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 it's 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. By using an optimization approach for SNE that follows the 'spectral direction' of descent, we are able to run the SNE variants and EE two orders of magnitude faster than with standard optimization. Finally, we demonstrate superior classification of data by first reducing its dimension then applying supervised learning using a C-SVC SVM and ν-SVC SVM.

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 cells corresponding type (1). 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.

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 (refered to henceforth as a map space) for ease of visualization and understanding. As cells change and evolve, so 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

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

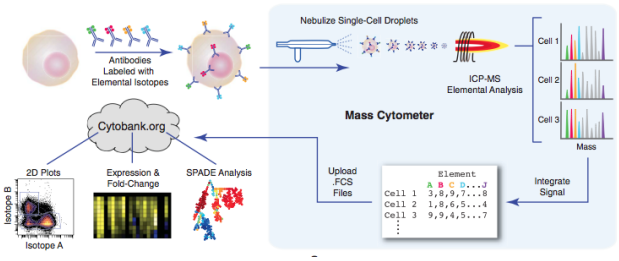


Figure 1. Overview of data acquisition, from extraction of cellular protein counts to storage in online databanks

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

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, taken from (1) shows this process. Table 1 lists the cell types and subtypes that are parsed using this method.

2.2. Feature Selection

The $p = 41$ types of proteins collected for each cell using mass cytometry are 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 different .FCS files we run our algorithms on equally sized portions of different cell data. In particular, if we let $\mathbb{S}$ 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 $M \in \mathbb{R}^{N|S| \times n}$. This allows us to run algorithms that accomplish dimensionality reduction from $\mathbb{R}^n$ onto $\mathbb{R}^2$ quickly and without giving unfair weighting to a particular cell sub-type.

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, rather counterintuitively, 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 celllular 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 (2), t-SNE (3) and s-SNE (4), along with Elastic Embedding (EE) (5) 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 image clusters 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.

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" (6). 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 definitie 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 ($< 10{,}000$ points) is two orders of magnitude faster than gradient descent and Newton's Method.

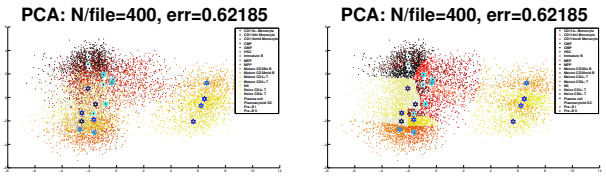
3.5. Metric on Quality of Visualization

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

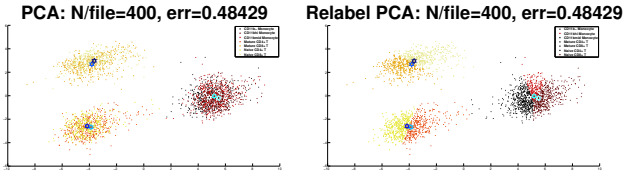
We consider PCA over all 20 cell sub-types with 400 cells per cell sub-type and obtain a 62.18% re-labeling error. Figure 2 below shows the visualized data. It is clear that the overlap between points renders the map meaningless. The colored stars in the figures represent the centroids of each sub-type.



(a) Original Labels (b) New Labels

Figure 2. Principal Components Analysis on all cell subtypes 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 3.

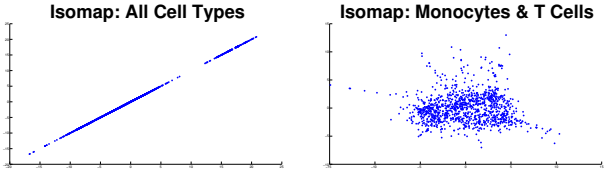


(a) Original Labels (b) New Labels

Figure 3. Principal Components Analysis on Monocytes and T Cells with 400 cells per subtype

4.2. Spectral Methods - ISOMAP

We illustrate in Figure 4 the spectral method ISOMAP on this data both for all cell sub-types (N/type = 100) and for just monocytes and T cells (N/type = 400) combined. In each case, the algorithm yields no discernible segmented high-level structure.



(a) All cell types (b) Monocytes and T Cells

Figure 4. 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 5 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 6), we obtain a 7.46% error - a 7-fold improvement on PCA for the same data set.

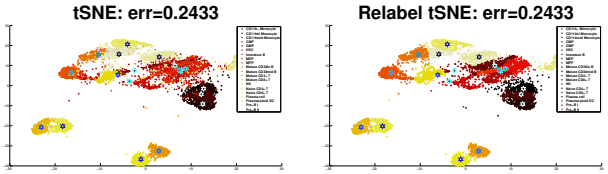
Next, we consider s-SNE on a similar sample of 400 cells taken from either all cell types (Figure 7) or Monocytes and T Cells (Figure 8). We see that the visualization appears to perform as well as tSNE, yet the metric shows that its performance is slightly worse. sSNE attained errors of 28.57% and 8.18% for all cell types and the subset of Monocytes and T Cells, respectively.

We now look to EE, the last of the nonlinear methods we are considering, and run the algorithm with the same inputs as the previous methods. Once again, we take a sample of 400 cells from (i) all cell types and (ii) Monocytes and T Cells. Figure 9 shows the visualization of all cell subtypes and Figure 10 shows it for Monocytes and T Cells. Though the algorithm works better than PCA, it performs substantially worse than the SNE variants with an error of 41.838% for all cells and 12.46% for just Monocytes and T Cells. In addition, there is less spacing between clusters than with the SNE variants.

Table 2 summarizes the errors for the different algorithms.

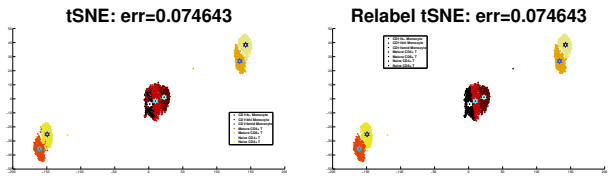
Table 2. Algorithm Errors

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%



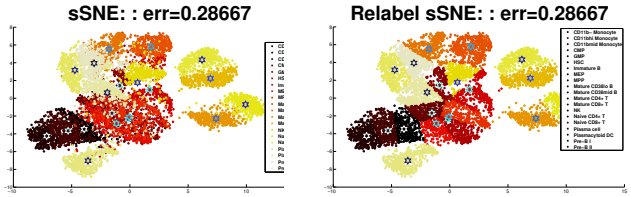
(a) Original Labels (b) New Labels

Figure 5. tSNE on all cell sub-types with 400 cells per subtype



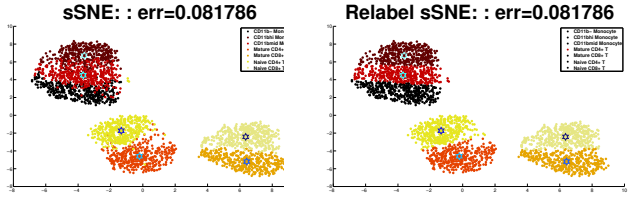
(a) Original Labels (b) New Labels

Figure 6. tSNE on Monocytes and T Cells with 400 cells per subtype



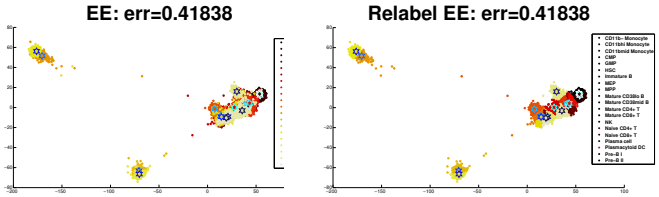
(a) Original Labels (b) New Labels

Figure 7. sSNE on all cell subtypes with 400 cells per subtype



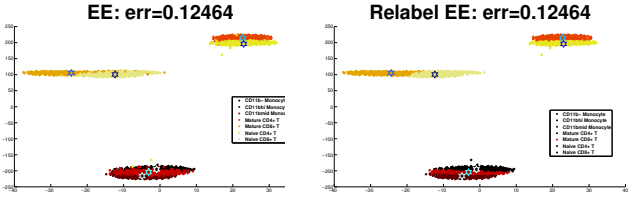
(a) Original Labels (b) New Labels

Figure 8. sSNE on Monocytes and T Cells with 400 cells per subtype



(a) Original Labels (b) New Labels

Figure 9. EE on all cell subtypes with 400 cells per subtype



(a) Original Labels (b) New Labels

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

Table 3. SVM Classification Rate

C-SVC	Original	PCA	t-SNE	EE
radial	5.2%	91.3%	78.3 %	59.1 %
polyn.	4.6%	92.3%	74.6 %	55.6 %
linear	5.53 %	93.3%	74.07%	56.9 %
sigmoid	5.8%	70.8%	8.4 %	7.6 %
nu-SVC	**Original**	**PCA**	**t-SNE**	**EE**
radial	5%	90%	76.4 %	58 %
polyn.	4.4%	96.2%	66.8 %	51.2 %
linear	5.8 %	91.4%	68.2%	46.8 %
sigmoid	4.6%	71%	26.4 %	9.8 %

4.4. Classification

Shifting our attention from visualization to classification, we compare the classification effectiveness of C-SVC and ν-SVC SVM (per the LIBSVM library (7)) run on the original data to the projected data, and reach a counter-intuitive result. The SVM applied to the original data fails entirely, with low success rates. The SVM applied to the projected data classifies with significantly higher accuracies, depending on the dimensionality reduction algorithm used. Table 3 shows the success rates of C-SVC and ν-SVC with various kernels (with s-SNE omitted as its values are very close to t-SNE). We find that the best classification is achieved, with a success rate of 96.2%, when using ν-SVC based SVM with a polynomial kernel on the PCA-reduced data.

5. Conclusion

Human immune system cells are comprised of a variety of cell types and sub-types which are characterized by the proteins contained within them. Proteins embedded on the cell's surface are particularly effective at characterizing the type of cell, and through 'gating', biologists can label a cell's type from these proteins by plotting various permutations of two types of proteins at time. We apply dimensionality reduction considering only the surface proteins as features to aid the visualization of cells in this multi-dimensional protein space. We apply Principle Components Analysis, ISOMAP, elastic embedding, symmetric-stochastic neighbor embedding, and student t-distributed stochastic neighbor embedding to this problem and use a centroid-based relabeling metric to quantify the effectiveness of each visualization. We demonstrate the superiority of the nonlinear methods, with t-SNE being the most effective, while simultaneously showing that PCA and ISOMAP yield poor results. We further conclude that the data is nonlinear and that it supports the belief that clusters of healthy cells in this protein space are gaussian-distributed.

In addition, we demonstrate a substantial improvement in SVM-based classification of data by first reducing the dimensionality of the data and then applying classification on the reduced data. We find that by combining PCA with ν-SVC based SVM and a polynomial kernel a 96.2% classification rate is achieved.

Our work provides a general tool which biologists can use to visualize and better understand multi-dimensional cellular protein data. This is a first step towards automated cell-labelling and further work could yield a scalable platform by which both healthy and unhealthy immune system cells could be easily categorized, leading to a cancerous-cell detection system. Additionally this can be used by researchers to track how immune system cells' protein content changes as they mature, evolve, and differentiate.

6. 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

[1] Bendall, S. C. *et al.* Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science* **332**, 687–696 (2011).

[2] Hinton, G. E. & Roweis, S. T. 1. Hinton, G. E. & Roweis, S. T. Stochastic neighbor embedding. Advances in neural information ... (2002). - Google Scholar. *Advances in neural information ...* (2002).

[3] Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 85 (2008).

[4] Cook, J., Sutskever, I. & Mnih, A. Visualizing similarity data with a mixture of maps. *International ...* (2007).

[5] Carreira-Perpinán, M. A. The Elastic Embedding Algorithm for Dimensionality Reduction. *ICML* (2010).

[6] Vladymyrov, M. & Carreira-Perpinan, M. Partial-Hessian strategies for fast learning of nonlinear embeddings. *arXiv preprint arXiv:1206.4646* (2012).

[7] Chang, C.-C. & Lin, C.-J. LIBSVM. *ACM Transactions on Intelligent Systems and Technology* **2**, 1–27 (2011).