \log \Gamma(n_k + 0.5) + \text{const.}$$

$n$  = number of data points

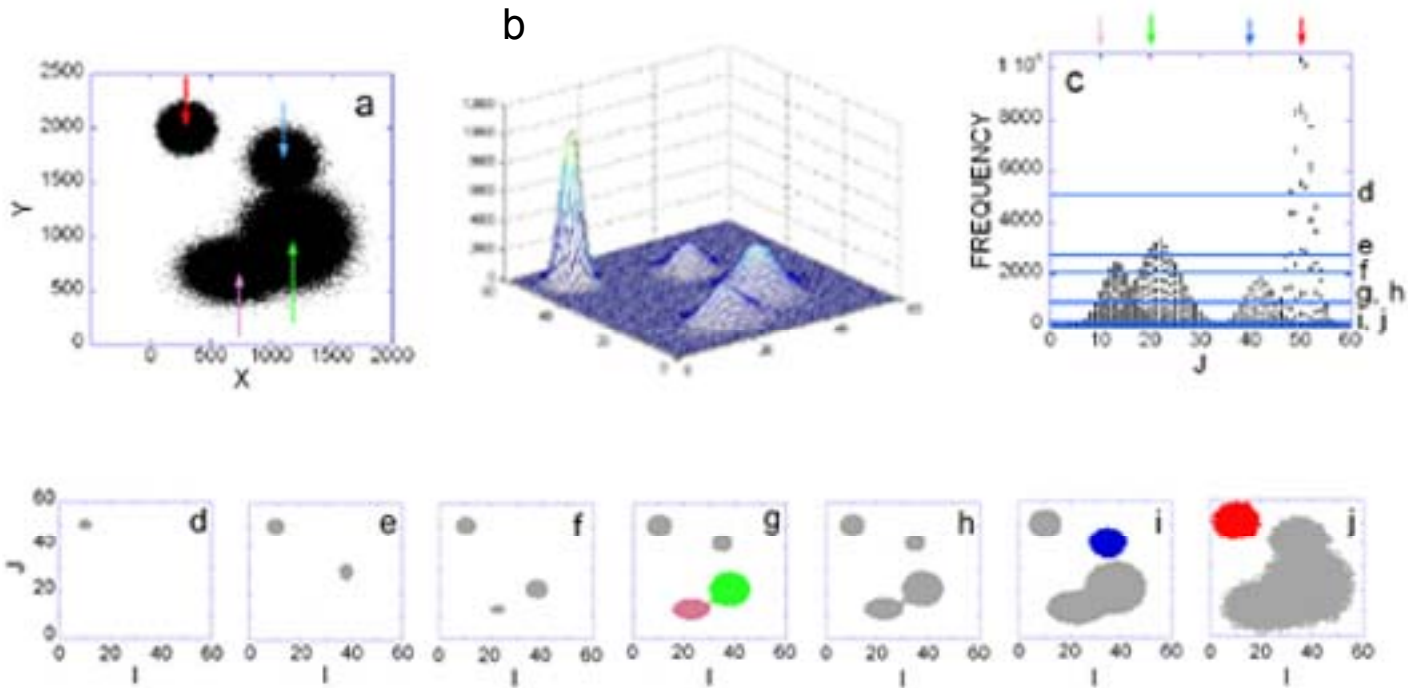
$n_k$  = number of data points in the  $k$ -th bin

$D$  = dimension of the data space

$p(N|\underline{d})$  = probability for the number of bins of similar shape at given data  $\underline{d}$ .

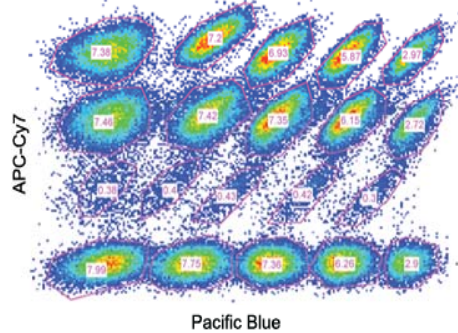
$\Gamma(x)$  = gamma function

# Misty Mountain clustering

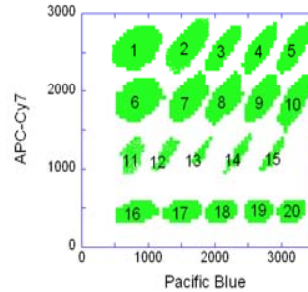


# Comparison of different methods by clustering the same 2D barcoding data set

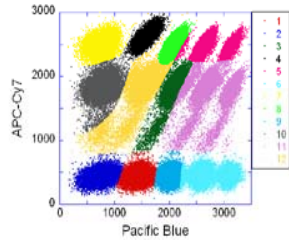
**A** Expert Manual Gating: 6 minutes



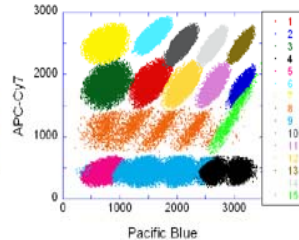
**B** Misty Mountain: 10 seconds



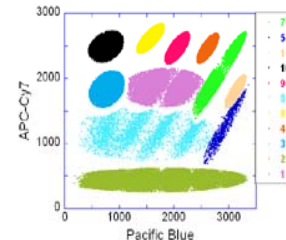
**C** FLAME: 14 h



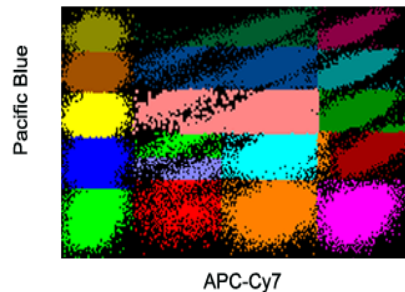
**D** flowClust: 14 h



**E** flowMerge: 35 h



**F** flowJo: ~10 seconds



Gated	Show	Events	Name	<Pacific_Blue-A>	<APC-Cy7-A>
*	*	180924	(Parent)		
*	*	15032	Cluster #01-111	++	++
*	*	14680	Cluster #02-111	--	--
*	*	14412	Cluster #03-112	+	--
*	*	14399	Cluster #04-117	--	++
*	*	14012	Cluster #05-110	+	+++
*	*	13717	Cluster #06-114	--	+++
*	*	13606	Cluster #07-113	++	--
*	*	13088	Cluster #08-119	+	+++
*	*	13023	Cluster #09-120	++	+++
*	*	12866	Cluster #10-117	+++	++
*	*	11633	Cluster #11-114	+++	--
*	*	10908	Cluster #12-121	+++	++
*	*	5510	Cluster #13-122	+++	++
*	*	5503	Cluster #14-115	+++	++
*	*	5485	Cluster #15-113	+++	++
*	*	1498	Cluster #16-216	--	+
*	*	721	Cluster #17-119	++	++
*	*	662	Cluster #18-118	+	++
*	*	170	Cluster #19-118	+	++

## Comparison of clustering accuracy

Clustering Method	Clustering accuracy	
	sensitivity (%)	specificity (%)
Misty Mountain	100	100
FLAME	20 <sup>a</sup> 60 <sup>b</sup>	33 <sup>a</sup> 50 <sup>b</sup>
flowClust	45 <sup>a*</sup> 60 <sup>b*</sup>	60 <sup>a*</sup> 55 <sup>b*</sup>
flowMerge	25	45
flowJo	45	47

$$\text{sensitivity} = \frac{\# \text{ of correctly assigned clusters}}{\# \text{ of clusters in gold standard}}$$

$$\text{specificity} = \frac{\# \text{ of correctly assigned clusters}}{\text{total } \# \text{ of assigned clusters}}$$

Gold standards were independent expert manual clustering for 2D barcoding data.

# Serial vs. Parallel Clustering

Model based clustering requires serial clustering for all cluster numbers within a user defined interval. Then the optimal cluster number is selected by an information criterion.

Misty Mountain is a parallel clustering method that finds every cluster after analyzing only once the cross sections of the histogram

## Performance of Misty Mountain clustering in flowCAP challenges #1

	<b>Stem (D=4)</b>	<b>GvHD (D=4)</b>	<b>DLBCL (D=3)</b>
Number of data sets	30	12	30
Average CPU per data set (sec)	0.284	0.623	0.184
Total CPU for all data sets (sec)	8.52	7.48	5.52
Cluster # deviates by 0 from manual clustering	67%	42%	40%
Cluster # deviates by 1 from manual clustering	27%	58%	43%
Cluster # deviates by 2 from manual clustering	6%	0%	17%

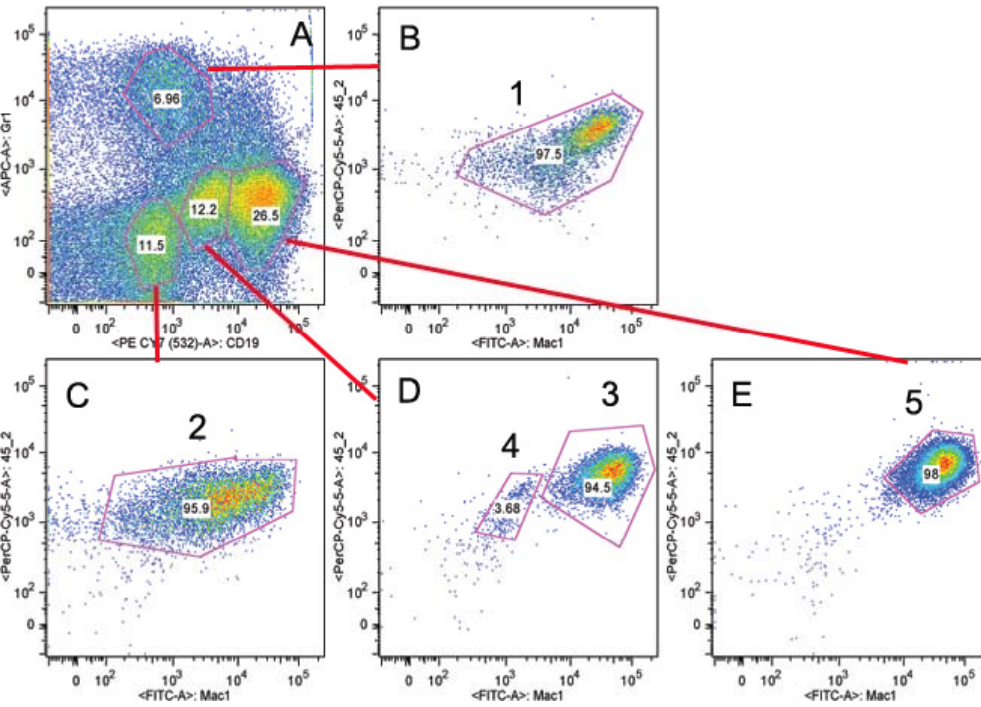
# Acknowledgements

We thank Profs. D. Stäuffer and B. Roysam for sending the source code of a Hoshen-Kopelman type cluster counting algorithm and spectral clustering, respectively. We also thank Prof. F. Hayot for the critical evaluation of the manuscript. We acknowledge Drs. B. Hartman and J. Seto for providing the FCM data and Dr. German Nudelman for making the program available on the web. Dr. Yongchao Ge for analyzing FCM data with flowClust and flowMerge. We are grateful for Prof. Ryan Brinkman for providing access to the GvHD flow cytometry data sets and to Prof. Hans Snoeck for providing the OP9 dataset. This work from the Program for Research in Immune Modeling and Experimentation (PRIME) was supported by contract NIH/NIAID HHSN266200500021C.

# Publication

Sugar, IP; Sealfon, SC (2010) Misty Mountain clustering: application to fast unsupervised flow cytometry gating, BMC Bioinformatics, in press

# Comparison of different methods by clustering the same 4D OP9 data set



## Manual gating of 4D OP9 data set

A) 4 clusters were gated in the APC/PE CY7 plane, B-E) elements of each of the 4 clusters are projected into the PerPC-CY5/FITC plane. In this plane only one of the four clusters splitted into two clusters, while the others remained single clusters. Thus the manual gating identified 5 clusters total.

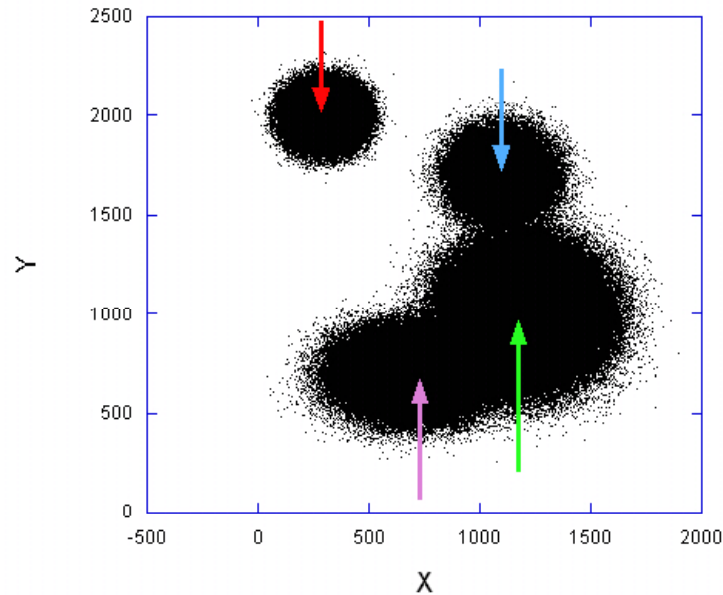
## Comparison of clustering accuracy

Clustering Method	Clustering accuracy		Cluster number	CPU (sec)
	sens (%)	spec (%)		
Misty Mountain	100	100	5	3.6
flowClust	60 60	75 38	4 8	3660
flowMerge	25	45	7	8400

$$\text{sensitivity} = \frac{\# \text{ of correctly assigned clusters}}{\# \text{ of clusters in gold standard}}$$

$$\text{specificity} = \frac{\# \text{ of correctly assigned clusters}}{\text{total } \# \text{ of assigned clusters}}$$

Gold standards were independent expert manual clustering for 4D OP9 data.



## Goal of the cluster analysis

Select from the experimental data separated clusters of data points where each cluster characterizes the respective group of data points