2D visualization of high-dimensional cellular protein data for cancer detection

Amit Badlani
Dept. of Electrical Engineering Stanford University
abadlani@stanford.edu

Andre Esteva
Dept. of Electrical Engineering Stanford University
andre.esteva@gmail.com

Anand Sampat
Dept. of Electrical Engineering Stanford University
asampat@stanford.edu

Abstract

1 Introduction and Problem Statement

In the field of cancer immunology, scientists use the protein content of immune system cells as a way to identify a cells 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, each type is uniquely identifiable by both intracellular proteins (IP) and surface proteins (SP). 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 is known as gating.

Is there order to this data? Does cell type, as a function of protein expression, have some structure in multi-dimensional protein space which can be understood and segmented? Can this data be mapped onto a different multi-dimensional space (MDS) where structure is better defined? These are questions of interest to cancer immunologists. If a well-defined structure exists which separates healthy immune system cells from other types of cells, such structure could be leveraged to identify cancer cells.

Cells evolve. Stem cells become progenitor cells, which become monocytes, etc. As they change, so do the types and quantities of proteins that they express. This leads to a shifting of their representation in some MDS, which can be tracked and understood.

The final goal of this project is two-fold:

- Develop machine learning algorithms that probabilistically tag single bone marrow cells as cancerous after being trained on the protein expression levels of healthy immune system cells.
- Understand the underlying structure in MDS of cell types as identified by their protein content.

2 Data Representation and Significance

There are two techniques which are used in the medical field to get the protein data for each of the cell types. These are flow cytometry and mass cytometry and they are described below.

Flow cytometry is a laser-based, biophysical technology employed in cell counting, cell sorting, biomarker detection and protein engineering, by suspending cells in a stream of fluid and passing them by an electronic detection apparatus. The flow cytometers are used for this purpose, which are able to analyze several thousand particles every second, in "real time," and can actively separate and isolate particles having specified properties.

Mass cytometry or CyTOF (DVS Sciences) 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.

Mass cytometry data is recorded in tables that list, for each cell, the signal detected per channel, which is proportional to the number of antibodies tagged with the corresponding channel's isotope bound to that cell. All this data is formatted as FCS files. We use an FCS that converts the data in these files into matrices, which we operate on.

Once we have the input data we need to read it out in a particular format in order before operating on it. There are two different kinds of data that the biologists like to look at. First is the normal cell readout, which we use as in input in our Matlab code; and the other is the super stimulated cell readout. Basal gives a normal cell readout whereas PVO-4 is a stimulus that produces a strong and exaggerated protein expression.

The table below shows the different names of the cells and their respective cell types.

The flowchart below shows how the different cell types evolve. The HSC are the first type of cells which divide and form the MPP cells. The MPP cells further divide and form LMP cells as well as GMP cells. The GMP cells further evolve into pDC and Monocytes, whereas the LMPs evolve produce B-cells, T-cells and NK type cells.

3 Machine Learning Techniques and Application

Our dataset contains a huge dataset of various cells and a metric for how much a given protein is represented in that cell. Specifically the data matrix has dimension $n \times p$ where $n =$ number of cells considered and $p =$ number of proteins for which have a number representing the amount of protein expression in the cell. The initial dataset is in $\mathbb{R}^p$. In order to reduce this to some $\mathbb{R}^k$ (e.g. $\mathbb{R}^2$ or $\mathbb{R}^3$), we use various linear and non-linear techniques to best map the distances of points in a higher dimensional space $\mathbb{R}^p$ to a lower dimension $\mathbb{R}^k$ we can visualize.

3.1 Linear Methods

3.1.1 Principal Component Analysis (PCA)

Principal component analysis allows us to find k principal components in order of decreasing influence of samples to the overall mean. In order to run this algorithm, we consider $\{x^{(i)}; i = 1, \ldots, m\}$ where $\{1, \ldots, m\} \in \mathbb{R}^p$ where the set of m proteins is some subset of the protein set p that are most representative in detecting cancerous cells (in our case 41 proteins). In particular each $x_j^{(i)}$ is a measure of how much the protein i is represented in cell j. Since we cannot visualize this data of j cells in $\mathbb{R}^m$ we use PCA to preserve the variance by converting these $x^{(i)}$'s into principal components (i.e. new unitless axes that contain most of the information within just a first few principal components).

Before we apply the algorithm, however, we need to preprocess the data by subtracting the mean and normalizing the values. The mean centering is key to ensure the first principal component doesn't just represent the mean and the normalizing ensures each protein is weighted equally in the algorithm (going forward we may want to weight them as some proteins may be more relevant in determining certain cell types).

Table 1: Cell names and their respective cell types

PART	**DESCRIPTION**
HSC	Stem Cells/Progenitors
MPP	Stem Cells/Progenitors
CMP	Stem Cells/Progenitors
GMP	Stem Cells/Progenitors
MEP	Stem Cells/Progenitors
Plasma	B Cells
pre-B-I	B Cells
pre-B-II	B Cells
Immature B	B Cells
Mature CD38 low B	B Cells
Mature CD38 mid	B Cells
Mature CD4 + T	T Cells
Mature CD8 + T	T Cells
Naive CD4 + T	T Cells
Naive CD8 + T	T Cells
NK	NK
Plasmacytoid DC	pDC
CD11B - Monocyte	Monocytes
CD11B high Monocyte	Monocytes
CD11B mid Monocyte	Monocytes

Finally, we apply PCA, which maximizes the variance of projections $\frac{1}{m}\sum_{i=1}^{m}(x^{(i),T}u)^2$ by finding the eigenvectors of the covariance matrix $\frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i),T}$ and finds the k eigenvectors $u_1,\ldots,u_k$ which then define the new axes (i.e. the principal components) where:

$$y^{(i)} = \begin{bmatrix} u_1^T x^{(i)} \\ \vdots \\ u_1^T x^{(i)} \end{bmatrix} = PC^{(i)}$$

3.1.2 Classical Multidimensional Scaling (CMDS)

Similar to PCA, MDS aims to replot each data point in a reduced dimension. As above, assume we want to reduce points in $\mathbb{R}^m$ to $\mathbb{R}^k$. Specifically we are considering $k = 2$ and $k = 3$. The approach of MDS however is to output an embedding of points $x_1,\ldots,x_j \in \mathbb{R}^k$ where j is the number of cells (i.e. number of samples).

For any two of the j given cells x_a and x_b in the original space $\mathbb{R}^m$we define a distance $\delta_{a,b}$ between the two points in the higher dimension. In classical CMDS, this is just the Euclidean distance between the points (e.g. $\sqrt{(x_{a,1}-x_{b,1})^2+(x_{a,2}-x_{b,2})^2}$ for $\mathbb{R}^2$). Thus the constraint on the embedding of points $x_1,\ldots,x_j \in \mathbb{R}^k$ is that $||x_a - x_b|| \approx \delta_{a,b}$. In other words we want to solve the optimization problem with the constraint:

$$\min_{x1,\ldots,x_j} sum_{a<b}(||x_a - x_b|| - \delta_{a,b})^2$$

3.1.3 Stochastic Neighbor Embedding (SNE)

3.2 Non-linear Methods

3.2.1 t-SNE

3.2.2 vi-SNE

4 Current Results

Below we present a PCA analysis of the Basal data. Since surface proteins are more standard proteins for cell type identification, we compare PCA run on surface protein data to PCA run on all protein data. Additionally, we show PCA for all cell types as well as PCA for individual subsets of cells (i.e. B Cells, TCells, Monocytes, and Stem Cells).

5 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

5.1 Citations within the text

Citations within the text should be numbered consecutively. The corresponding number is to appear enclosed in square brackets, such as [1] or [2]-[5]. The corresponding references are to be listed in the same order at the end of the paper, in the **References** section. (Note: the standard BIBTEX style `unsrt` produces this.) As to the format of the references themselves, any style is acceptable as long as it is used consistently.

Figure 1: Sample figure caption.

Acknowledgments

We would like to thank Karen Sachs and Andrew Gentles for providing the data and guidance for the project.

References

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to ‘small’ (9-point) when listing the references. **Remember that this year you can use a ninth page as long as it contains *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.