CS194-302

Joon Kim

Background

RNA is a good proxy measure for protein!

Classical (Bulk) RNA Seq: Systematic errors, (Pseudo) Alignments

↳ What can we say about the gene expression? Methods such as Library normalization, Dimensionality Reduction, Differential Expression.

Why Single Cell? Multiple cell measurements result in average.

↳ Single cell profiles recover heterogeneity! (we don't know about the diversity btwn cells)

Immunology readily adopted sc-omics due to immune cells being singular

Intrinsic Noise from biological data is due to stochastic processes

---

Week 2: Tumors

Immune System: protect the body from pathogens. — adaptive, innate, complement → today!

innate → short term, fixed. adaptive → long term, needs learning

There are different types of immune cells. We are mainly interested in how these function together (systems immunology).

ex) CD4+ T cells have helper subsets, can be beneficial/deleterious.

↳ conduct in vitro (in lab) experiments that control the environments that affect the "evolution" of the cells, and scatterplot markers.

* In vivo, relationships between subsets are not very clean.

ex) In vitro shows an XOR relation, but in nature, they can both exist.

Self vs Not-Self → checker for whether we should attack the cell

Two bins of tumors: Hot(Lymphocyte infiltrating) / Cold(no Lymph. inf.)

↳ CD8 T cells can infiltrate(attack) hot tumor cells

↳ CD4 T cells are not very helpful, sometimes even backfiring!

→ Tregs supress other cells that can attack tumors → not good!

T cell exhaustion → bleeding edge, how to do it is unanswered

Cancer cells try to trick attackers by sending "don't eat me" signals.

↳ CTLA4, PD1, CD47

Week 3: Batch Effects

CAR-T Therapy: train patient's T cells to recognize tumors

↳ Blood tumors work well, solid tumors are in progress

Batch Effects: systematic (experimental) bias → how to overcome?

Curse of Dimensionality: All points are "equally" distant to each other

Principle Component Analysis: Linear transformation for dim. reduction

$X :=$ N by d matrix, covariance $\Sigma := X^T X$ → not diagonal

↳ we need lin. transform P s.t. $(PX)^T(PX)$ is diag$(\lambda_1, \lambda_2, \ldots)$.

Some confounding variables will have biological significance!

Harmony Algorithm: K-means, then shift data closer to centroids

Local Inverse Simpson's Index (LISI) → quantifies diversity of neighbors

↳ integration LISI: effective # of datasets (should → # datasets)

↳ cell type LISI: effective # of cell types (should → ~1)

Reciprocal PCA (rPCA): doesn't assume clusters from datasets!

↳ find "anchor" cells that are similar across datasets, then their difference vector is the batch effect.

1) Normalize dataset, select set of most common genes

2) Reciprocal Projection: PCA, then reciprocally superimpose datasets

3) Find anchors via Mutual Nearest Neighbors

4) Heuristically evaluate the quality of anchors (filtering, weighting)

5) Integration: pick a reference dataset, move others accordingly

Evaluation: how do we evaluate these methods?

Silhouette Scores: $a(i)$:= mean distance in same batch,

$b(i)$:= minimum mean distance to another batch,

$s(i) := \frac{b(i)-a(i)}{\max\{a(i), b(i)\}}$ if $|B_I| > 1$, 0 if $|B_I| = 1$.

↳ we would want this score to decrease (close to 0)

Week 4: Cell Trajectories

Metacells: group cells into a few representatives for strong signals

↳ helps with reducing the sparsity of the matrix.

MC2: Divide-and-Conquer, 2-sided Stability Score

↳ highly parallelizable, $O(N \log N)$ vs $O(N^2)$ ←(MC1)

- how do we break up into subcomponents?
- how do we account for the fact that biology is not uniform?

Preliminary Phase: split cells into random piles, solve each one.

Metagroup Phase: partition metacells into metagroups, recurse.

Final Phase: treat outliers/rare cells seperately

* also detects rare genes and seperately processes them out of DaC.

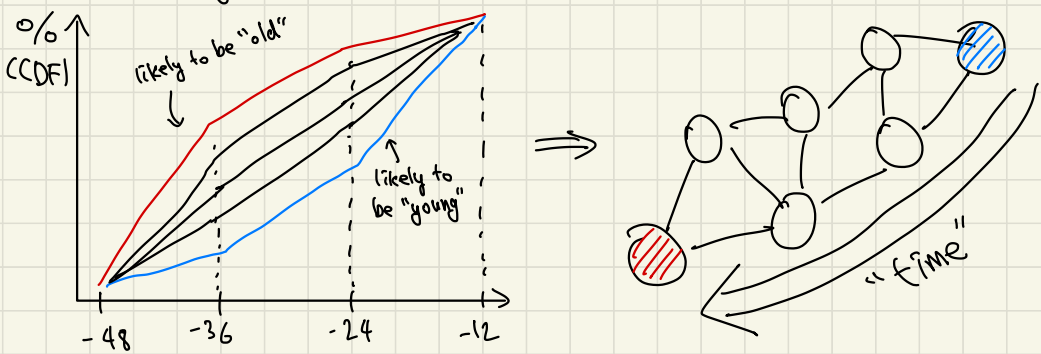
2-sided Stability Score: use iterative updates for accuracy.

Zman-seq: how do we analyze temporal (→ over courses of day) processes in scRNA?

↳ add fluorescent pulse labels for temporal information.

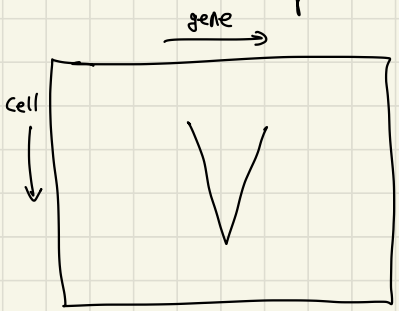
↳ inject different colors at different timestamps to know when a particular cell entered the tumor!

cTET: approximate which metacell entered at what time by specifying a CDF of time entered (12, 24, 36, ... hrs) and calculating the AUC of each function → sense of time!



Week 5: (Consensus) Nonnegative Matrix Factorization

Bulk RNA-seq data → very large and sparse (dropout is a factor) matrix V

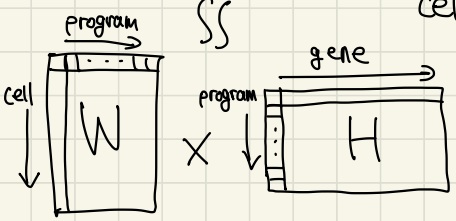


traditional NMF is obfuscated by noise!

also, assignments are single & deterministic

linear analysis is efficient but not rigorous

"cell type" → discrete, "cell activity" → continuous



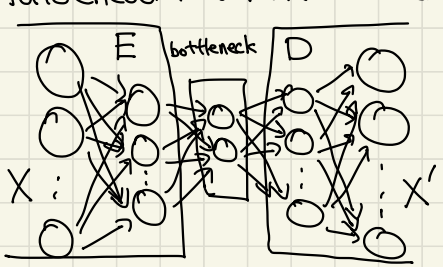
cNMF tries to build a consensus H matrix over various starting points to raise empirical accuracy. Programs need to be grouped intelligently!

Week 6: VAE

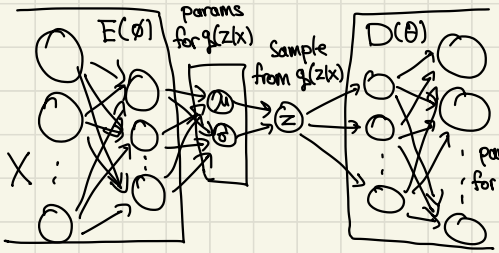
So far: Dimensionality Reduction → Clustering → Diff. Exp.

With VAE: Neural Networks → Clustering → Generative Exp.

Autoencoders: NN for dimensionality reduction; the NN tries to minimize $\|x - x'\|$ while having a "bottleneck" that learns a latent representation of data

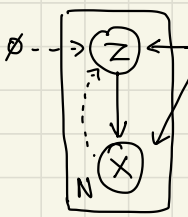


VAE: Probabilistic version of AE, predict a distribution on latent space



↳ this is a generative model, where we try to learn the joint distribution of x(data) and y(label).

* Math: $\log p_\theta(X^{(i)}) = \underbrace{D_{KL}(q_{\phi}(z|x^{(i)}) \| p_\theta(z|x^{(i)}))}_{\text{how far are we away from the answer?}} + \underbrace{\mathcal{L}(\theta, \phi; x^{(i)})}_{\text{ELBO}}$



$$\mathcal{L}(\theta, \phi; x^{(i)}) = -D_{KL}(\underbrace{q_\phi(z|x^{(i)})}_{\text{variational post. (encoding)}} \| \underbrace{p_\theta(z)}_{\text{prior}}) + E_{q_\phi(z|x^{(i)})}\left[\underbrace{p_\theta(x^{(i)}|z)}_{\text{generative model (decoding)}}\right]$$

scVI: z as latent cell representation, x as gene data

"Systems" Biology: Consider interactions between components!

Week 7: Multi-Modal Omics

(CD4 vs CD8 fate)

Today's Story: CITE-seq + totalVI → new bio discovery!

Flow Cytometry is powerful but is a manual procedure.

CITE-seq: RNA-seq + surface protein abundance quantification in sc.

↳ the readout is still sequencing, so a lot of caveats & noise

totalVI: How can we explicitly incorporate protein data in modeling?

↳ "Joint" probability with a VAE! Proteins are modeled as mixture of Gaussians to better fit for foreground/background bimodalness

* Gamma → Poisson mixture is just Negative Binomial, like scVI

Week 8: Cohort Studies

// Our presentation, refer to slides

Week 9: Spatial Omics I

(in-situ)

What if we were able to keep spatial information intact in sequencing?

Two flavors: imaging / sequencing

Imaging: single-cell resolution, but limited in # of genes

↳ Codex, MERFISH, ExSeq

Sequencing: transcriptome-wide measurement, but only near-single-cell

↳ Slide-Seq, (Spot Deconvolution)

Spot Deconvolution: "Separate" information. What is the proportion of cells under this specific bead?

TACCO: spatial deconvolution & categorical annotation methods

↳ "semi-unbalanced entropic optimal transport"

Optimal Transport: $\underset{\gamma \in \mathbb{R}^{|p| \times |q|}}{\operatorname{argmin}} \sum_{a,b} \gamma_{a,b} M_{a,b}$, s.t. $\gamma_{a\emptyset} = p_a$, $\gamma_{\emptyset b} = q_b$ (marginals)

↳ "convex" ⇒ LP!! but # of constraints is pretty large...

Entropic OT: regularize entropy of γ, add $\varepsilon \cdot \sum_{ab} \gamma_{ab} \log(\gamma_{ab})$

Semi-Unbalanced Entropic OT: if we don't know one of the marginal?

⇒ add another term w.r.t. a "prior" $\tilde{q}$: $\lambda \cdot D_{KL}(q \| \tilde{q})$!

TACCO OT: "Objects" → cell-like objects, "Categories" → cell types

* OT doesn't explicitly use spatial data!

↳ TACCO incorporates both spatial & compositional annotation data

TACCO Spatial Framework: (boosters) → comp. annotation → deconvolution

Boosters: Platform Normalization ("change of basis"), Multicenter (k-means), Bisectioning (annotate, then subtract from data)

Object Splitting: derive several virtual observations for each one real

---

Week 10: Spatial Omics II

Physical tumors are 3-dimensional. How to analyze in 3D?

↳ Locations? Local communications? During homeostasis/disease?

Cellular Niches: Zones of tissue defined by mixture of cells & programs

2D → 3D: slicing, take 2D images at multiple places
↳ but how do we stitch these together?

CellCharter: Preprocess with scVI/scArches (VAE), then identify clusters incorporating neighbor information (Delaunay Triangulation!)
↳ Delaunay maximizes minimum angles → empty circumcircles!
Take n-neighbors, aggregate them for final cell representation.
Use EM algorithm on GMM for clustering. ~ # of clusters?
↳ do a parameter sweep and compare overlaps with $k' = k \pm 1$. (...?)
# of clusters ≃ # of subjects can lead to clustering based on subjects.
" >> " eventually leads to subclusters for subjects!

Week 11: scCRISPR Screens

Genetic knockout: originally very difficult & time-consuming
RNAi, ZFN, TALEN, ... CRISPR!
CRISPR: enables easier genome editing, modular & efficient

CRISPR + scRNASeq? cannot read which guide is in which cell

↳ create a "barcode" mRNA ... but this breaks the virus!

↳ we can reverse the entire cassette to bypass this issue

Gene Regulatory Networks: nodes are regulators & targets, directed edges and types show the regulation

↳ Knockout of a TF can lead to many downstream effects

Double knockout experiments can give some evidence (hopefully)

↳ some technical/biological variations, some KOs are better than others

Linear Regression: $Y = X\beta + \varepsilon$ (Y: gene exp., X: predictor, β: effect size, ε: noise) naturally extends to $X \in \mathbb{R}^d$.

MIMOSCA: Y, X, β are all matrices, Y is (cells x expressions)

↳ uses elastic net regularization, $\lambda_1 \sum^{(lasso)} |\beta_j| + \lambda_2 \sum^{(ridge)} \beta_j^2$, L_1 & L_2 penalties.

↳ ridge helps with correlation, lasso helps with sparsity, general denoising

↳ use random permutation to simulate no associations, then compare

* t-tests/p-values don't work because of correlations!

⇒ find TF co-modules/gene programs, genetic expressions

Now we want to look at <u>in vivo</u> cells via CRISRR screening

What genes control T cell states in TME? Which TFs push cells?

Controlled model: 1) tumor with known antigen 2) T cells that recognize it

↳ How to choose which TF to KO? DE/DA to curate 180 TF-library

---

Week 12: Foundation Models

Motivation: How can we scale system biology better?

Challenges: 1) Knowledge is not shared amongst different datasets.

2) Each model is task-specific. 3) Poor out-of-distribution generalization

⇒ Can we build a "universal" model for sc biology?

Foundation Model: self-supervised feature extractor, tokenizes genes!

↳ Core philosophy: Pre-train universally, Fine-tune on demand.

* scRNA data is not sequence but a tablature, how to fit it in?
* some criticism on zero-shot performance being worse than specific models

scGPT: input of (expression values, gene tokens, condition tokens) for a cell

Nicheformer: FM for spatial omics