Comp683: Computational Biology

Lecture 6

January 29, 2025

Today

- Graph embedding with node2vec
- Intro to single-cell; cell-types, gating, featurizing

Graph Representation Learning

- There is a lot of recent hype in the graph world about feature learning from graphs or learning a representation for each node in a continuous vector space
- For a particular node, u, the objective is to learn some representation in d dimensions, $f(u)$, with $f(u) \in \mathbb{R}^d$

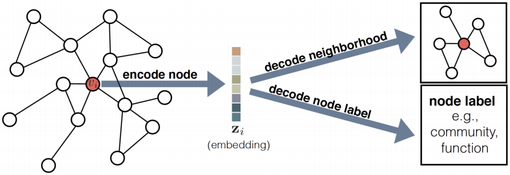


Figure: from Hamilton *et al.* ArXiv. 2018

Get Graph Partitioning for Free from Embedding

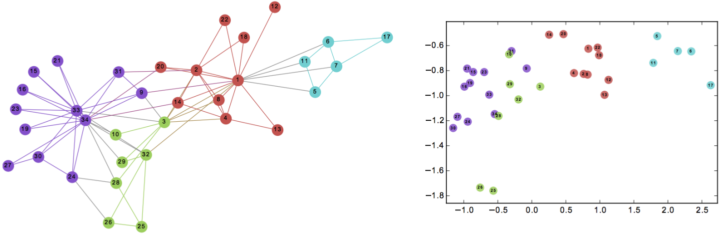


Figure: from Perozzi *et al.* KDD. 2014

The Random Walk Based Approach

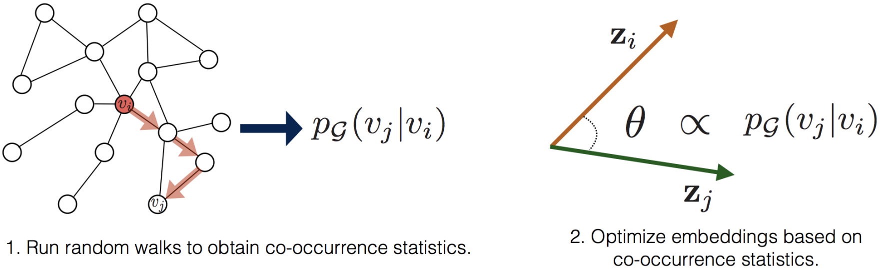


Figure: from Hamilton *et al.* ArXiv. 2018

Encoding the 'Role' of a Node with Node2Vec

- Node2Vec allows for control of whether the embedding captures between-node structural similarity (neighborhoods) or role (the types of nodes in connection with)

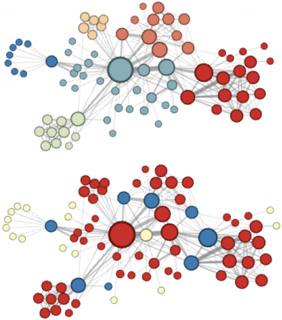


Figure: from Grover and Leskovec. KDD. 2016

Other Similar Approaches

- LINE $\rightarrow$ `https://arxiv.org/abs/1503.03578`
- DeepWalk $\rightarrow$ `https://arxiv.org/abs/1403.6652`
- Node2Vec $\rightarrow$ `https://arxiv.org/abs/1607.00653`
- Splitter $\rightarrow$ `https://arxiv.org/pdf/1905.02138.pdf`

Flexible Notion of Neighborhood (Breadth First)

When defining a neighborhood set N_s of some node u, there are two possible strategies. The following example assumes sampling $k = 3$ nodes.

- **Breadth-first Sampling:** Find the most immediate neighbors of node u and sequentially sample additional nodes according to their distance from u. In the image below, if we are selecting k neighbors, this selects s_1, s_2, s_3.

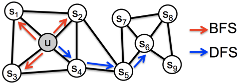


Figure 1: BFS and DFS search strategies from node u ($k = 3$).

Figure: from Leskovec and Grover. KDD. 2016

Flexible Notion of Neighborhood (Depth First)

- **Depth-first Sampling:** Select nodes sampled at sequentially further distances from the source node, u. According to the example below, nodes s_4, s_5, s_6 are selected.

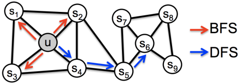


Figure 1: BFS and DFS search strategies from node u ($k = 3$).

Figure: from Leskovec and Grover. KDD. 2016

Node2Vec Objective

Starting with a graph, G with $G = (V, E)$, then the task if to learn a mapping function, f with $f : V \rightarrow \mathbb{R}^d$. Here, d is the number of dimensions of of f. Ultimately, f is a matrix of size $|V| \times d$. $\mathcal{N}_s(u)$ is further defined as the nodes that are sampled in the walk from u. Then the objective is to find an f such that,

$$\sum_{u \in V} \log P(\mathcal{N}_s(u) \mid f(u)) \tag{1}$$

is as large as possible. Here, $\mathcal{N}_s(u)$ is conditioned on the feature representation of u, $f(u)$.

Conditional Independence Assumption

$$P(\mathcal{N}_S(u)|f(u)) = \prod_{n_i \in \mathcal{N}_s(u)} P(n_i \mid f(u)). \tag{2}$$

Conditional Likelihood for each Source, Neighborhood Pair

To model each source, neighborhood pair and given the conditional independence assumption introduced on the previous slide,

$$P(n_i \mid f(u)) = \frac{\exp(f(n_i) \cdot f(u))}{\sum_{v \in V} \exp(f(v) \cdot f(u))} \tag{3}$$

Varying Neighborhood According to DFS or BFS

As we have pointed out, $\mathcal{N}_S(u)$ is not restricted to the immediate neighbors of u. The authors define a flexible notion of a random walk on the graph that allows for sampling in either a DFS or BFS way. Assuming the random walk has just traversed edge (t, v) and now resides at node v. The transition probability from v to x (π_{vx}) can be written as,

$$\pi_{vx} = \alpha_{pq}(t, x) \cdot w_{vx} \tag{4}$$

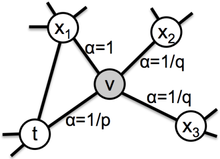


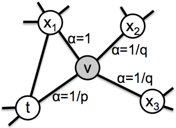
Figure: from Grover and Leskovec, KDD. 2016

Varying Neighborhood According to DFS or BFS

$$\pi_{vx} = \alpha_{pq}(t,x) \cdot w_{vx} \tag{5}$$

Here w_{vx} is the edge weight between nodes v and x or simply $\{0,1\}$ if the graph is unweighted. d_{tx} denotes shortest path distance between t and x (we are at v). A user can specify p or q.

$$\alpha_{p,q}(t,x) = \begin{cases} 1/p & \text{if} d_{tx} = 0 \\ 1 & \text{if} d_{tx} = 1 \\ 1/q & \text{if} d_{tx} = 2 \end{cases}$$



Intuition Behind p

- p controls the likelihood of immediately revisiting a node in the walk.
  - A **low** value of p would keep the walk local, close to the starting node, u. A low value of p would encourage local walks next to a source node, u.

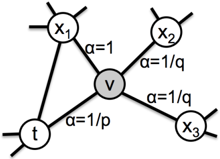


Figure: from Grover and Leskovec, KDD. 2016

Intuition Behind q

item q allows for differentiation between inward and outward nodes

- For $q < 1$, the walk is inclined to visit nodes that are further from node t. (Depth First)

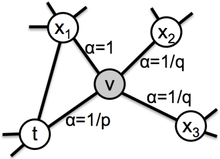
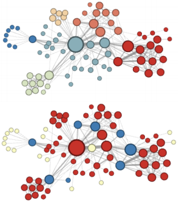


Figure: from Grover and Leskovec, KDD. 2016

Allow for Variation in What Defines Nodes as Being Similar

- **Application of Capturing Local Similarities:** Identify proteins involved in a common pathway. Or parts of the brain that are correlated to each other.
- **Usefulness of Capturing More Global Similarities:** Identify proteins that are regulators of many other proteins but are not related extensively among them selves.
- Validating a graph representation (**project idea!!!**).



Embedding as a Tool for a Variety of Tasks on Graphs

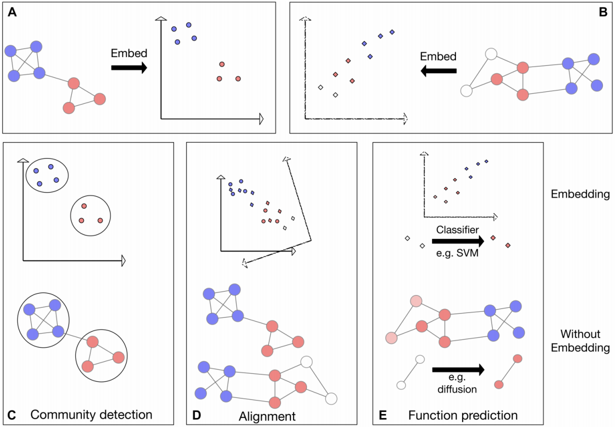


Figure: from Nelson *et al.* Frontiers in Genetics. 2019. Protein families (C). Cross-Species Alignment (D), Function Prediction (E)

Recap of graph preliminaries and opportunities in biomedicine

- Graphs are useful for representing relational information
  - Cells, Patients, etc.
- There are a lot of well-defined tasks on graphs that we can use to answer our questions
  - Graph partitioning
  - Diffusion for smoothing (which we will see later), and partitioning if we were to work with Graph Laplacian
  - Visualization- connecting visualization with the scientific problem
  - Other tasks not discussed: link prediction, collaborative filtering, label propagation.

Single Cell Intro Day

- What are single-cell assays?
- Mass Cytometry (CyTOF) Bioinformatics
- Automating Human Gating with Spade
    - What is a gate?
- Practical Considerations (normalization, batch effects)

What is a Single-Cell Assay?

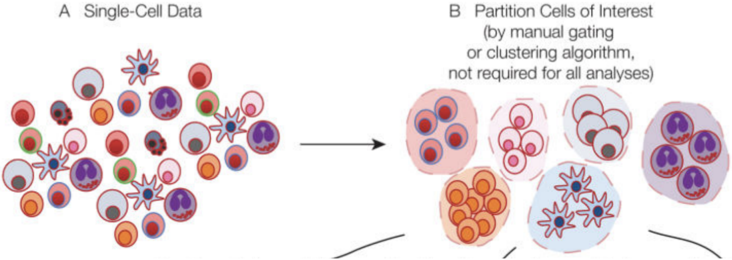


Figure: from Spitzer *et al.* Cell. 2016. In general, we are taking a biological sample (blood, tissue, etc.) and measuring properties of individual cells. We then seek to track and understand entire cell-populations.

What Level of Biology do You Want to Measure?

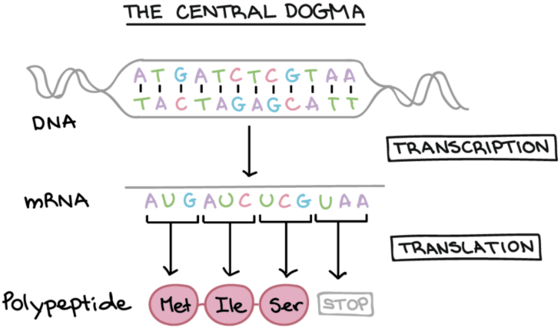


Figure: Remember, DNA → RNA → protein

scRNA seq for measuring gene expression

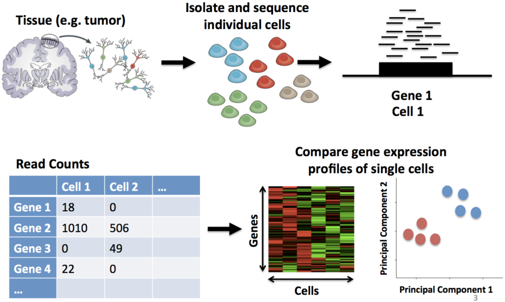


Figure: from `https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/`.
Measure the expression of 10s of thousands of genes in 10s of thousands of cells.

Starting first with single-cell proteomics

Here we will discuss analysis of single-cell proteomics assays because they are useful for immunophenotyping, or identifying, tracking, and characterizing the diverse immune cell-types.

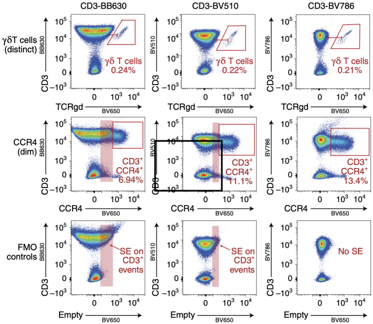


Figure: from Nature Immunology. 2021. We use protein expression in cells to figure out what they are.

Example use case of immune profiling clinically

Example 1: The immune system adapts during pregnancy. We can use it to predict someone's likelihood of pre-mature birth.

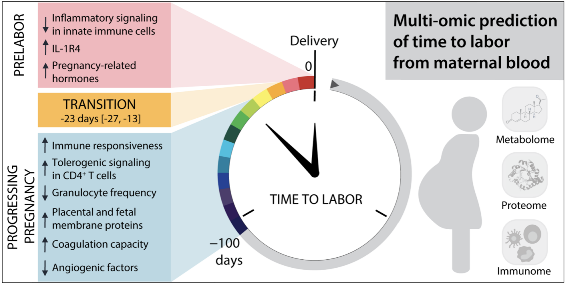


Figure: from Science Translational Medicine. 2021.

SC Protemics: Flow and Mass Cytometry

Flow → 18-20 proteins per cell at a rate of 10,000 cells per second!

Mass → 36-45 proteins at a rate of 1,000 cells per second

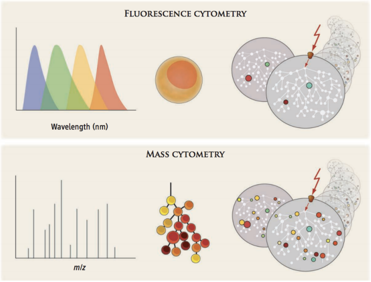


Figure: from Benoist and and Hacohen. Science. 2011

Cytometry by Time of Flight (CyTOF)

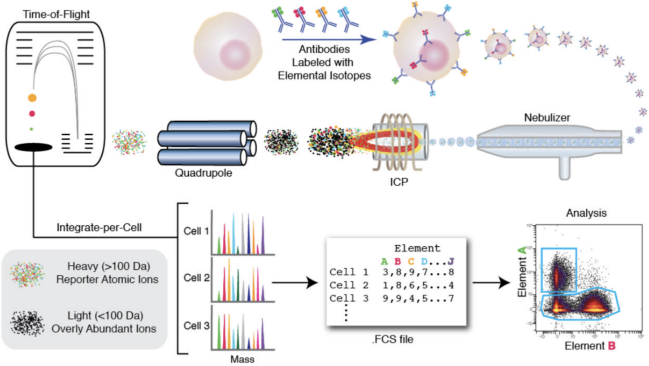


Figure: from Bendall *et al.* Trends in Immunology. 2012

CyTOF : the specific technology for mass cytometry

Manual Gating

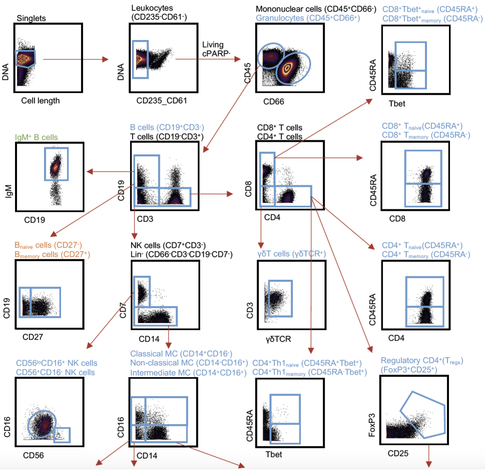


Figure: Nature Communications. 2020

Problems with Manual Gating

- Requires a human expert with extensive knowledge about how to characterize cells
- Time consuming. If you measure 36 proteins, then you have 36 choose two scatterplots. Not all combinations of markers is meaningful and the gating is hierarchical in nature.
- Biased towards characterizing cell-populations that have already been well described (e.g. you only find what you are looking for).
- A coming attraction : fully automating this analysis!

I have my manually gated cell-populations, now what?

We look at particular immune cell-types to understand how the immune system is responding in particular circumstances (for example, pre-term birth).

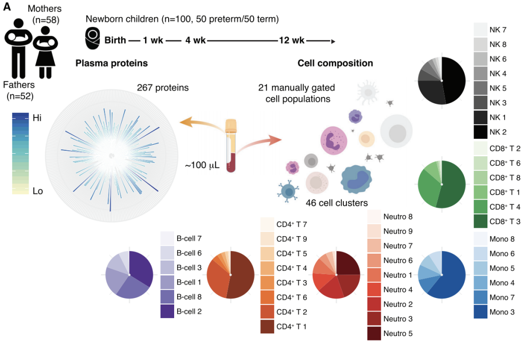


Figure: from Olin *et al.* Cell. 2018

What Properties of Cell-Populations do we Study?

You can count cells in each population (known as frequency, or fq), or characterize signaling activity across multiple stimulations (e.g. experimental perturbations).

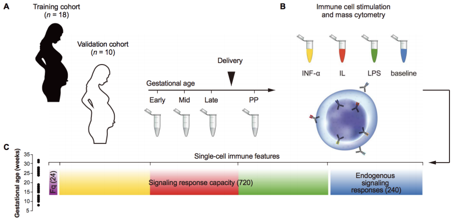


Figure: from Aghaeepour *et al.* Science Immunology. 2017

Extracting Features for Prediction Tasks

- Start with a matrix of cells × protein expression for each patient sample
- Define populations through gating
- Compute function or frequency based features for each population

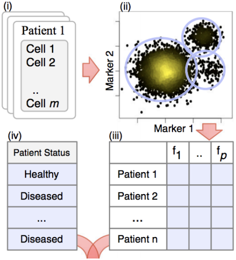


Figure: from Bruggner *et al.* PNAS. 2014

What Are Cell Frequencies?

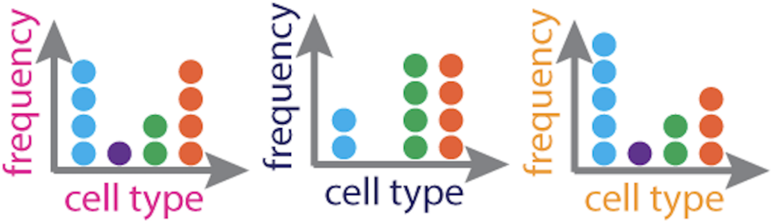


Figure: Simply count the number of cells in each profiled sample assigned to each specific type. This is a very simple hand engineered feature.

Practical Concerns. Converting data profiled with CyTOF into a matrix that can be used to understand a biological question.

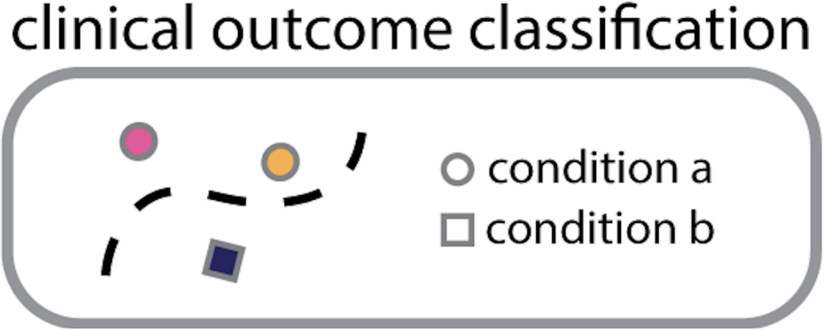


Figure: Each point is representing a different patient sample. How do we ultimately specify a set of features for each sample that can be used to classify them correctly?