Comp790-166: Computational Biology

Lecture 14

March 1, 2023

Today

- Trajectory inference, with SLICER
- Spatial immunophenotyping
- LEAPH for identifying cellular microenvironments.

Project Related Announcements

- Please sign up with your group members and topics here, `https://docs.google.com/document/d/1x9mIJCZAkeogAhmGlpqJkXwuoAKlB_0gewV1LpXZDoU/edit?usp=sharing`
- Writeup due before spring break on March 8.
- Presentations will be the week after spring break. Stay tuned for your date and time.

Do You Remember Question

1. What is a main limitation of volume-dependent sketching, like hopper or geometric sketching and what can it tend to over-select for?
2. What are the main components required for computing diffusion distance?

Review $\rightarrow$ Using Eigenvectors of $\tilde{\mathbf{P}}$

- $\tilde{\mathbf{P}}$ has n ordered eigenvalues, $\lambda_0 = 1 > \lambda_1 \geq \cdots \geq \lambda_{n-1}$
- Corresponding to these eigenvalues are eigenvectors $\mathbf{\Psi}_0 \ldots \mathbf{\Psi}_{n-1}$
- As we saw a few weeks ago, powering $\tilde{\mathbf{P}}$ to $\tilde{\mathbf{P}}^t$ represents the probability of transitioning between two cells with a walk of length t.

The diffusion distance between a pair of cells $\mathbf{x}$ and $\mathbf{y}$ can be written in terms of the eigenvectors of $\tilde{\mathbf{P}}$ as,

$$D_t^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n-1} \lambda_i^{2t} \left(\psi_i(\boldsymbol{x}) - \psi_i(\boldsymbol{y})\right)^2$$

Unpacking Diffusion Distance

$$D_t^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n-1} \lambda_i^{2t} \left(\boldsymbol{\psi}_i(\boldsymbol{x}) - \boldsymbol{\psi}_i(\boldsymbol{y})\right)^2$$

- The eigenvector to the largest eigenvalue, λ_0 is a constant vector, $\boldsymbol{\Psi}_0 = \mathbf{1}$. Therefore, it contributes 0.
- The eigenvalues of $\tilde{\mathbf{P}}$ determine the diffusion coefficients in the direction of the corresponding eigenvector
- After the first l prominent directions, the diffusion coefficients typically drop to a noise level.
- When you find the l such that there is a large difference between l and $l+1$ eigenvalues (an elbow), you can use the sum up to the l-th term as an approximation for diffusion distance. The first l eigenvectors correspond to the diffusion components.

The diffusion map is capturing transitions between cell types that are not reflected in PCA or tSNE.

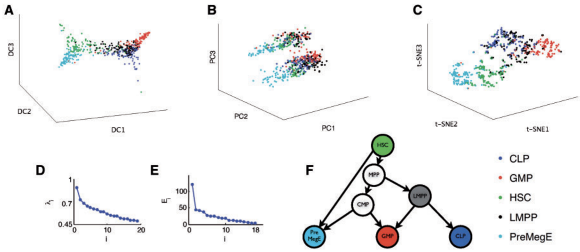


Figure: from Haghverdi *et al.* Bioinformatics. 2015.

This diffusion map approach was the beginning of thousands of people starting to think about cellular differentiation.....

Welcome SLICER

SLICER builds on and expands the very early Diffusion based techniques through the following

- Automatically select genes to use for building the trajectory (or in establishing the ordering between cells)
- Use locally linear embedding to capture non-linear relationships between gene expression levels and progression through a process
- Define ‘geodesic entropy‘ and use it to define branches
- Capture unique trajectory patterns such as bubbles.

SLICER Overview

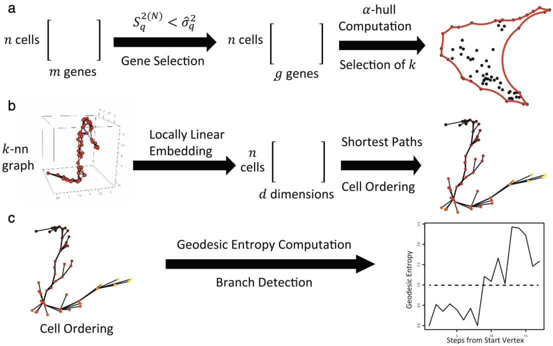


Figure: from Welch *et al.* Genome Biology. 2016

Step 1: Selecting Features to Use (Intuition)

Establishing some intuition about what makes a good ‘trajectory feature‘

- If a feature is involved in progression along a trajectory, expect gradual change in that feature along the trajectory
- A feature not involved should not fluctuate along the trajectory.
- In real life, we have no idea what is happening with this trajectory. Use similarity within neighborhoods to study ‘segments‘ of a trajectory.

Neighborhood Variance

Interesting features are those whose variance is greater than some level of neighborhood variance. Specifically, for the gth feature, we can compute its variance within a neighborhood and compare it to the overall variance. The neighborhood variance is defined as,

$$S_g^{2(N)} = \frac{1}{nk_c - 1} \sum_{i=1}^{n} \sum_{j=1}^{k_c} \left(e_{ig} - e_{N(i,j)g}\right)^2$$

- k_c is the number of nearest neighbors needed for each node for the graph to be connected.
- Each e_{ig} is representing the value of feature g in cell i.
- $e_{N(i,j)g}$ is representing the feature value of the jth nearest neighbor in cell i.

Selecting Genes Most Likely to Be Involved in a Trajectory

Given computed neighborhood variance, $\hat{\sigma}_g$ and neighborhood variance $S_g^{2(N)}$, we seek genes varying more globally than within a neighborhood so that $\hat{\sigma}_g > S_g^{2(N)}$.

Local Linear Embedding ($d = 2$)

Step 1: Find the weights (w_{ij}s) that can best reconstruct the original data (e.g. the Es cell $\times$ feature) in terms of k nearest neighbors as,

$$W = \text{argmin}_W \sum_{i=1}^{n} \left| E_i - \sum_{j=1}^{k} w_{ij} E_j \right|_2^2$$

Step 2: Find optimal d-dimensional embedding, so in this case, L

$$L = \text{argmin}_L \sum_{i=1}^{n} \left| L_i - \sum_{j=1}^{k} w_{ij} L_j \right|_2^2$$

k-NN graph and shortest path

- Compute k-nearest neighbor graph between cells in terms of the LLE-determined coordinates.
- Specify a starting point (like a stem cell), and use a shortest path algorithm like Dijkstra to find the shortest path to some cell of interest.

Detecting Branches with Geodesic Entropy Measure

- Let $t_i = \{s = v_1, \ldots, v_k, \ldots, v_l = i\}$ be the shortest path from the starting point s to some cell, i.
- Denote the kth node on the shortest path from s to i by $t_i(k)$.
- Define f_{jk} as the number of paths passing through point j at distance k, $f_{jk} = \sum_i^n I[t_i(k) = j]$
- Then compute the fraction of all paths in S that pass though node j at distance k, $p_{jk} = \frac{f_{jk}}{\sum_{i=1}^n f_{ik}}$
- $H_k = -\sum_{i=1}^n p_{ik} \log_2 p_{ik} \rightarrow$ look at high entropy

SLICER Applied to Synthetic Data

Studying geodesic entropy over k. Higher entropy in terms of steps corresponds to the 'bubbles' in the data.

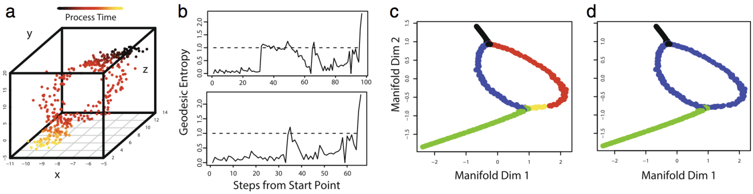


Figure: from Welch *et al.* Genome Biology. 2016.

Neural Stem-Cell Differentiation Data

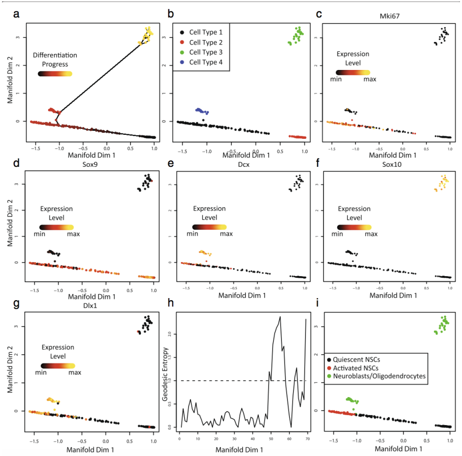
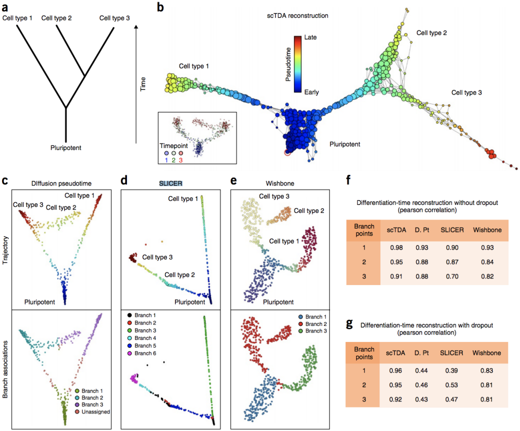


Figure: from Welch *et al.* Genome Biology. 2016.

SLICER Compared



f Differentiation-time reconstruction without dropout (pearson correlation)

Branch points	scTDA	D. Pt	SLICER	Wishbone
1	0.98	0.93	0.90	0.93
2	0.95	0.88	0.87	0.84
3	0.91	0.88	0.70	0.82

g Differentiation-time reconstruction with dropout (pearson correlation)

Branch points	scTDA	D. Pt	SLICER	Wishbone
1	0.96	0.44	0.39	0.83
2	0.95	0.46	0.53	0.81
3	0.92	0.43	0.47	0.81

Figure: from Rizvi *et al.* Nature Biotechnology. 2016.

Switching gears to spatial immunophenotyping with imaging cytometry modalities.....

CyTOF + Spatial Resolution

An upgrade of regular CyTOF to image 32 proteins and their modifications at cellular resolution.

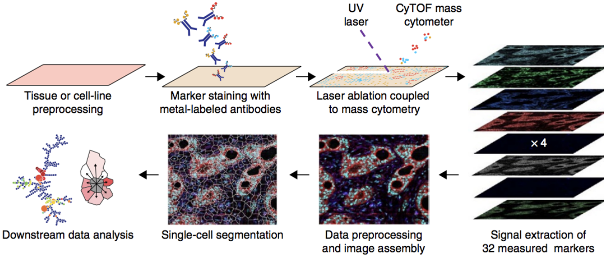


Figure: from Giesen *et al.* Nature Methods. 2016

Understanding the spatial organization of cells (for example, tumor and immune cells) can provide a more mechanistic understanding of the underlying biology. This can further translate to more accurate prediction of prognostic outcomes.

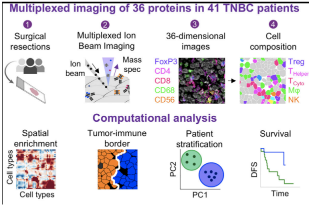


Figure: from Keren *et al.* Cell. 2018.

Recent Advances in Study The Relationship Between Immune Cells and Tumor

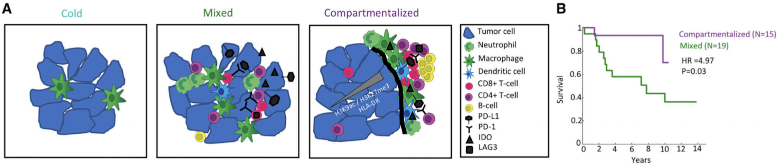


Figure: from Keren *et al.* Cell. 2018.

Studying Aging

Older mice were observed to have infiltrating T-cells in their neurogenic niches (the collection of neuronal progenitor cells)

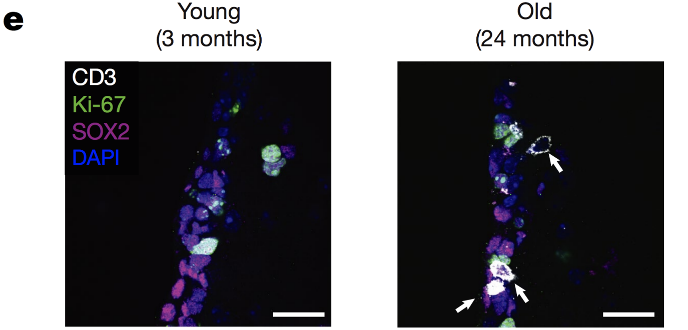
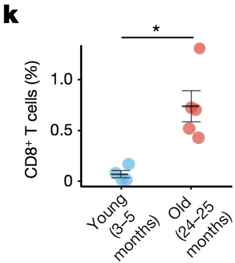


Figure: from Dulken *et al.* Nature 2019

Counting CD8+ T-cells

You can even compare the proportion of CD8+ T-cells there are in neurogenic niches between young and old mice. It's a pretty striking difference.



General Steps in Analyzing These Data

- Segmentation of cells
- Phenotype cells
- Identify microenvironments or characteristic co-occurences of particular cell-types within a region.

Example-Cell Phenotype Map

Cells are clustered and phenotyped according to protein expression.

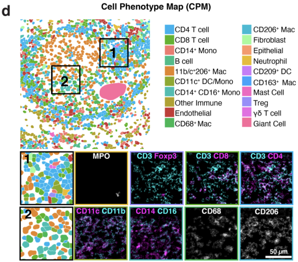


Figure: from https://www.biorxiv.org/content/10.1101/2020.06.08.140426v1.full.pdf

End-Goal of Identifying Particular Microenvironments

Ultimately, an objective is to identify ‘micro-environments‘ or spatially-localized subsets of cells with characteristic frequency patterns that are predictive of some outcome of interest.

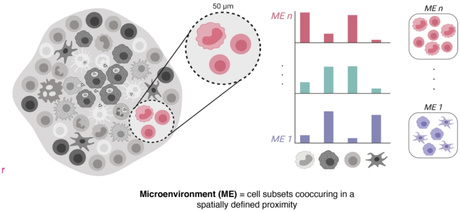


Figure: from `https://www.biorxiv.org/content/10.1101/2020.06.08.140426v1.full.pdf`

A New Problem: Identifying Microenvironments

Welcome LEAPH. One of the first methods out there to identify phenotypically distinct microdomains of spatially configured cell phenotypes.

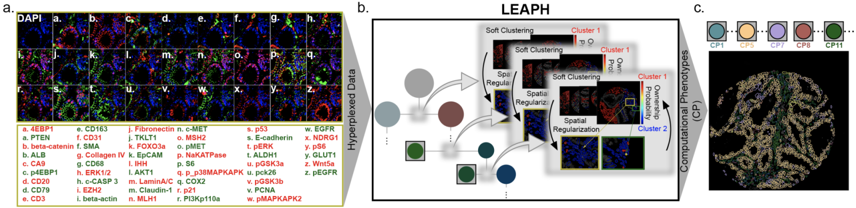


Figure: from Furman *et al.* Cell Systems. 2021.

LEAPH Overview

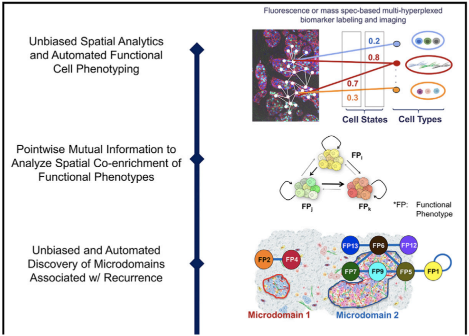


Figure: from Furman *et al.* Cell Systems. 2021.

Notation in LEAPH

- For cell i, let its protein expression be represented as $\mathbf{x}_i \in \mathbb{R}^p$.
- Mixture of factors setup, with k dimensions in the latent space, with $\mathbf{x}_i = \Lambda\mathbf{z} + \boldsymbol{\mu} + \mathbf{v}$
  - Loadings in $\Lambda \in \mathbb{R}^{p \times k}$
  - Latent variables, $\mathbf{z} \in \mathbb{R}^{k \times 1}$
  - Noise term via, $\mathbf{v} \sim \mathcal{N}(0, \Psi)$
  - Mean vector, $\boldsymbol{\mu} \in \mathbb{R}^{p \times 1}$

Mixture Model

Each $p(\mathbf{x}_i)$ is computed as

$$p(\mathbf{x}_i) = \sum_{j=1}^{M} \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \Lambda_j \Lambda_j^T + \Psi)$$

- π_j is the mixing weight for cluster j.

Practicalities

- Overall, parameters being estimated are $\{\pi_j, \boldsymbol{\mu}_j, \Lambda_j\}_{j=1}^{M}, \Psi)$.
- They 2-dimensions for each latent space, so, $k = 2$.
- Ultimately, they get a prediction that each cell belongs to of the M components, and in particular for class j, $p(j \mid \mathbf{x}_i) = \frac{p(\mathbf{x}_i|j)p(j)}{\sum_{c=1}^{M} p(\mathbf{x}_i|c)p(c)}$
- Use the estimated probability between a cell i and a cluster c and create a matrix, $\Omega \in \mathbb{R}^{N \times M}$ where Ω_{ic} gives the probability that cell i belongs to cluster c.
- This gives a soft clustering interpretation for each cell.

Spatial Regularization Intuition

- Based on prior biological knowledge, there are known properties that for example, epithelial/tumor cells should be surrounded by or spatially proximal to other epithelial/tumor cells.
- There should also be some allowance for tumor-infiltrating cells, such as lymphocytes and other stromal cells.

A new Ω is optimized that encodes spatial information as follows,

$$\min_{\Omega} - \sum_{i=1}^{N} \sum_{j=1}^{M} \Omega_{ij} \log_2 (\Omega_{ij}) + \lambda \sum_{(j,k)} w_{jk} ||\Omega_j - \Omega_k||_2$$

Unpacking

$$\min_{\Omega} -\sum_{i=1}^{N}\sum_{j=1}^{M}\Omega_{ij}\log_2(\Omega_{ij}) + \lambda\sum_{(j,k)} w_{jk}||\Omega_j - \Omega_k||_2$$

- w_{jk} is a weight, calculate as the reciprocal of distance between cells j and k in the image
- The first term is basically an entropy term of ownership confidence
- The second term is promoting spatial coherence.
- λ controls the tradeoff between spatial coherence and membership confidence.

Effect of Spatial Regularization

In particular in the first example, a cell with a highly predicted assignment towards CP1 transitioned towards a phenotype of CP2 after spatial regularization.

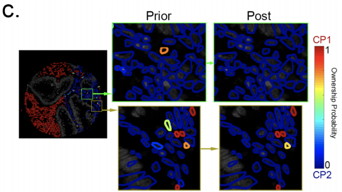


Figure: from Fig. 2 of `https://www.biorxiv.org/content/10.1101/2020.10.02.322529v3.full.pdf`

Determining Specialized Cells

- Based on the Ω, assign each cell to one of the M phenotypes based on the j that gives the maximum probability.
- For a particular patient, p, create a feature vector $\mathbf{f}_p$ which gives the proportion of its cells assigned to each of the cell phenotypes.
- At times, the authors refer to specialized cell-types (membership probability $> 95\%$) in contrast to transitional and rare cells.

Recap and Transition

- The clustering part is straight-forward : Assume each cell is from one of M 2-dimensional latent factors
- Calculate a probability that each cell was from each of these latent factors
- Add penalties that enforce spatial coherence and certainty of assignment
- **Next step:** Identify microdomains with a collection of cells that are predictive of some penotype of interest.

Predicting Time to Recurrence in Breast Cancer

- Consider cohorts of patients with the following properties.
    - 45 patients in 'NED-8' category that have no evidence of disease for over 8 years
    - 46 patients in 'NED-3', where cancer came back within 3 years.

The goal is to translate the distributions of cell phenotypes that spatially co-occur to a signal that can be used for prediction.

Constructing a Cell Network For Each Patient

- Connectivity is determined by proximity in the image of the tissue
- For a pair of cells, m, and n, connect them with a weights, $w_{mn} = 1$ if their spatial distance, $d_{mn} < 1$.
- Otherwise, $w_{mn} = 0$ and there are no edge between the cells

Identifying Spatial Co-Occurrence Between Cell Phenotype Pairs

Consider two phenotypes, f_i and f_j for a given set (e.g. a subset of patients, etc). The pairwise mutual information between these two phenotypes is defined as,

$$\mathrm{PMI}_s\left(f_i, f_j\right) = \log_2 \left(\frac{p\left(f_i^s, f_j^s\right)}{p\left(f_i^t\right) p\left(f_j^t\right)} \right)$$

- $p(f_i^s)$ is the probability of a particular phenotype, i occurring in a network set, s.
- $p(f_i^t)$ is the background probability of phenotype i.

Calculating Joint Phenotypic Probability for a Single Patient

Letting Ψ encode the set of edges for a particular patient, the joint probability of phenotypes i and j is given as,

$$p\left(f_i^s, f_j^s\right) = \frac{1}{z}\left(\sum_{(m,n)\in\Psi} w_{mn}\left(\vec{\Omega}_{mf_i}\vec{\Omega}_{nf_j} + \vec{\Omega}_{mf_j}\vec{\Omega}_{nf_i}\right)\right)$$

*Here z is a normalization over all combinations of i and j according to the computational phenotypes.

Specifying a Background Distribution

The background probability for a phenotype, i is simply the mean assignment probability over all cells, or,

$$p(f_i^t) = \frac{1}{N} \sum_{c=1}^{N} \Omega_{ci}$$

Ultimately, for each cell phenotype pair, (f_i, f_j) compute the PMI for each sample and consider how this relates to the patient re-occurence outcomes.

Looking at Significant Microdomains Between Groups

There were a few cellular phenotypes that tended to co-occur between the two patient groups.

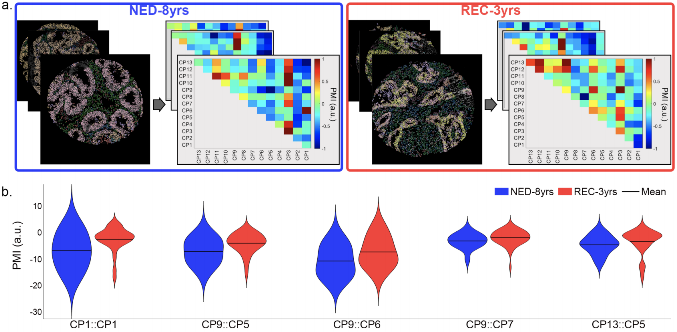


Figure: from Fig. 4 in `https://www.biorxiv.org/content/10.1101/2020.10.02.322529v3.full.pdf`