
2D Visualization of Immune System Cellular Protein Data by Nonlinear Dimensionality Reduction

Andre Esteva

ESTEVA@STANFORD.EDU

Stanford University, Electrical Engineering, 496 Lomita Mall, Durand 196, Stanford, CA 94305 USA

Anand Sampat

ASAMPAT@STANFORD.EDU

Stanford University, Electrical Engineering, 450 Serra Mall, Stanford, CA 94305 USA

Amit Badlani

ABADLANI@STANFORD.EDU

Stanford University, Electrical Engineering, 450 Serra Mall, Stanford, CA 94305 USA

Abstract

We present in this paper a way to effectively visualize multi-dimensional immune system cellular data by means of nonlinear methods. We find that Stochastic Neighbor Embedding (SNE), and its variations, t-SNE and s-SNE, to be most effective at successfully mapping clusters of points into a two dimensional embedding space while preserving both the structure between similar points and the disparity between different clusters. Using a centroid-based metric that relabels points according to the cluster centroid to which they are closest, we conclude that SNE works significantly better than linear and spectral methods. In addition, by using an optimization approach for SNE similar to Newton's Method, but with the Hessian of the objective function, $\nabla^2(E)$, replaced by its first term, we are able to run the SNE variants and EE two orders of magnitude faster than with standard optimization.

1. Introduction

1.1. Immune Cell Data

In the field of cancer immunology, scientists use the protein content of immune system cells as a way to identify a cell's corresponding type. For example, immune system cells, which are contained in bone marrow, are comprised of a variety of cell types, and to a

large degree are uniquely identifiable by the proteins they contain. Highly sophisticated methods have been developed that process cells and return information on the types and quantities of proteins expressed in those cells. This data can then be viewed by an expert in the field and categorized. The laborious process of viewing the different dimensions of protein expression and categorizing a cell's type is known as gating. Figure 1.1, below, taken from (Amir et al., 2013) shows this graphically.

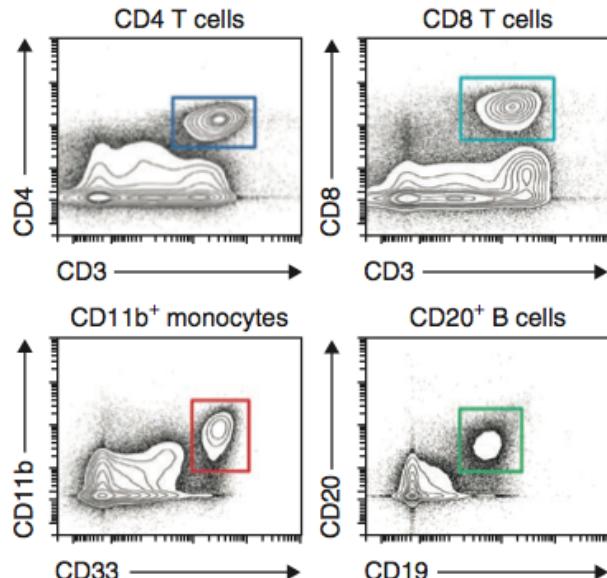


Figure 1. Strategy for cell gating: Two single dimensions of a cell are viewed at a time and through an iterative process the cell is classified

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.
Copyright 2012 by the author(s)/owner(s).

1.2. Project Goals and Metrics

It is of interest to cancer immunologists to find structure within multi-dimensional protein expression space and map it onto a lower dimensions (referred to henceforth as a map space) for ease of visualization and understanding. As cells change and evolve, so too do the types and quantities of the proteins they express. This leads to a shifting of their representation in multi-dimensional space which can be tracked. Dimensionality reduction of original biological data coupled with a metric for how well the projection represents the original data would provide biologists with a powerful tool for understanding the structure of their data. To address these challenges we demonstrate:

- The application of linear and non-linear methods of dimensionality reduction of multi-dimensional protein data
- A metric-based comparison of how each algorithm performs

2. Data Representation

2.1. Data Acquisition

Mass cytometry is a single-cell multiparametric protein detection technology based on inductively coupled plasma mass spectrometry. It is an extension of flow cytometry in which antibodies are tagged with isotopically pure rare earth elements allowing simultaneous measurement of greater than 40 parameters while circumventing the issue of spectral overlap. In single-cell droplet form, the cells are passed through an elemental mass spectrometer and an integrator to generate an $m \times p$ matrix where m is the number of cells processed and p is the number of distinct proteins contained in the cell set. These matrices are stored as .FCS files in online databanks which we have been granted access to. Figure 2.1 shows this process.

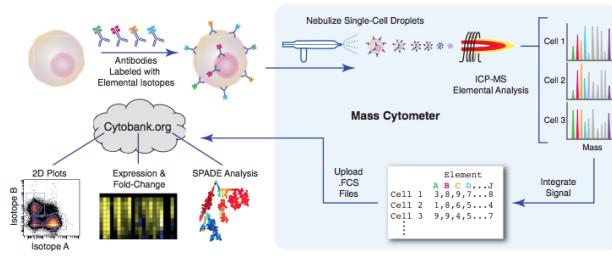


Figure 2. Overview of data acquisition, from extraction of cellular protein counts to storage in online databanks

Table 1. Classification of Cell Types and their Corresponding Sub-Types

CELL TYPES	SUB-TYPES
STEM CELL	HSC, MPP, CMP, GMP, MEP
B CELLS	PLASMA, PRE-B-I, PRE-B-II, IMMATURE, MATURE CD38 LOW, MATURE CD38 MID
T CELLS	MATURE CD4+, MATURE CD8+ NAIVE CD4+, NAIVE CD8+
NK	-
PDC	-
MONOCYTES	CD11B -, CD11B HIGH, CD11B MID

Table 1 lists the cell types and subtypes that are parsed using this method.

2.2. Feature Selection

The $p = 41$ protein counts collected for each cell that passes through mass cytometry is comprised of both intracellular and surface proteins. These two types play fundamentally different roles in cell identification. Surface proteins are semi-permanent markers that last for significant periods of time relative to the lifetime of a cell, whereas intracellular proteins are highly transient and can change quickly. This is analogous to classifying a person based on where they live (semi-permanent) versus what they wore on a particular day (transient). Understanding this, we select as our feature space the $n = 17$ surface protein markers of the cell data.

3. Methods

The datafiles provided contain cell counts on the order of tens of thousands where we consider each cell to be a point in \mathbb{R}^n . To simplify our algorithms and account for matrix size differences in the difference .FCS files we run our algorithms on equally sized portions of different cell data. In particular, if we let S be the set of all cell sub-types as defined in Table 1, $S \in \mathbb{S}$ be some subset of interest with cardinality $|S|$, and N some fixed positive integer, then by taking N rows from each sub-type $s \in S$ we form a matrix $M \in \mathbb{R}^{N|S| \times n}$ on which we can run algorithms quickly and without giving unfair weighting to a particular cell sub-type.

We consider various algorithms which project sets of data in \mathbb{R}^n onto \mathbb{R}^2 for easy visualization.

3.1. Linear Methods

Linear methods such as Principal Components Analysis (PCA) and multidimensional scaling are straightforward and standard ways of achieving dimensionality reduction. The caveat is the requirement that the data be linearly separable in the space being considered. Unlike SVMs, which have the ability to project data to higher dimensions in order to linearly separate it, PCA runs in the original dimension of the data. Given the nature of this data and the fact that cellular protein counts can vary largely between cell types, PCA is not an optimal method for visualizing these sorts of problems. We apply PCA and demonstrate that its visualization is quite poor, as expected.

3.2. Spectral Methods

Spectral methods like Locally Linear Embedding (LLE), ISOMAP, and Laplacian Eigenmaps (LE) seem like better candidates for dimensionality reduction of problems of this sort. For example, LE tries to preserve the local structure of the data while keeping the scale of the embedding data fixed, leading to good results on toy problems such as the swiss roll. We apply LLE and ISOMAP to this data set and show that the results are even worse than PCA.

3.3. Nonlinear Methods

Nonlinear methods, as demonstrated in the section below, are the most effective of the algorithms used for visualization of cellular protein data. Cellular protein data is largely believed to exist in multi-dimensional space in gaussian-distributed clumps that can have a high degree of overlap. These gaussians can exhibit high degrees of variation in their average and variance values. As such, we consider the variants of SNE, t-SNE and s-SNE, along with Elastic Embedding (EE) as prime candidates for effective visualization. These algorithms minimize objective functions that mimick attractive and repulsive forces in the mapping space, which keeps the images of nearby objects close while pushing all images apart from each other. They take the form $E(X, \lambda) = E^+(X) + \lambda E^-(X)$ where $E^+(X)$ is the attractive term, $E^-(X)$ the repulsive term, $\lambda \geq 0$ is a fixed parameter, and $X = \{x_n\}$ is the set of points in the mapping space. The repulsive term significantly improves them over spectral methods like LE. For comparison, we outline the objective functions E of these algorithms.

Given two neighborhood graphs,

$$w_{nm}^+ = \exp(-\frac{1}{2}\|y_n - y_m\|^2/\sigma^2) \quad (1)$$

$$w_{nm}^- = \|y_n - y_m\|^2 \quad (2)$$

And conditional probability matrices $\{p_{nm}\}$ and $\{q_{nm}\}$ in both the original data space (points y_n) and mapping space (points x_n), respectively that a point selects any other point as its neighbor:

$$p_{nm} = \frac{\exp(-\|y_n - y_m\|^2/2\sigma^2)}{\sum_{k \neq n} \exp(-\|y_n - y_k\|^2/2\sigma^2)} \quad (3)$$

$$q_{nm} = \frac{\exp(-\|x_n - x_m\|^2/2\sigma^2)}{\sum_{k \neq n} \exp(-\|x_n - x_k\|^2/2\sigma^2)} \quad (4)$$

The objective functions for LE, SNE, t-SNE, s-SNE, and EE are given by,

$$\begin{aligned} E_{LE}(X) &= \sum_{n,m=1}^N w_{nm} \|x_n - x_m\|^2 \\ E_{SNE}(X) &= \sum_{n,m=1}^N p_{nm} \|x_n - x_m\|^2 + \sum_{n=1}^N \log \sum_{m \neq n}^N \exp(-\|x_n - x_m\|^2) \\ E_{s-SNE}(X) &= \sum_{n,m=1}^N p_{nm} \|x_n - x_m\|^2 + \log \sum_{n,m=1}^N \exp(-\|x_n - x_m\|^2) \\ E_{t-SNE}(X) &= \sum_{n,m=1}^N p_{nm} \log(1 + \|x_n - x_m\|^2) \\ &\quad + \sum_{n,m=1}^N (1 + \|x_n - x_m\|^2)^{-1} \\ E_{EE}(X) &= \sum_{n,m=1}^N w_{nm}^+ \|x_n - x_m\|^2 \\ &\quad + \lambda \sum_{n,m=1}^N w_{nm}^- \exp(-\|x_n - x_m\|^2) \end{aligned}$$

3.4. Fast Learning of Nonlinear Embeddings

In order to speed up the SNE and EE algorithms, we employ a strategy for optimization of the objective functions in the "spectral direction" (Vladymyrov & Carreira-Perpinan, 2012). Optimization typically involves computing the gradient $g = \nabla E$ of the objective function and computing a descent direction as p by solving the linear system $Bp = -g$ using some positive definite B . When $B = I$ this is gradient descent, and for $B = \nabla^2 E$ we have Newton's Method. Descending in the spectral direction is accomplished by letting B be the first term of $\nabla^2 E$ and, for small data sets ($\approx 10,000$ points) is two orders of magnitude faster than gradient descent and Newton's Method.

3.5. Metric on Quality of Visualization

We quantify the quality of the various algorithms at accurately segregating different cell types in the map space using a simple re-labeling technique. After projection, the averages of all the points (centroids) of each cell sub-type are calculated. Each point is then relabeled based on whichever centroid it is closest to. Formally, for cell sub-type i with $P^{[i]}$ points $\{x^{[i]}\}$ in the projection space, the centroid is given by $\mu^{[i]} = \frac{1}{P^{[i]}} \sum_j x_j^{[i]}$, and each point x is relabeled as sub-type $i = \arg \min_j \|\mu^{[j]} - x\|$. The error of a visualization with N points is then calculated as $err = \frac{1}{N} \sum_j^N 1\{\text{originalLabel}(x_j) \neq \text{newLabel}(x_j)\}$.

4. Results

4.1. Linear Methods - PCA

We consider PCA over all 20 cell sub-types with 400 cells per cell sub-type and obtain a 62.18% re-labeling error. Figure 3 below shows the visualized data. It is clear that the overlap between points renders the map meaningless. The colored stars in the figures represent the centroids of each sub-type.

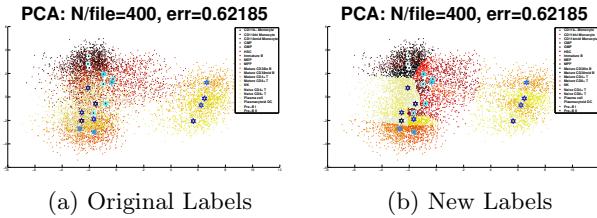


Figure 3. Principal Components Analysis on all cell sub-types with 400 cells per subtype

Even when we reduce the number of sub-types to 7 (Monocytes and T cells), the performance is still abysmal at 48.43% error, as shown in Figure 4.

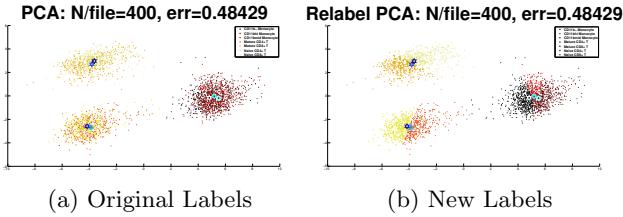


Figure 4. Principal Components Analysis on Monocytes and T Cells with 400 cells per subtype

4.2. Spectral Methods - ISOMAP

We illustrate in Figure 5 the spectral method ISOMAP on this data both for all cell sub-types ($N/\text{type} = 100$) and for just monocytes and T cells ($N/\text{type} = 400$) combined. In each case, the algorithm yields no discernible segmented high-level structure.

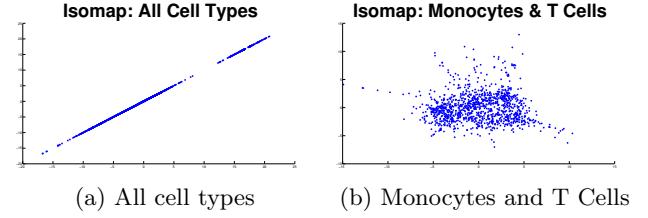


Figure 5. ISOMAP with 400 cells per subtype

4.3. Nonlinear Methods

We begin by considering t-SNE and sampling 400 cells from each cell type. We see in Figure 6 that the algorithm clusters the points reasonably well, even though it attains a relatively high error of 24.33%. On the other hand, when we only work with Monocytes and T Cells (Figure 7), we obtain a 7.46% error - a 7-fold improvement on PCA for the same data set.

Next, we consider s-SNE on a similar sample of 400 cells taken from either all cell types (Figure 8) or Monocytes and T Cells (Figure 9). We see that the visualization appears to perform as well as tSNE, yet the metric shows that its performance is slightly worse. sSNE attained errors of 28.57% and 8.18% for all cell types and the subset of Monocytes and T Cells, respectively.

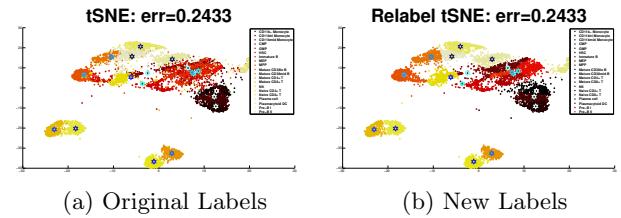


Figure 6. tSNE on all cell sub-types with 400 cells per subtype

We now look to EE, the last of the nonlinear methods we are considering. Once again, we take a sample of 400 cells from (i) all cell types and (ii) Monocytes and T Cells.

Table 2 summarizes the errors for the different algorithms.

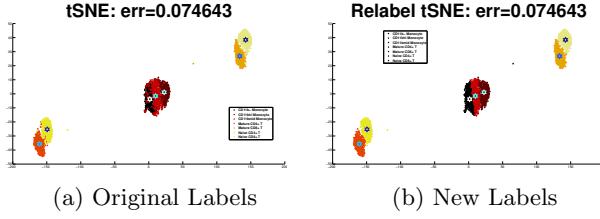


Figure 7. tSNE on Monocytes and T Cells with 400 cells per subtype

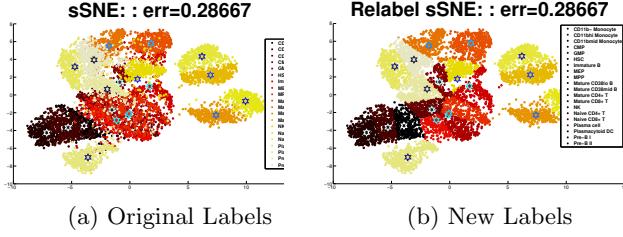


Figure 8. sSNE on all cell subtypes with 400 cells per subtype

4.4. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the L^AT_EX bibliographic facility, use `natbib.sty` and `icml2012.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (2011). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (Bendall et al., 2011). List multiple references separated by semicolons (Bendall et al., 2011; Crammer & Singer, 2002; Bendall et al., 2011). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (Bendall et al., 2011).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section ?? for detailed

Table 2. Algorithm Errors

ALGORITHM	ALL CELLS	MONOCYTES & T CELLS
PCA	62.18%	48.42%
tSNE	24.33%	7.46%
sSNE	28.67 %	8.18%
EE	41.84%	12.46%

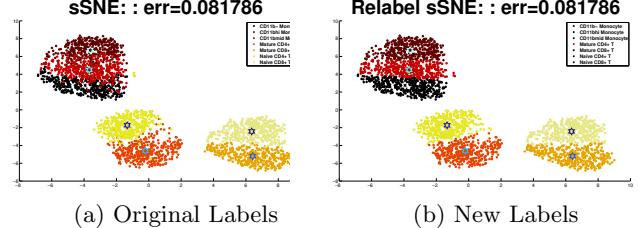


Figure 9. sSNE on Monocytes and T Cells with 400 cells per subtype

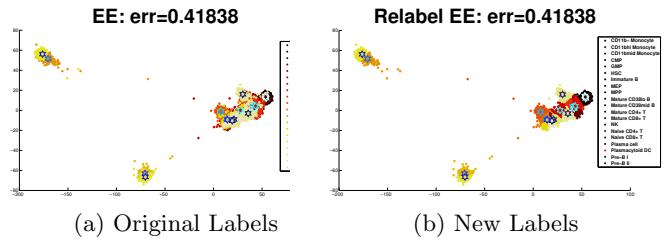


Figure 10. EE on Monocytes and T Cells with 400 cells per subtype

instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (Bendall et al., 2011), conference publications (Bendall et al., 2011), book chapters (Bendall et al., 2011), books (Bendall et al., 2011), edited volumes (Bendall et al., 2011), technical reports (Crammer & Singer, 2002), and dissertations (Bendall et al., 2011).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

5. Acknowledgments

We would like to thank Karen Sachs (Stanford University) and Andrew Gentes (Stanford University) for their support and guidance throughout this project, as well as for providing the data required for its completion.

References

Amir, El-ad David, Davis, Kara L, Tadmor, Michelle D, Simonds, Erin F, Levine, Jacob H, Ben-

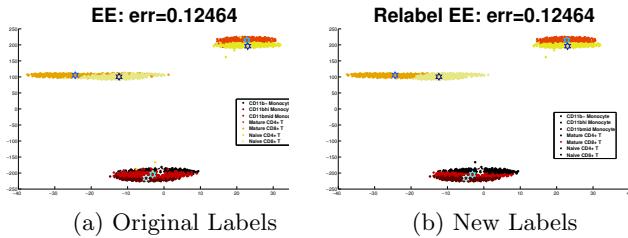


Figure 11. EE on Monocytes and T Cells with 400 cells per subtype

dall, Sean C, Shenfeld, Daniel K, Krishnaswamy, Smita, Nolan, Garry P, and Pe'er, Dana. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, 31(6):545–552, June 2013.

Bendall, S C, Simonds, E F, Qiu, P, Amir, E a D, Krutzik, P O, Finck, R, Bruggner, R V, Melamed, R, Trejo, A, Ornatsky, O I, Balderas, R S, Plevritis, S K, Sachs, K, Pe'er, D, Tanner, S D, and Nolan, G P. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science*, 332(6030):687–696, May 2011.

Crammer, Koby and Singer, Yoram. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, March 2002.

Vladymyrov, Max and Carreira-Perpinan, Miguel. Partial-Hessian strategies for fast learning of nonlinear embeddings. *arXiv preprint arXiv:1206.4646*, 2012.