Challenges in single-cell data analysis

Single-cell meeting

Aziz Fouché, PHD student

Thursday, September 24th

Outline

Mathematical context of single-cell analysis

Single-cell analysis theory

- The state of a cell = a high dimensional object
- A cell population can be seen as the distribution of a high dimension R.V.
- Connections with algebra, analysis, geometry & probability theory

Mathematical context of single-cell analysis

Single-cell analysis theory

- The state of a cell = a high dimensional object
- A cell population can be seen as the distribution of a high dimension R.V.
- Connections with algebra, analysis, geometry & probability theory

High dimensional riddles

- Defining & assessing the complexity of a cell population and sampling quality of an experiment
- Identify the common cell subpopulations between datasets
- Integrate global information between data types (RNA-seq, ATAC-seq...)

Q1: How to define the complexity of a distribution?

Intuition: What is a *complex* distribution? A *simple* one?

- Correlation complexity - number of parameters
- Correlation complexity - intrinsic dimensionality
- But, how to define complexity formally? Information theory?

Q1: How to define the complexity of a distribution?

Intuition: What is a *complex* distribution? A *simple* one?

- Correlation complexity - number of parameters
- Correlation complexity - intrinsic dimensionality
- But, how to define complexity formally? Information theory?

Approaches

- Top-down: coarse-grain the distribution iteratively until no changes [1]
- Bottom-up: approximate the distribution using 1, 2, ..., n normal distributions

Q2: How to assess if you need more samples?

A connected but different question

- Given a sample, do you need to sample more?
- Important question in all data science, as sampling = money
- For now, rule of thumbs in most cases

Q2: How to assess if you need more samples?

A connected but different question

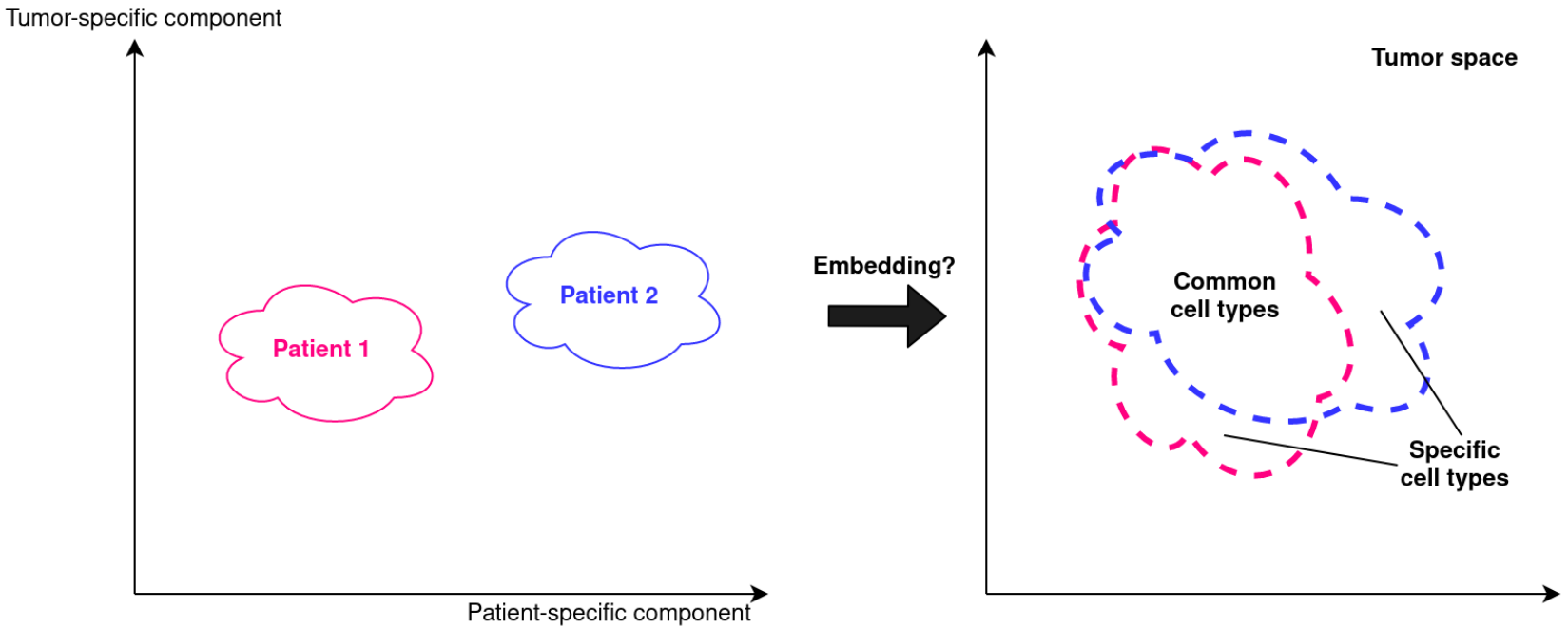
- Given a sample, do you need to sample more?
- Important question in all data science, as sampling = money
- For now, rule of thumbs in most cases

Intuition: predictability is the key

- We need to sample more if we can get *surprised* by new data
- We need a procedure, maybe linked with bootstrapping or classification

Horizontal data integration (multi-patients)

How to get rid of patient-specific information?



Approaches (1)

Graph-based methods

Use algorithms such as MNN to identify *anchors* between datasets, which can be used as reference points for the alignment [2]

Component-based methods

Seek relevant subspaces with insightful basis vectors (PCA, CCA, ICA...), in which correcting biases is easy

Procrustean methods

State the problem as an optimization problem, allow translation, rotation and scaling of datasets

Approaches (2)

Latent space methods

Use kernels and abstract feature spaces such as Reproducing Kernel Hilbert Space (RKHS) to embed datasets in a more convenient space

Optimal transport based metods

Work directly on the distributions *mass* and region densities to identify common cell populations.

And probably a lot more!

A naive prototype to integrate cell cycle

Figure: A typical cell population in cell cycle space $\{G1/S, G2/M\}$

A naive prototype to integrate cell cycle

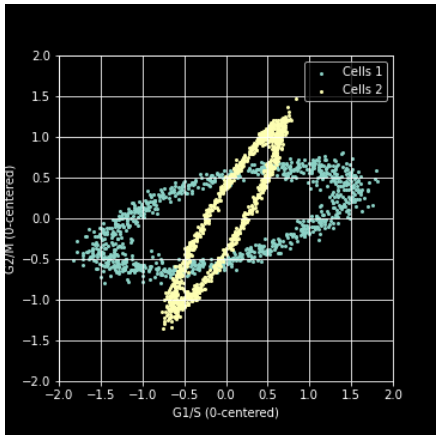
Figure: A typical cell population in cell cycle space $\{G1/S, G2/M\}$

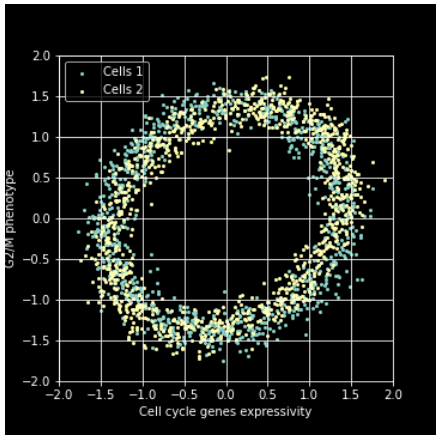
Idea: Procrustean elliptic correction

- Center the loop
- Detect the long axis angle with linear regression
- Rotate the ellipse along the x-axis (rotation)
- Standardize its standard deviations (scaling)

Results, limitations

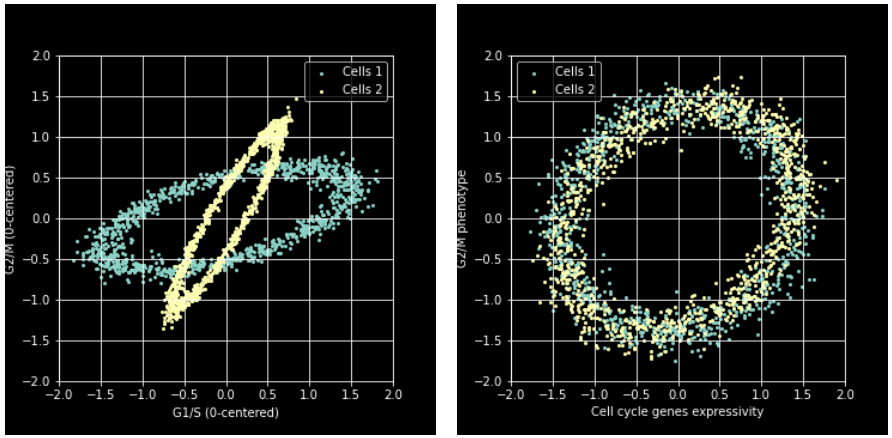
On synthetic data, works as expected



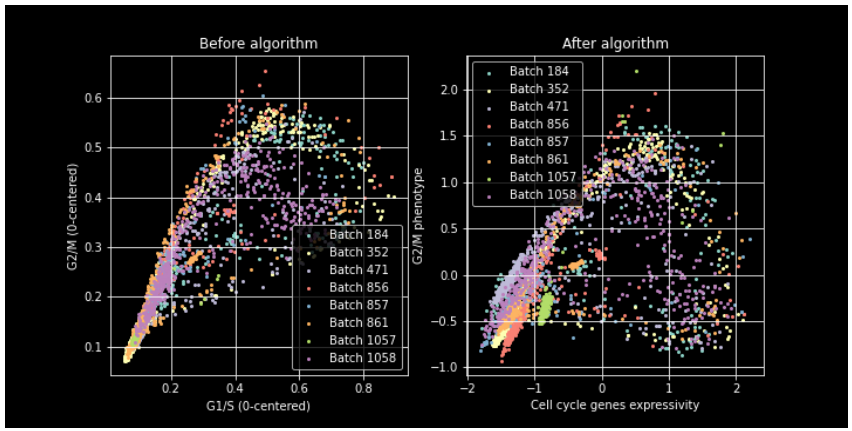


Results, limitations

On synthetic data, works as expected

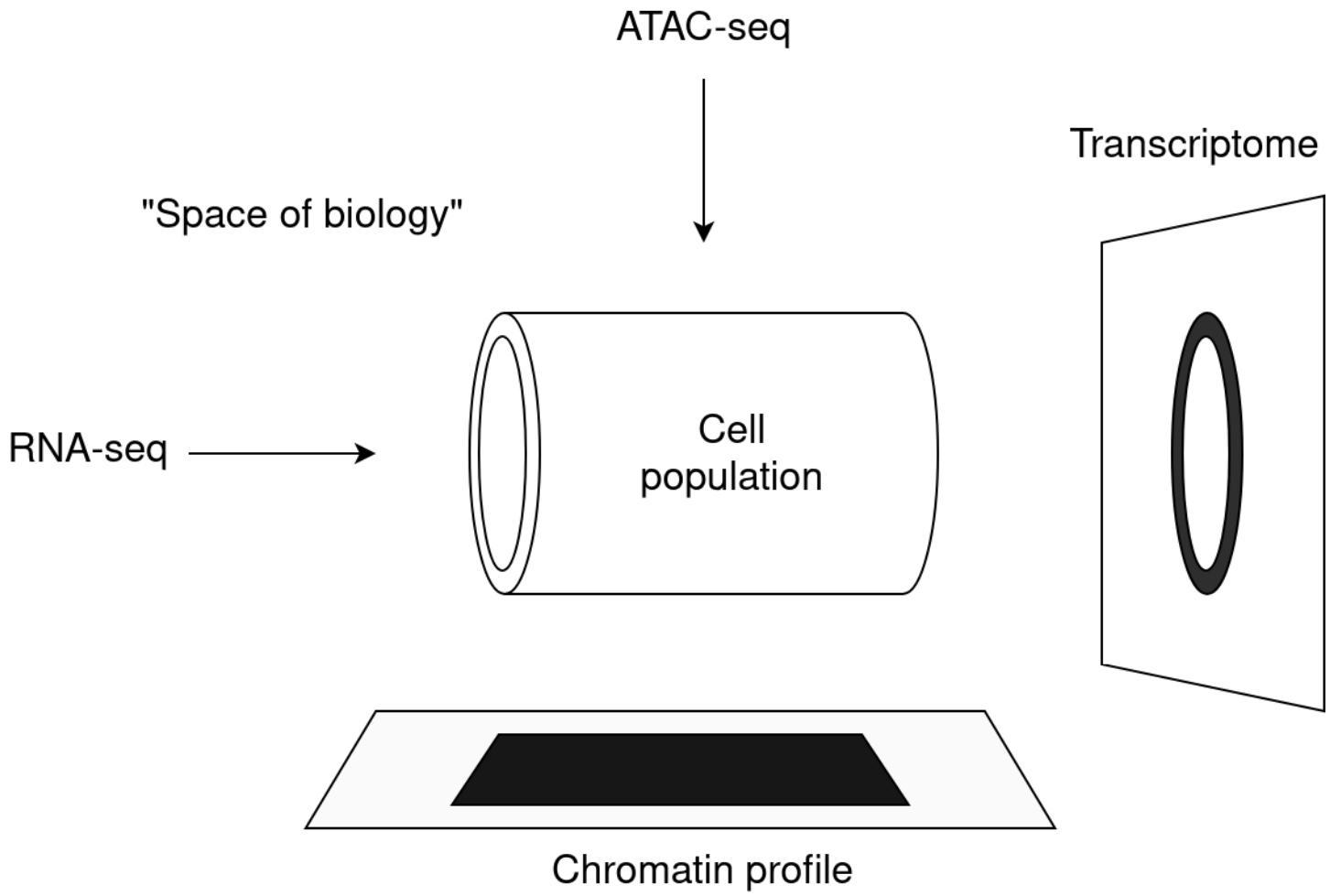


On real data, some limitations (and maybe no real purpose?)



Vertical data integration (multi-omics)

How to reconstruct the whole object from a set of projections?



Approaches

The problem is difficult

- Datasets do not live in the same world
- Datasets are not coupled (same distribution but cells are different)

Discover a common feature space

For instance, a RKHS, but which dimensionality? [3]

Using the locality assumption

Nearest neighbor structure is preserved between the different views
Density-based anchoring algorithms?

Conclusion

- There are a lot of exciting unsolved problems related to single-cell analysis
- There can be found connections to many mathematics/computer science fields
- Hard problems, but with direct applications in reality
- A hard problem may not imply a complex solution...

Bibliography

Andrey A. Bagrov, Ilia A. Iakovlev, Mikhail I. Katsnelson, and Vladimir V. Mazurenko.

Multi-scale structural complexity of natural patterns.

CoRR, 2020.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Marlon Stoeckius, Peter Smibert, and Rahul Satija.

Comprehensive integration of single cell data, 2018.

Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble.

Jointly embedding multiple single-cell omics measurements.

BioRxiv, page 644310, 2019.

Thank you!

Additional figure: centering the loop