Comp790-166: Computational Biology

Lecture 12

March 9, 2021

Announcements

- Please send me your project proposals by the end of the day!
- The proposal is worth 60 points total → did you answer the questions (50 points) and did you do the presentation (10 points)
- No class on Thursday as we have another wellness day.
- Homework returned. Let me know if you have any questions.

Today

- Comments on homework
- Detour into multiomics
- Subspace Merging for Multiple Omics Datasets
- Project presentations from $\{$Tarek,Tianyi$\}$, Taksir, and Misha

Homework Observations

- **Language of Choice:** Submissions were mostly in Python, we had 4-5 R submissions, and 1 Julia submission. :D Very interesting

Homework Observations

- **Language of Choice:** Submissions were mostly in Python, we had 4-5 R submissions, and 1 Julia submission. :D Very interesting
- The most frequently wrong question was the adjacency matrix math question, writing the number of edges in terms of $\mathbf{A}$ and $\mathbf{1}$

Homework Observations

- **Language of Choice:** Submissions were mostly in Python, we had 4-5 R submissions, and 1 Julia submission. :D Very interesting
- The most frequently wrong question was the adjacency matrix math question, writing the number of edges in terms of **A** and **1**
- The answer is, $(1/2)\times\mathbf{1}^T \times \mathbf{A} \times \mathbf{1}$

Homework Observations

- **Language of Choice:** Submissions were mostly in Python, we had 4-5 R submissions, and 1 Julia submission. :D Very interesting
- The most frequently wrong question was the adjacency matrix math question, writing the number of edges in terms of $\mathbf{A}$ and $\mathbf{1}$
- The answer is, $(1/2)\times\mathbf{1}^T \times \mathbf{A} \times \mathbf{1}$
- Some of the R users had an issue reading in their graph (because their Louvain NMI was very low). Whenever reading an edgelist, I would always implement several checks and make sure it is indeed undirected, etc.

Homework Observations, Continued

- The majority of you showed that graph-based clustering achieved higher NMI than clustering on original cell × marker matrix. Louvain and Leiden were the most popular methods used, but I also saw some others as well.

Homework Observations, Continued

- The majority of you showed that graph-based clustering achieved higher NMI than clustering on original cell × marker matrix. Louvain and Leiden were the most popular methods used, but I also saw some others as well.
- Classifying T-cells from monocytes was definitely too easy (I knew it would be easy, but not that easy.....)

Homework Observations, Continued

- The majority of you showed that graph-based clustering achieved higher NMI than clustering on original cell × marker matrix. Louvain and Leiden were the most popular methods used, but I also saw some others as well.
- Classifying T-cells from monocytes was definitely too easy (I knew it would be easy, but not that easy.....)
- Most people did not see an advantage to clustering on the node embeddings (in terms of NMI) as opposed to partitioning the original graph

The Overall Problem: Combining Multiple Sets of Features

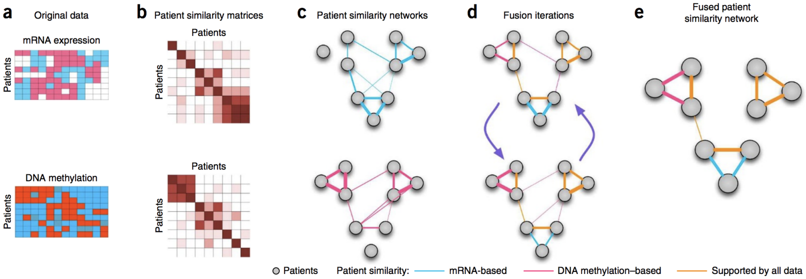


Figure: from Wang *et al.* Nature Methods 2014. The problem is to learn a joint representation of all patients that respects each modality.

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.

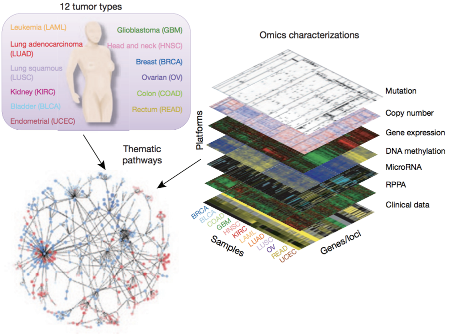
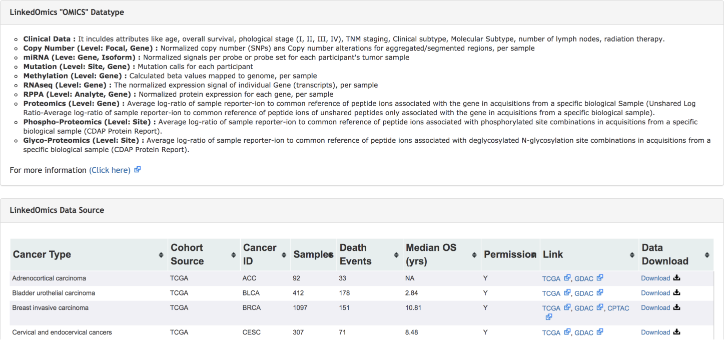


Figure: from TCGA, Nature Genetics. 2013.

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.

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 $p_m \times N$

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

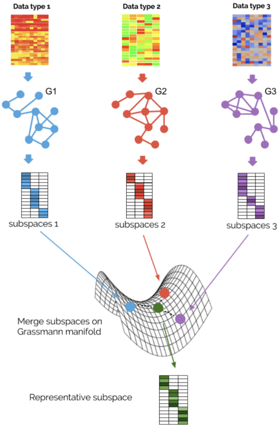


Figure: from Ding *et al.* Bioinformatics. 2019.

Build a Similarity Graph Between Patients in Each Modality

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

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

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, $W_{ij}^m = S_{ij}^m$

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

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \tag{1}$$

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

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,

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

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

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

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,

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

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,

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

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

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.

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, **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,

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

Optimization Problem for Multiple Subspaces

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

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

The authors showed that this simplifies to,

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

Rayleigh Ritz Again....

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

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.

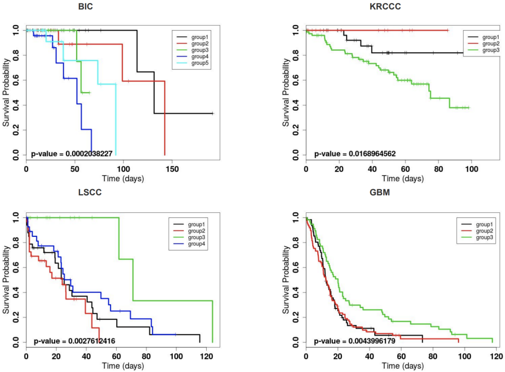


Figure: from Ding *et al.* Bioinformatics. 2018.

Another View : Between Patient Similarity

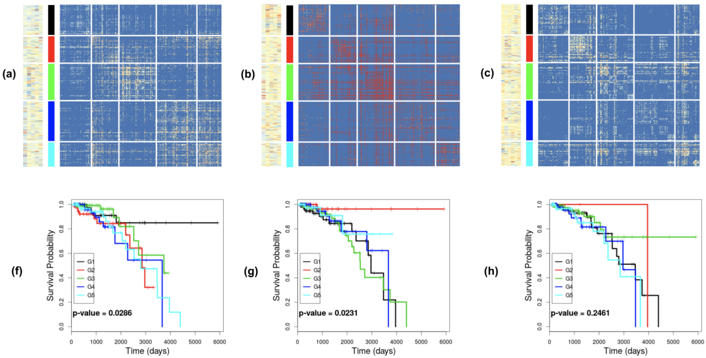


Figure: from Ding *et al.* Bioinformatics. 2018. Here we are viewing adjacency matrices between patients, based on all features jointly.