2D Visualization of Immune System Cellular Protein Data using Dimensionality Reduction

Andre Esteva
esteva@stanford.edu
Anand Sampat
asampat@stanford.edu
Amit Badlani
abadlani@stanford.edu

Introduction

Immune system cells found in bone marrow come in many types.

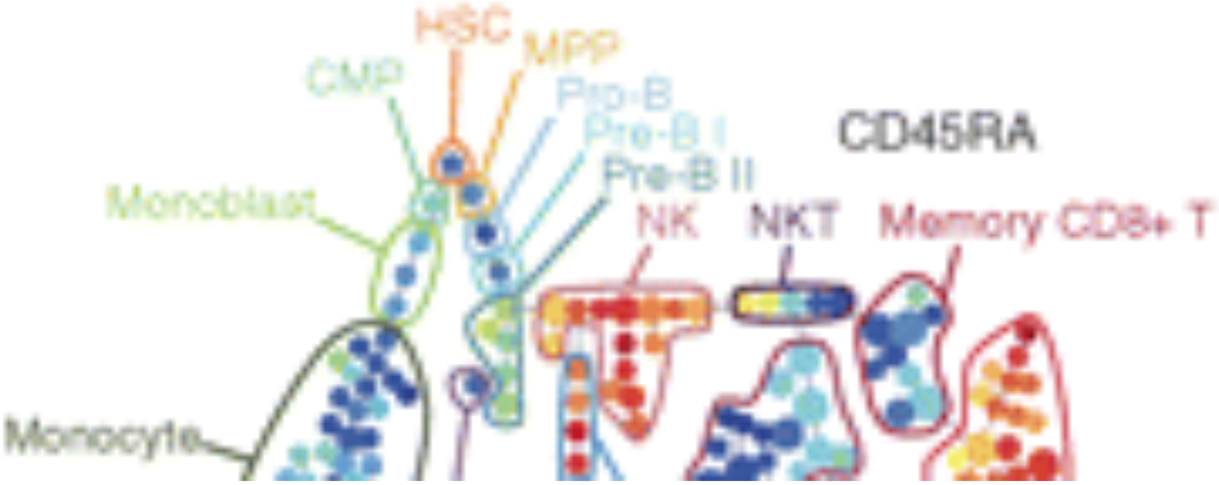
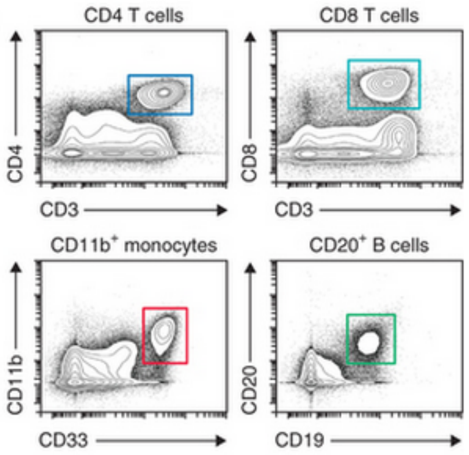


Figure 1: Various immune cell types

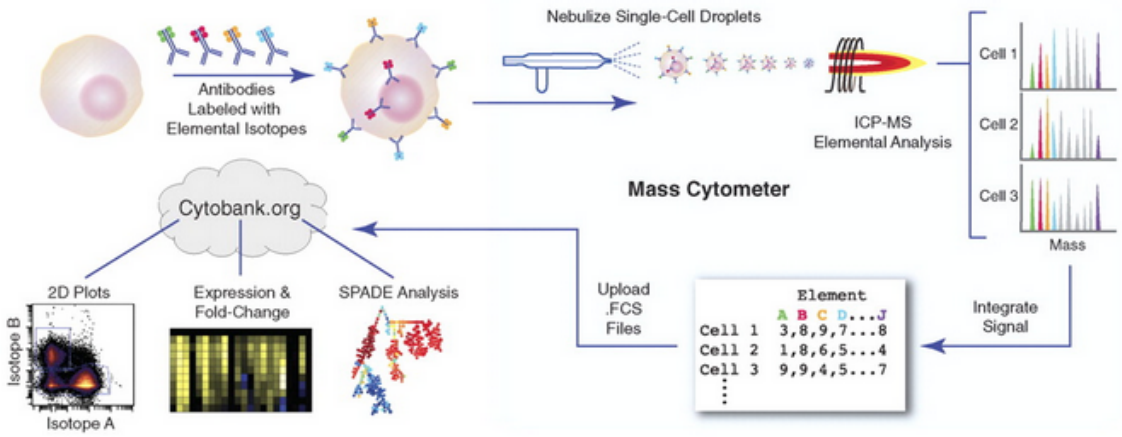
Each of them is identified by both intracellular proteins (IP) and surface proteins (SP).

Motivation

- There are 17 SP and 21 IP expressions that define each cell, which is too much to visualize properly.
- These cells are manually classified in a process called “gating”, shown in the figure to the right. .



Data Format



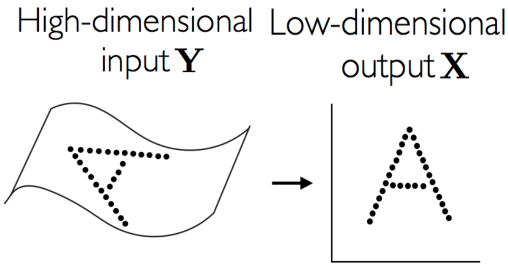
```
         Element
         A B C D...J
Cell 1   3,8,9,7...8
Cell 2   1,8,6,5...4
Cell 3   9,9,4,5...7
  ⋮
```

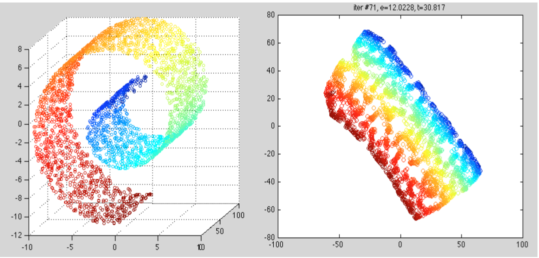
Source:
Bendell et al.

Algorithms

Dimensionality Reduction Methods

- Linear Methods
  - Principal Component Analysis
  - Multidimensional scaling
- Spectral Methods
  - Laplacian Eigenmaps
  - ISOMAP
  - Locally Linear Embedding
- Nonlinear Methods
  - Stochastic Neighbor Embedding (SNE)
  - symmetric-SNE (s-SNE)
  - t-distributed-SNE (t-SNE)
  - Elastic Embedding (EE)

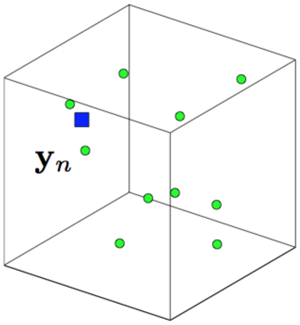

High-dimensional input $\mathbf{Y}$
Low-dimensional output $\mathbf{X}$

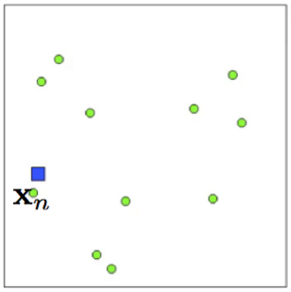


SNE Algorithm (Hinton & Roweis, 2003)

Define a conditional probability in both spaces that a point selects any other point as its neighbor

$$p_{m|n} = \frac{\exp(-\|(\mathbf{y}_n - \mathbf{y}_m)/\sigma^2\|)}{\sum_{k \neq n} \exp(-\|(\mathbf{y}_n - \mathbf{y}_k)/\sigma^2\|)} \qquad q_{m|n} = \frac{\exp(-\|\mathbf{x}_n - \mathbf{x}_m\|^2)}{\sum_{k \neq n} \exp(-\|\mathbf{x}_n - \mathbf{x}_k\|^2)}$$





Minimize the KL Divergence between the two distributions

$$E_{SNE}(\mathbf{X}) = \sum_{n=1}^{N} \mathrm{D}\left(P_n \| Q_n\right) = \sum_{n,m=1}^{N} p_{n|m} \log \frac{p_{n|m}}{q_{n|m}}$$

s-SNE (Cook et al, 2007)
symmetric matrices
normalizes pdfs over all
interactions

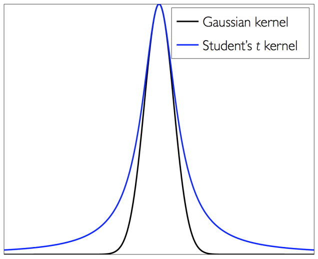
$$p_{nm} = \frac{\exp(-\|(\mathbf{y}_n - \mathbf{y}_m)/\sigma\|^2)}{\sum_{k,l=1}^{N} \exp(-\|(\mathbf{y}_k - \mathbf{y}_l)/\sigma\|^2)}$$

$$q_{nm} = \frac{\exp(-\|\mathbf{x}_n - \mathbf{x}_m\|^2)}{\sum_{k,l=1}^{N} \exp(-\|\mathbf{x}_l - \mathbf{x}_k\|^2)}$$

t-SNE (van der Maaten & Hinton, 2008)
low-d pdf uses Student's t-distribution

$$p_{nm} = \frac{\exp(-\|(\mathbf{y}_n - \mathbf{y}_m)/\sigma\|^2)}{\sum_{k,l=1}^{N} \exp(-\|(\mathbf{y}_k - \mathbf{y}_l)/\sigma\|^2)}$$

$$q_{nm} = \frac{(1 + \|\mathbf{x}_n - \mathbf{x}_m\|^2)^{-1}}{\sum_{k,l=1}^{N} (1 + \|\mathbf{x}_l - \mathbf{x}_k\|^2)^{-1}}$$



Objective Function Comparison for SNE varieties:

$$E_{SNE}(\mathbf{X}) = \sum_{n,m=1}^{N} p_{nm} \|\mathbf{x}_n - \mathbf{x}_m\|^2 + \sum_{n=1}^{N} \log \sum_{m \neq n}^{N} \exp(-\|\mathbf{x}_n - \mathbf{x}_m\|^2)$$

$$E_{s\text{-}SNE}(\mathbf{X}) = \sum_{n,m=1}^{N} p_{nm} \|\mathbf{x}_n - \mathbf{x}_m\|^2 + \log \sum_{n,m=1}^{N} \exp(-\|\mathbf{x}_n - \mathbf{x}_m\|^2)$$

$$E_{t\text{-}SNE}(\mathbf{X}) = \sum_{n,m=1}^{N} p_{nm} \log(1 + \|\mathbf{x}_n - \mathbf{x}_m\|^2) + \sum_{n,m=1}^{N} (1 + \|-\mathbf{x}_n - \mathbf{x}_m\|^2)^{-1}$$

Elastic Embedding

Define two neighborhood graphs

$$w^{+}_{nm} = \exp(-\frac{1}{2}\|(\mathbf{y}_n - \mathbf{y}_m)/\sigma\|^2) \qquad w^{-}_{nm} = \|\mathbf{y}_n - \mathbf{y}_m\|^2$$

Minimize with respect to **X**:

$$E_{EE}(\mathbf{X}, \lambda) = \sum_{n,m=1}^{N} w^{+}_{nm} \|\mathbf{x}_n - \mathbf{x}_m\|^2 + \lambda \sum_{n,m=1}^{N} w^{-}_{nm} \exp(\|-\mathbf{x}_n - \mathbf{x}_m\|^2)$$

Nonlinear Embedding Optimization

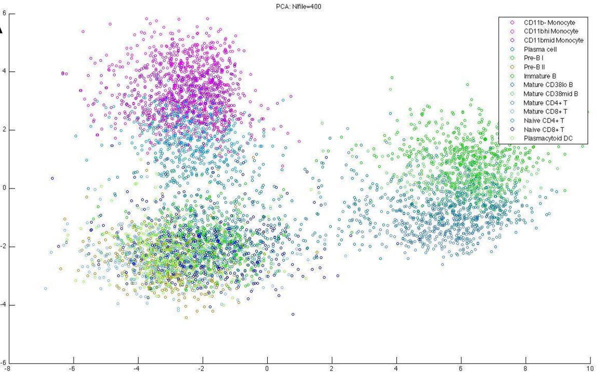
Minimize objective function by iterating over k:

1. Solve $B_k p_k = g_k$ for B_k = psd matrix, p_k = search direction, g_k = gradient
   Choose B_k as the Hessian of the green terms (aka *spectral directions*)
   Is faster than B_k=I (gradient descent) or B_k=Hessian(E) (Newton Met.)
2. Find step size n using line search for the next iteration: $x_{k+1} = x_k + n*p_k$

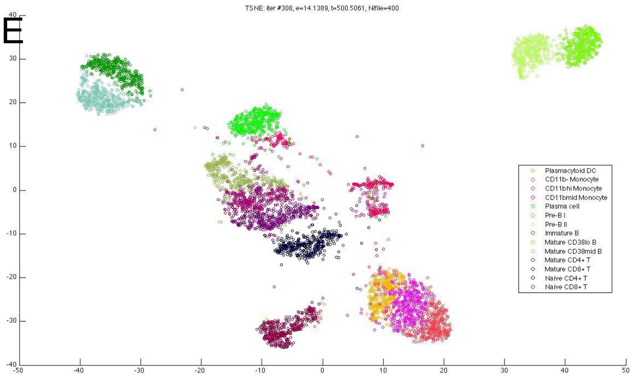
Results

Cell Types: Monocytes, pDC, B Cells, & T Cells 400 Samples/type, Surface Proteins

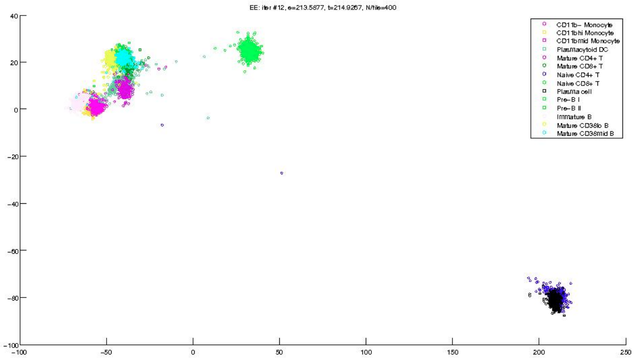
PCA



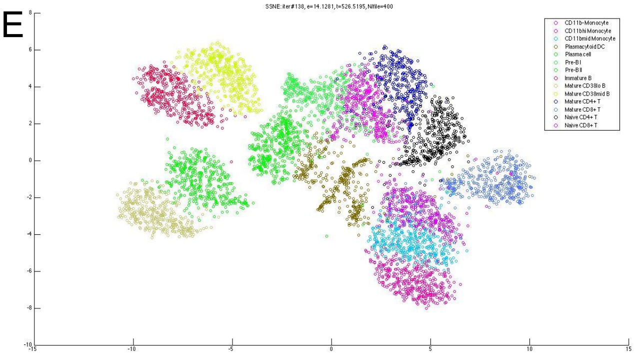
t-SNE



EE

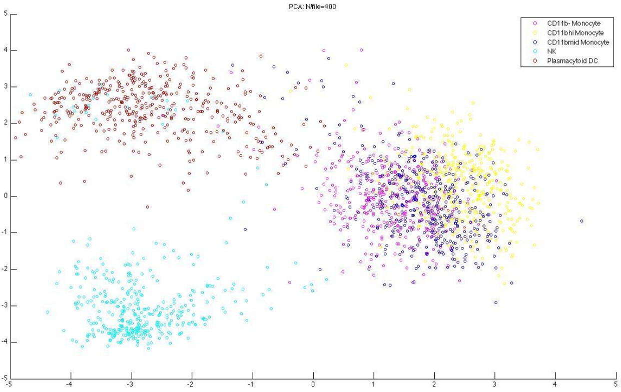


s-SNE

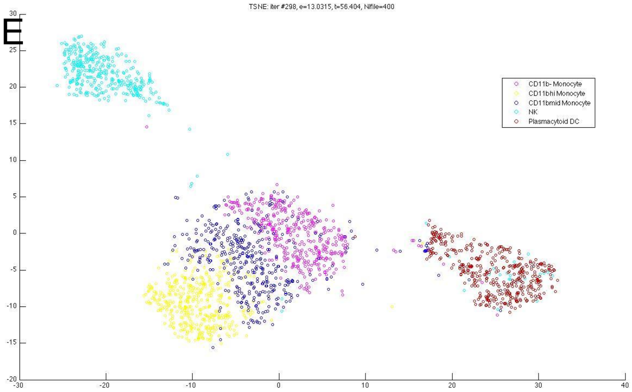


Few Cell Types: Monocytes, pDC, & NK Cells, 400 Samples/type, Surface Proteins

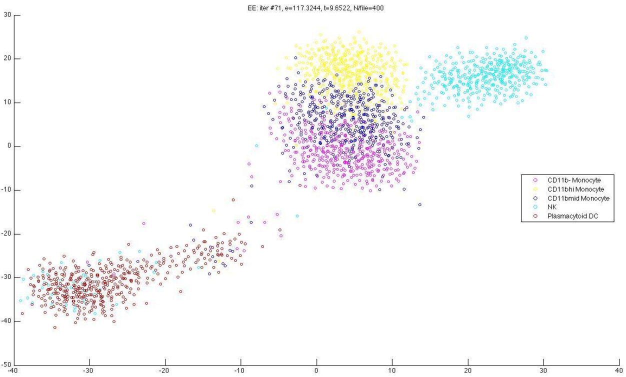
PCA



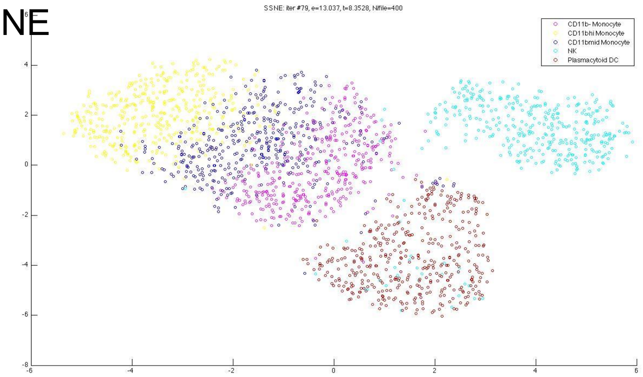
t-SNE



EE



s-SNE



Conclusion

- Linear methods such as PCA do not work well on this data, leading us to believe that the data is not linearly separable, as is common for many cell data sets

- t-SNE results in the most segregated clusters as per our metrics (not shown)

- s-SNE and EE are similarly well-segregated but had lower segregation metrics compared to t-SNE.

Future Work

- Classification of cell types using Crammer SVM for non-binary labelings
- Application of these techniques with the inclusion of gated cancerous cells
- Use of these algorithms on unlabeled data for prediction classification

References and Acknowledgements

1. Bendall SC, Nolan GP, et al. *Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum,* Science (New York, N.Y.), May 6, 2011.
2. El-ad David Amir et al, *viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia,* Nature Biotechnology, May 19, 2013.
3. Laurens van der Maaten and Geoffrey Hinton, *Visualizing Data using t-SNE*, Journal of Machine Learning Research, August, 2011.
4. Max Vladymyrov and Miguel A. Carreira-Perpi*, Partial-Hessian Strategies for Fast Learning of Nonlinear Embeddings,* Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012.
5. Geoffrey Hinton and Sam Roweis*, Stochastic Neighbor Embedding.*

6. The authors would like to acknowledge and thank Karen Sachs and Andrew Gentes for guidance in the project and for access to the data.