Comp790-166: Computational Biology

Lecture 20

March 27, 2022

Good Morning Question

1. What are examples of the multiple single-cell datasets that could be merged with Conos?
2. What is some intuition about how within and between dataset edges are determined?

Today

- SLICER and trajectories
- Begin multimodal integration in biology

Intermission for Announcements

- Next homework to be assigned $\sim$ April 6. Now is a good time to work on projects!
- Don't forget about reading summaries

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

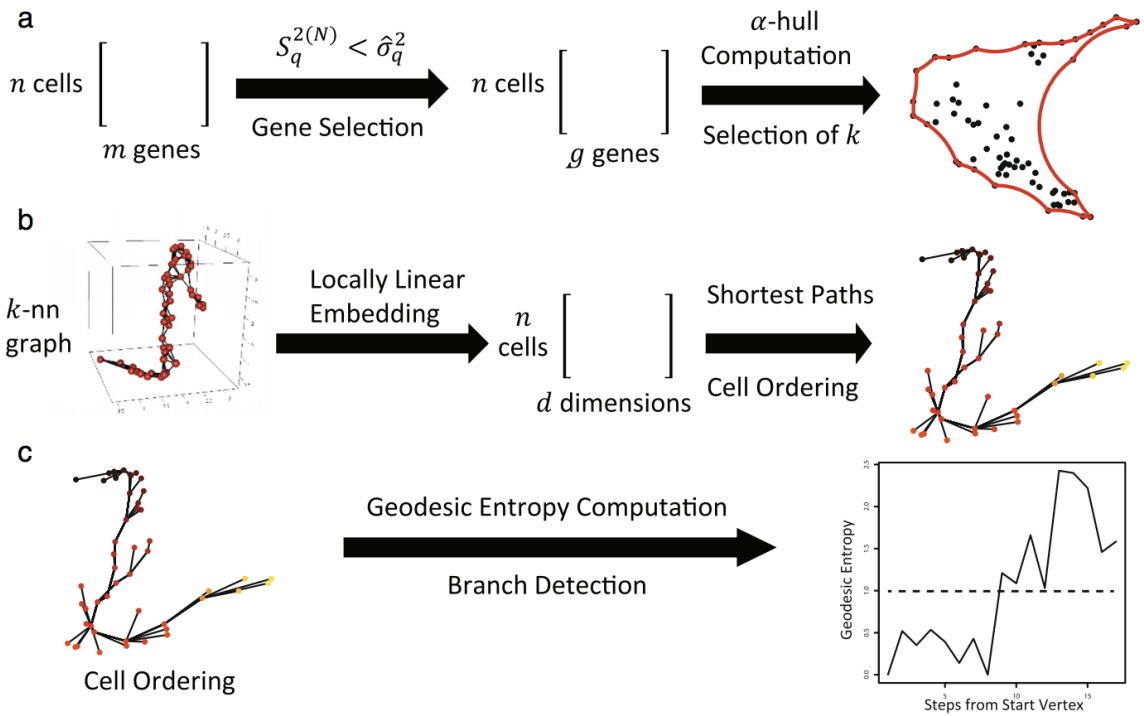


Figure: from Welch *et al.* Genome Biology. 2016

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 (σ_g^2) across samples and compare it (making sure it is at least as large as) to this defined neighborhood 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.

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.

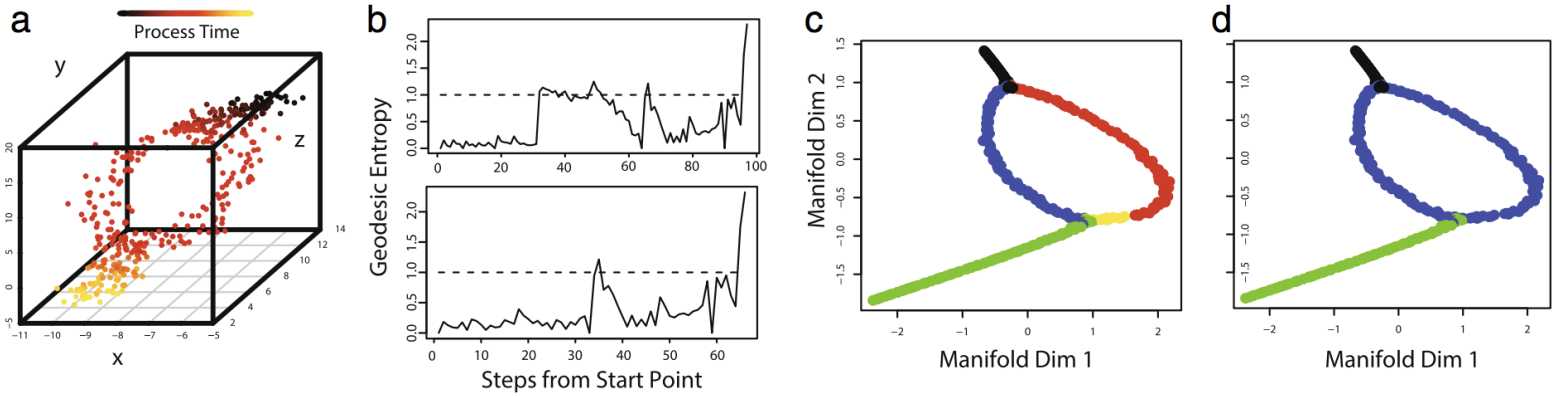


Figure: from Welch *et al.* Genome Biology. 2016.

Neural Stem-Cell Differentiation Data

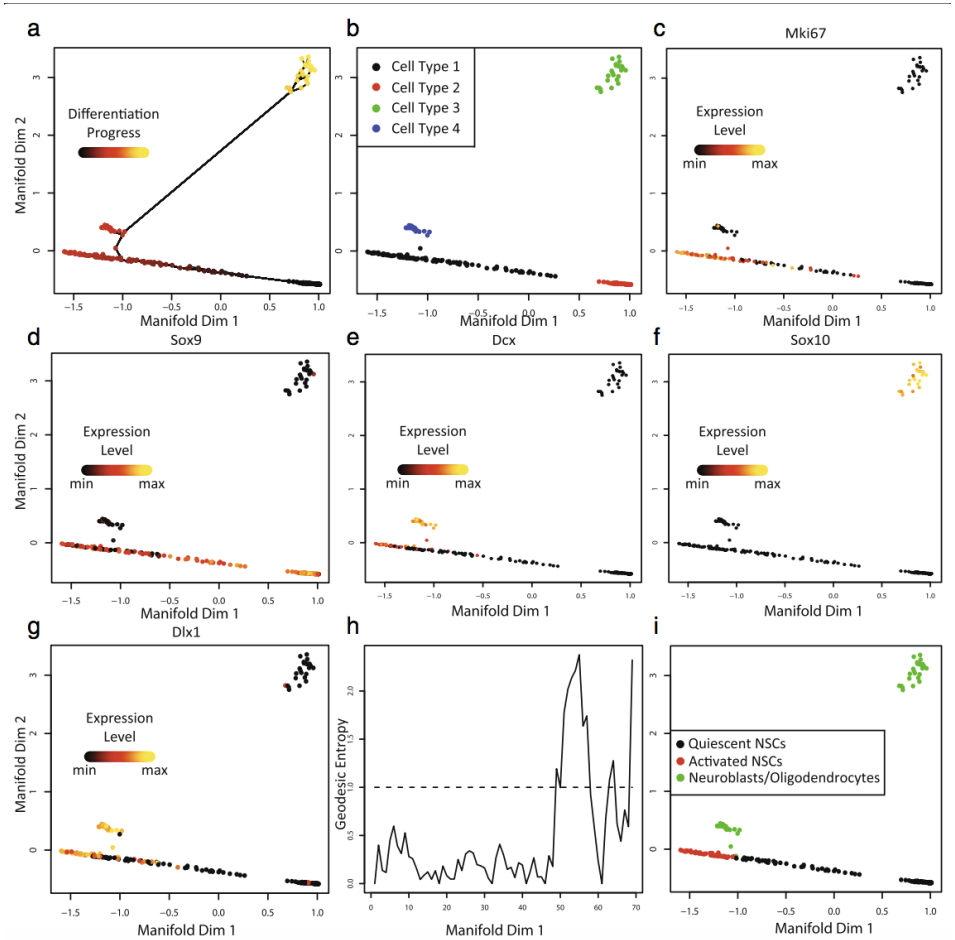
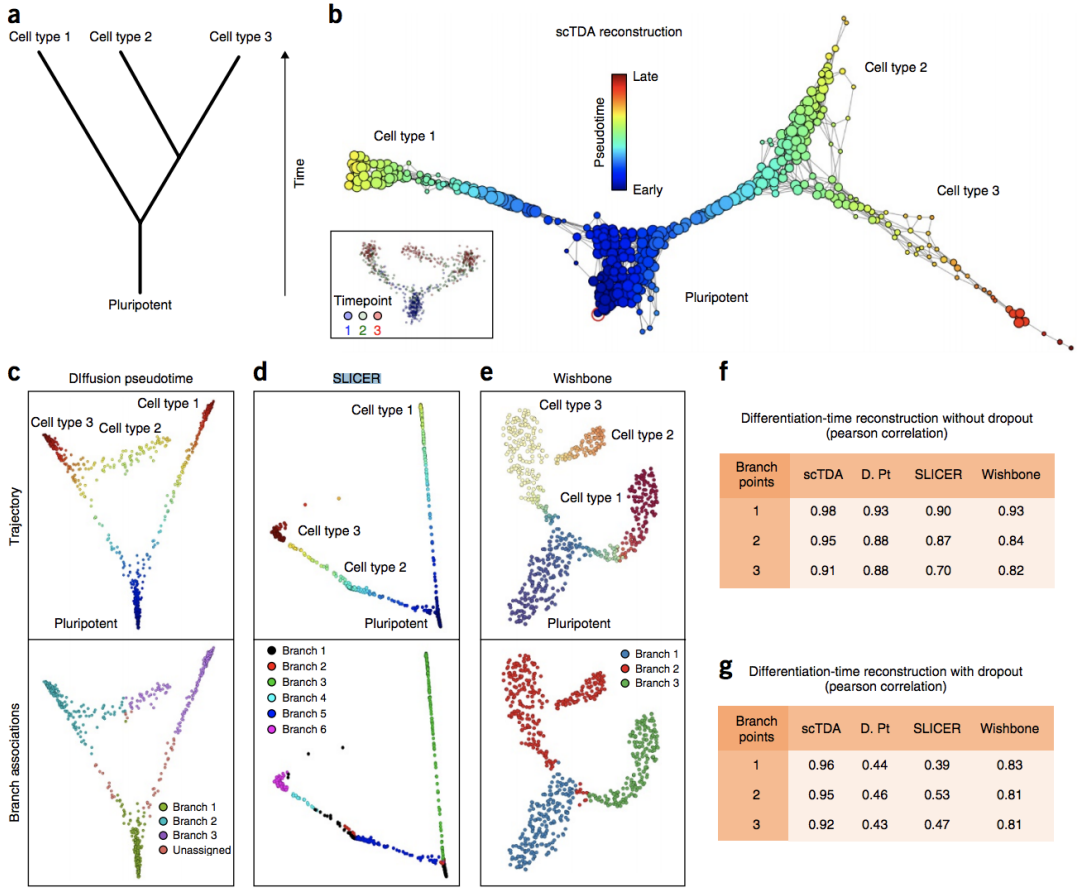


Figure: from Welch *et al.* Genome Biology. 2016.

SLICER Compared



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

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.

Where we are going now...

The Overall Problem: Combining Multiple Sets of Features

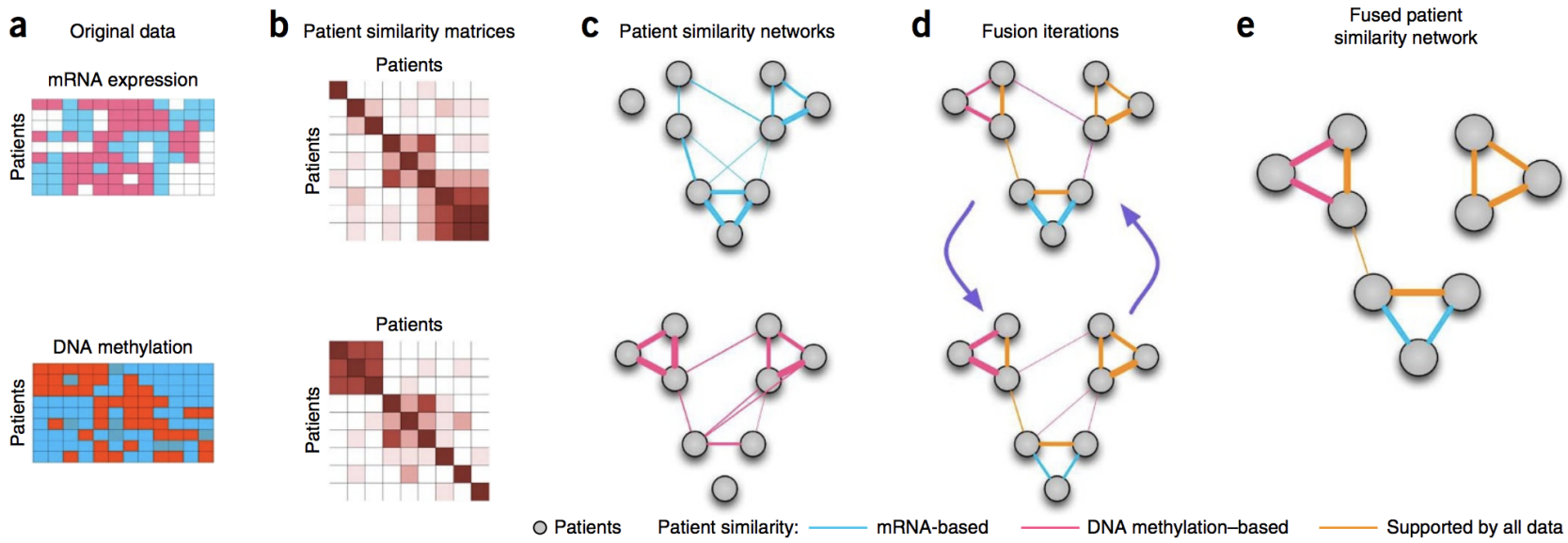


Figure: from Wang *et al.* Nature Methods 2014. The problem is to learn a joint representation of all patients that respects each modality.

The Cancer Genome Atlas (TCGA)

The focus on merging multiple datasets was inspired by The Cancer Genome Atlas, an effort to profile large patient cohorts of patients with various cancer types, with several modalities.

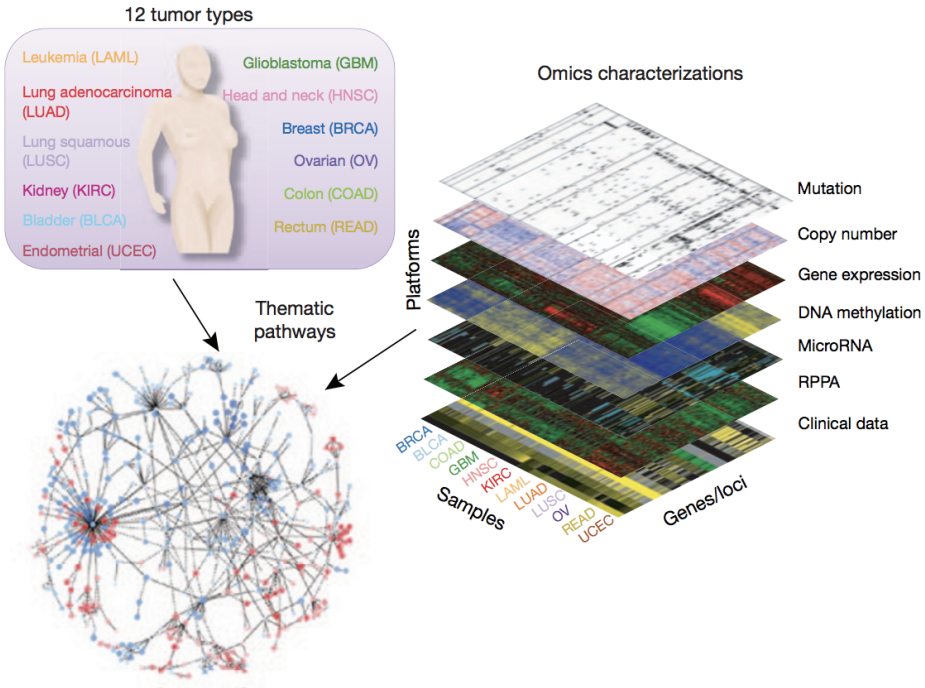


Figure: from TCGA, Nature Genetics. 2013.

LinkedOmics for Human Readable Data

- Download TCGA data here across many different cancers
- `http://www.linkedomics.org/login.php`

LinkedOmics "OMICS" Datatype

- **Clinical Data :** It inculdes attributes like age, overall survival, phological stage (I, II, III, IV), TNM staging, Clinical subtype, Molecular Subtype, number of lymph nodes, radiation therapy.
- **Copy Number (Level: Focal, Gene) :** Normalized copy number (SNPs) ans Copy number alterations for aggregated/segmented regions, per sample
- **miRNA (Leve: Gene, Isoform) :** Normalized signals per probe or probe set for each participant's tumor sample
- **Mutation (Level: Site, Gene) :** Mutation calls for each participant
- **Methylation (Level: Gene) :** Calculated beta values mapped to genome, per sample
- **RNAseq (Level: Gene) :** The normalized expression signal of individual Gene (transcripts), per sample
- **RPPA (Level: Analyte, Gene) :** Normalized protein expression for each gene, per sample
- **Proteomics (Level: Gene) :** Average log-ratio of sample reporter-ion to common reference of peptide ions associated with the gene in acquisitions from a specific biological Sample (Unshared Log Ratio-Average log-ratio of sample reporter-ion to common reference of peptide ions of unshared peptides only associated with the gene in acquisitions from a specific biological sample).
- **Phospho-Proteomics (Level: Site) :** Average log-ratio of sample reporter-ion to common reference of peptide ions associated with phosphorylated site combinations in acquisitions from a specific biological sample (CDAP Protein Report).
- **Glyco-Proteomics (Level: Site) :** Average log-ratio of sample reporter-ion to common reference of peptide ions associated with deglycosylated N-glycosylation site combinations in acquisitions from a specific biological sample (CDAP Protein Report).

For more information (Click here)

LinkedOmics Data Source

Cancer Type	Cohort Source	Cancer ID	Samples	Death Events	Median OS (yrs)	Permission	Link	Data Download
Adrenocortical carcinoma	TCGA	ACC	92	33	NA	Y	TCGA, GDAC	Download
Bladder urothelial carcinoma	TCGA	BLCA	412	178	2.84	Y	TCGA, GDAC	Download
Breast invasive carcinoma	TCGA	BRCA	1097	151	10.81	Y	TCGA, GDAC, CPTAC	Download
Cervical and endocervical cancers	TCGA	CESC	307	71	8.48	Y	TCGA, GDAC	Download

Notation and Problem Formulation

- Consider M types of omics data measurements $\{\mathbf{X}^m\}_{m=1}^M$ from the same set of N patients.
- For a modality, m, there are p_m measured features and the dimensions of the data matrix are therefore $p_m \times N$
- We will let G^m be the graph for modality m

Comment

Before we had node2vec, we just used nice theorems from linear algebra! :D (graph embedding for old people)

Overview of Subspace Merging

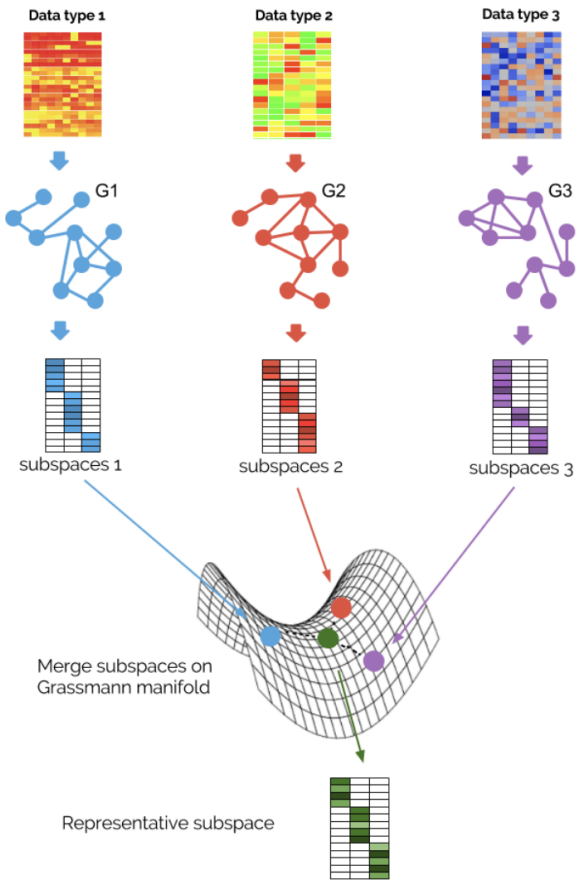


Figure: from Ding *et al.* Bioinformatics. 2019.

Build a Similarity Graph Between Patients in Each Modality

Use our 'favorite' rule for calculating edge weights as,

$$S_{ij}^{m} = \exp\left(-\frac{\left\|\mathbf{x}_i^m - \mathbf{x}_j^m\right\|^2}{2t^2}\right), i = 1, \ldots, N, j = 1, \ldots N$$

From here, retain the top k edges for each node based on S_{ij} and use W_{ij} for the notation of the edge weights retained, such that, $W_{ij}^m = S_{ij}^m$

Connection to Some GSP Conversation from a Few Weeks Ago

We already talked about the total variation of a signal in terms of the Graph Laplacian, or the variation of a signal around neighbors as,

$$\mathbf{x}^T \mathbf{L} \mathbf{x} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} (x_i - x_j)^2 \tag{1}$$

Pause for Rayleigh Ritz Theorem

Let $\mathbf{A}$ be a square, symmetric matrix, $N \times N$ matrix with eigenvalues, $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_n$ and corresponding eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_n\}$. Then define

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \tag{2}$$

Then the minimum value of $R_{\mathbf{A}}(\mathbf{x})$ is λ_1 and it's taken for $\mathbf{x} = \mathbf{v}_1$

Matrix Extension

We will be seeing a lot on the form of $\mathbf{X}^T\mathbf{L}\mathbf{X}$. We can talk about the trace of that matrix product as the distance in vectors of adjacent nodes.

$$\text{trace}(\mathbf{X}^T\mathbf{L}\mathbf{X}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}||\mathbf{x}_i - \mathbf{x}_j|| \tag{3}$$

An extension of Rayleigh Ritz says that the minimum k-dimension matrix $\mathbf{X}$ of trace$(\mathbf{X}^T\mathbf{L}\mathbf{X})$ is $\lambda_1 + \lambda_2. + \cdots + \lambda_k$ and corresponds to the first k eigenvectors of $\mathbf{L}$.

Specify Optimization Problem in terms of Normalized Graph Laplacian

$$\mathbf{L}^m = \mathbf{D}^{m^{-\frac{1}{2}}} \left(\mathbf{D}^m - \mathbf{W}^m\right) \mathbf{D}^{m^{-\frac{1}{2}}}$$

Written out this gives us,

$$L_{i,j}^{\text{sym}} := \begin{cases} 1 & \text{if } i = j \text{ and } \deg(v_i) \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i)\deg(v_j)}} & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise.} \end{cases}$$

Writing Down the Objective Function

The goal is to specify a $\mathbf{U}^m$ for each modality. The optimal graph embedding in k dimensions can written as,

$$\min_{\mathbf{U}^m \in \mathbb{R}^{N \times k}} \operatorname{tr}\left(\mathbf{U}^{m\prime} \mathbf{L}^m \mathbf{U}^m\right), \quad \text{s.t. } \mathbf{U}^{m\prime} \mathbf{U}^m = I$$

- It turns out the solution is the first k eigenvectors of the Graph Laplacian $\mathbf{L}^m$ by the Rayleigh–Ritz theorem

Merging Subspaces on a Grassmann Manifold

- With the subspace representations $\mathbf{U}_{m=1}^{M}$ from each data type, these will be merged on a Grassmann manifold
- A Grassmann manifold is defined as a set of linear subspaces of a euclidean space.
- To merge all $\mathbf{U}^m$, we seek to define an integrative subspace, span($\mathbf{U}^m$) that should also preserve connectivity in each G^m.

Defining a Projection Distance Between The Integrative Subspace and Individual Modality Subspaces

$$
\begin{aligned}
d^2_{\text{proj}}\ \left(\mathbf{U}, \{\mathbf{U}^m\}_{m=1}^M\right) &= \sum_{m=1}^{M} d^2_{\text{proj}}\ \left(\mathbf{U}, \mathbf{U}^m\right) \\
&= \sum_{m=1}^{M} \left[k - \text{tr}\left(\mathbf{U}\mathbf{U}'\mathbf{U}^m\mathbf{U}^{m\prime}\right)\right] \\
&= kM - \sum_{i=1}^{M} \text{tr}\left(\mathbf{U}\mathbf{U}'\mathbf{U}^m\mathbf{U}^{m\prime}\right)
\end{aligned}
$$

The subspace, $\mathbf{U}$ that minimizes this is close to all individual subspaces, $\{\mathbf{U}^m\}_{i=1}^M$

Optimization Problem for Multiple Subspaces

The optimization problem for merging multiple subspaces finally can be written as,

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times k}} \sum_{m=1}^{M} \text{tr}\left(\mathbf{U}'\mathbf{L}^m\mathbf{U}\right) + \alpha\left[kM - \sum_{m=1}^{M} \text{tr}\left(\mathbf{U}\mathbf{U}'\mathbf{U}^m\mathbf{U}^{m\prime}\right)\right], \quad \text{s.t. } \mathbf{U}'\mathbf{U} = I$$

The authors showed that this simplifies to,

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times k}} \text{tr}\left[\mathbf{U}'\left(\sum_{i=1}^{M}\mathbf{L}^m - \alpha\sum_{m=1}^{M}\mathbf{U}^m\mathbf{U}^{m\prime}\right)\mathbf{U}\right], \quad \text{s.t. } \mathbf{U}'\mathbf{U} = I$$

Rayleigh Ritz Again....

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times k}} \text{tr}\left[\mathbf{U}'\left(\sum_{i=1}^{M}\mathbf{L}^m - \alpha\sum_{m=1}^{M}\mathbf{U}^m\mathbf{U}^{m\prime}\right)\mathbf{U}\right], \quad \text{s.t. } \mathbf{U}'\mathbf{U} = I$$

Hopefully you recognize the form of the objective. We can define a new matrix, $\mathbf{L}_{\text{mod}}$ and again the first k eigenvectors are the optimal solution. Or,

$$\mathbf{L}_{mod} = \sum_{m=1}^{M}\mathbf{L}^m - \alpha\sum_{m=1}^{M}\mathbf{U}^m\mathbf{U}^{m\prime}$$

Clustering on Merged Subspace

When you cluster on the merged subspace, you get groups with different prognostic interpretations.

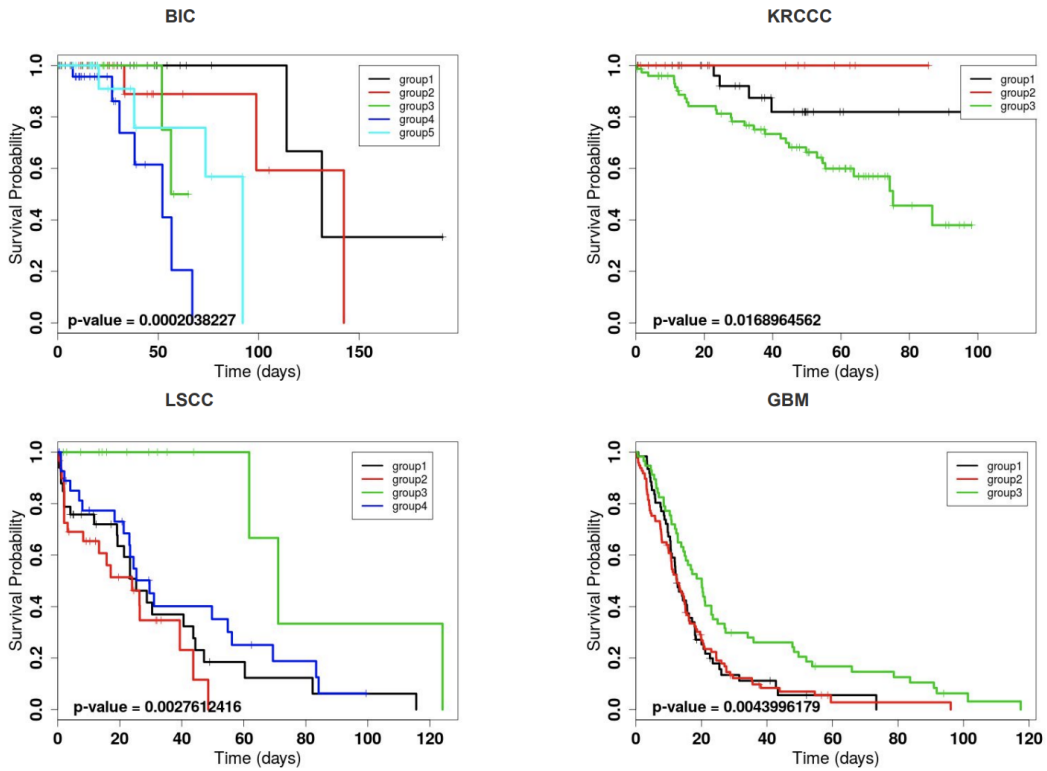


Figure: from Ding *et al.* Bioinformatics. 2018.

Another View : Between Patient Similarity

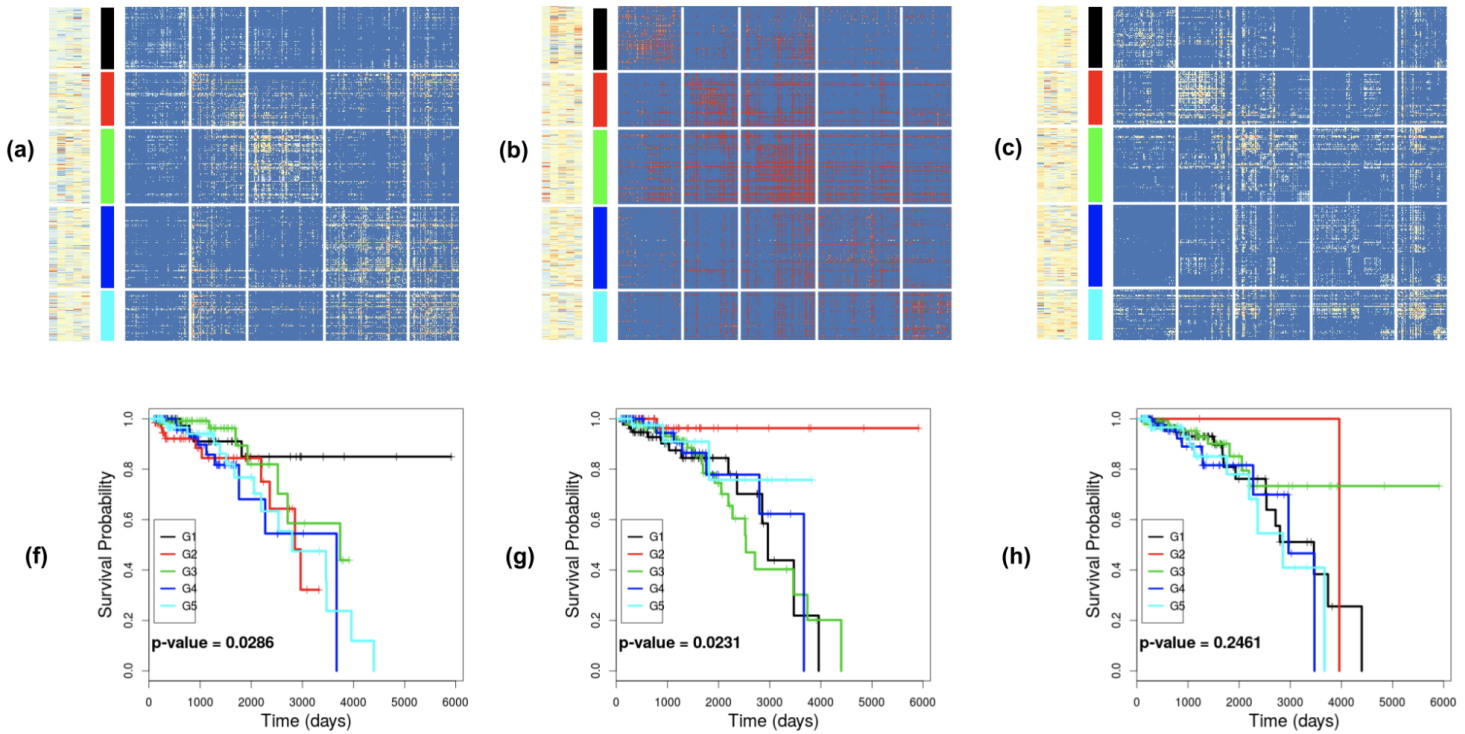


Figure: from Ding *et al.* Bioinformatics. 2018. Here we are viewing adjacency matrices between patients, based on all features jointly.