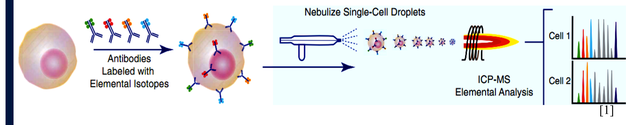
Comparative Analysis of Linear and Nonlinear Dimension Reduction Techniques on Mass Cytometry Data

Nathan Jekel[1*], Emily Vidal[2*], Anna Konstorum[3], Reinhard Laubenbacher[3]
[1]Penn State Harrisburg, [2]Angelo State University, [3]Center for Quantitative Medicine, UCONN Health
*These authors contributed equally to this work

Introduction

❖ Mass cytometry is a newly developed technology for quantification and classification of immune cells that can analyze up to 100 markers per cell. High dimensional data resulting from these experiments require innovative methods for analysis and visualization.



❖ Mathematical dimension reduction techniques map data from the input space to a lower dimensional subspace. Each technique aims to preserve a specific characteristic of the data during this process.

HIGH DIMENSIONAL SPACE

	Biomarker 1	Biomarker 2	...	Biomarker m
Cell 1	#	#	...	#
Cell 2	#	#	...	#
⋮	⋮	⋮	⋱	#
Cell n	#	#	...	#

DIMENSION REDUCTION

LOW DIMENSIONAL SPACE

	Principal Variable 1	Principal Variable 2
Cell 1	#	#
Cell 2	#	#
⋮	⋮	⋮
Cell n	#	#

Objectives

1. Implement 4 dimension reduction techniques on a benchmark manually gated (divided into cell subtypes) mass cytometry data set
2. Compare techniques using 3 metrics: Computation Time, Neighborhood Proportion Error, and Residual Variance
3. Apply best techniques to non-gated mass cytometry data

Methods

METHOD	TYPE	PRESERVED CHARACTERISTIC	COST FUNCTION
Isomap [3]	Nonlinear	Geometric structure of data	$\| \tau(D_G) - \tau(D_Y) \|_{L^2}$
t-SNE [4,5]	Nonlinear	Gaussian based similarity measure	$KL(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$
Diffusion Maps (D-Maps) [6]	Nonlinear	Distance based on Gaussian random walk	$\sum_i \sum_j \left(D^{(t)}(x_i, x_j) - \| y_i - y_j \|\right)^2$
PCA [2]	Linear	Variability	$\underset{\|\mathbf{w}\|=1}{\mathbf{argmax}\ \mathbf{var(t)}}$

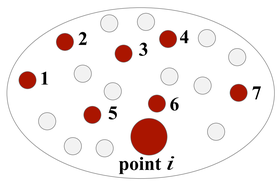
Comparison Metrics

1. Computation Time

2. Neighborhood Proportion Error (NPE):

- Find the k nearest neighbors for each data point and count the number of neighbors that are of the same cell subtype

$k = 20$
# like neighbors = 7
Percentage = 35%



- Convert the counts for each subtype into a probability distribution
- Calculate Total Variation distance between the high and low dimensional distributions

$$\delta(P, Q) = \sup_{A \in F} \mid P(A) - Q(A) \mid$$

- Sum the Total Variation for each subtype to get NPE

$$E = \sum_{i=1}^{n} \delta(P_i, Q_i)$$

3. Residual Variance:

Residual variance is defined as

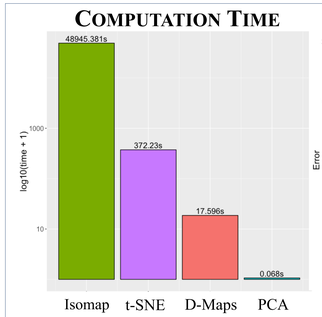
$$1 - r^2(D_M, D_Y)$$

where r is the Pearson correlation coefficient between a distance matrix in the original space (D_M) and a distance matrix in the dimension reduced space (D_Y)[7].

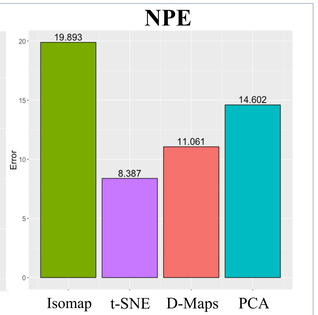
Results (Benchmark Data)

❖ All dimension reductions were performed using the statistical computing software R [2]. The benchmark data set consists of 167,044 cells from healthy human bone marrow. Its purpose is to measure healthy human hematopoiesis [1].

COMPUTATION TIME

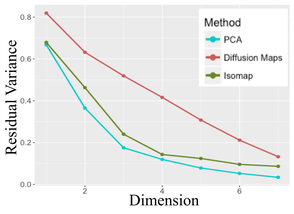


NPE



✧ *The NPE and Computation Time calculations were measured on reductions from 13 to 2 dimensions with random subsets of 10,000 cells.*

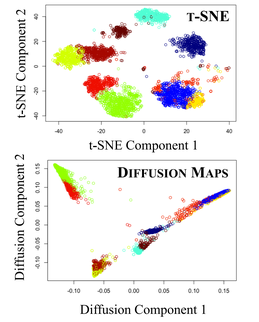
RESIDUAL VARIANCE



✧ *Residual Variance was calculated using reductions from 13 dimensions performed on a 5,000 cell subset*

✧ *Residual Variance is not applicable to t-SNE.*

2D VISUALIZATIONS



✧ *Colors denote distinct manually gated cell subtypes.*

Conclusion I (Benchmark Data)

❖ Diffusion Maps and t-SNE's performances in NPE suggest that they best preserve the local structure of the data.

❖ In 2D visualization, t-SNE showed well-defined phenotypic clustering, and Diffusion Maps showed structure indicative of cell differentiation.

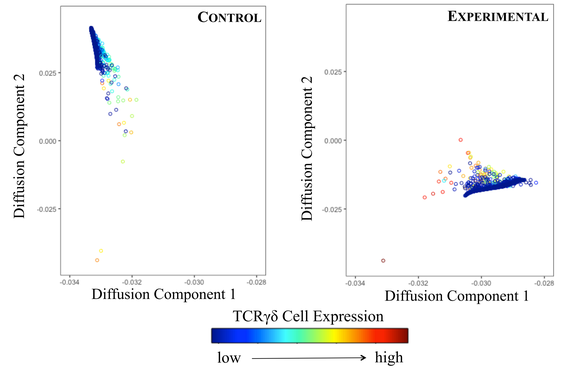
❖ PCA and Isomap displayed low residual variance, indicating that they preserve the most information globally.

❖ Overall, Diffusion Maps and t-SNE provide the best insights into cell phenotype and differentiation.

Results (ART Data)

❖ The acute response to toxin (ART) dataset consists of 66,662 control and toxin-stimulated immune cells. Its purpose is to examine the underlying molecular basis of TCRγδ cell activation. Upon infection, a special type of T cell, TCRγδ, is activated immediately, subsequently secreting cytokines that activate other immune cells.

DIFFUSION MAPS VISUALIZATIONS



Conclusion II (ART Data)

❖ The cells in the control and experimental data occupy different regions of the low dimensional space, indicating that dimension reduction preserves differences in experimental conditions.

❖ High TCRγδ cells are projected to the edge of the dimension reduced space, which indicates that they represent a phenotype that is distinct from the other cell types.

References

[1] Bendall SC et al. (2011). *Science* **332**, pp. 687-696
[2] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
[3] Bartenhagen C (2014). *RDRToolbox: A package for nonlinear dimension reduction with Isomap and LLE.* R package version 1.22.0
[4] Becher B et al. (2014). *Nature Immunology*, **15**, pp. 1181–1189.
[5] Wong MT et al. (2015). *Cell Reports*, **11**, pp. 1822–1833.
[6] Philipp Angerer et al. (2015): destiny – diffusion maps for large-scale single-cell data in R.
[7] Duan M et al. (2013). *J. Chem. Theory Comput.*, 9(5), pp 2490–2497

Acknowledgements

Funded by National Science Foundation award #1460967. Special thanks to Dr. Miranda Lynch of the UCONN Center for Quantitative Medicine for statistical consultation and Dr. Antoine Menoret of the UCONN Health Immunology Department for providing the ART data.