Mathematical Frameworks for Integrative Analysis of Multi-omics Biological Data

This manuscript (permalink) was automatically generated from BIRSBiointegration/whitePaper@f9acb26 on June 24, 2020.

Authors

- **Kim-Anh Lê Cao**
  0000-0003-3923-1116 · mixOmicsTeam · mixOmicsTeam
  Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Australia · Funded by Grant National Health and Medical Research Council Career Development fellowship (GNT1159458)

- **Aedin C Culhane**
  0000-0002-1395-9734 · aedin · AedinCulhane
  Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA; Biostatsitics, Harvard TH Chan School of Public Health, Boston, MA, USA · Funded by Chan Zuckerberg Initative, NIH, DoD (need to get grant IDs)

- **Elana Fertig**
  0000-0003-3204-342X · ejfertig · FertigLab
  Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA; Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA; Department of Applied Mathematics and Statistics, Johns Hopkins University Whiting School of Engineering, Baltimore, MD, USA · Funded by National Institute of Health, National Cancer Institute; National Institute of Health, National Institute of Dental and Craniofacial Research; Lustgarten Foundation; Emerson Foundation; Allegheny Health Network

- **Jane Roe**
  XXXX-XXXX-XXXX-XXXX · janeroe (PLEASE COPY/PASTE DO NOT EDIT THIS ONE) · XXX
  Department of Something, University of Whatever; Department of Whatever, University of Something · Funded by XX

Abstract

Introduction to single cell and imaging multi-omics

Current multi-omic technologies

Challenges for interpretation

Need for technology-specific questions and analysis methods vs one size fits all data blender

Case studies

scNMT-seq as a case-study for epigenetic regulation

Overview and biological question

Computational challenges

- Identification of multi-omics signatures that characterise lineage, stage or both.
- Handling missing values
- Do epigenetic changes in some genomic contexts affect cell fate decision more than others? If so, how?

Methods for stats/maths analyses and results summary

scRNA-seq + FISH as a case study for spatial transcriptomics

Overview and biological question

Computational challenges

- Can scRNA-seq data be overlaid onto seqFISH for resolution enhancement
- What is the minimal number of genes needed for data integration?
- Are there signatures of cellular co-localization or spatial coordinates in the non-spatial scRNA-seq data?

Methods for stats/maths analyses and results summary

Spatial proteomics and cross-study analysis

Overview and biological question

Computational challenges

- Integrating partially-overlapping proteomic data collected on different patients with similar phenotypes
- Integration of spatial x-y coordinate co-location and co-expression
- Integration with other ’omics datasets (e.g., scRNA-seq) to support the results of these proteomic analyses
- Can we predict the spatial expression patterns of proteins measured on mass-tag but not measured in the MIBI-TOF data?
- What additional information can we learn about the different macrophage and immune populations in breast cancer by conducting integrated analyses of these datasets?

Methods for stats/maths analyses and results summary

Overview of common analytical methods spanning technologies / case studies

- matrix factorization
- neural network / autoencoders

Software strategies to enable analyses of multimodal single cell experiments

Key questions

- How should multimodal single cell data be managed for interactive and batch analyses?
- What methods will help software developers create scalable solutions for multimodal single cell analysis?
- How can we ensure that visualization methods that are central to multimodal single cell analysis are usable by researchers with visual impairments?

Data management strategies

- Abstract data type: “multiassay experiment”. This reflects the idea that each mode will be characterized by a different collection of features on possibly non-overlapping collections of samples. The metadata on features should be clearly and conventionally defined. For example, genes and transcripts are enumerated using Ensembl catalog identifiers; regions of accessibility are defined using genomic coordinates in a clearly specified reference build. Metadata on samples must include all relevant information on experimental conditions such as treatment, protocol, and date of technical processing.
- Serializations and data access methods for
  - spatial transcriptomics
  - scNMT-seq ...

Scalability strategies

Reducing barriers to interpretable visualizations

An overview of the issues with impaired color perception

US Government tools for accessibility

Details of working components

Type	Brief name	Description	URL	Author email
R data class	MultiAssayExperiment	unify multiple experiments	bioconductor.org	many
R package	Giotto	Spatial transcriptomics	...	...
python library	PyTorch	deep learning	...	...

Techniques and challenges for benchmarking methods

Outline:

- We must first define what we are benchmarking
  - recovery of cell types / clusters
  - discovery of relationships between data modalities, e.g. gene regulatory relationships observed between chromatin accessibility and gene expression
  - ...
- Strategies for benchmarking
  - simulation (and we can discuss the difficulties with simulating covariance structure across features and data modalities)
  - benchmarking datasets
  - cross-validation within study (and we can discuss issues in matching dimensions of latent space across folds). For this Mike has a lot of literature in Google Doc to include on papers that have performed either permutation or cross-validation to assess model performance.
  - cross-study validation (are relationships discovered in one dataset present in other datasets, potentially looking across single cell and bulk)

Discussion

Emerging analytical methods and technologies

Community needs for data structures, analysis methods, etc

Glossary

Consensus term	Synonyms	Description
Network	Graph	A set of *nodes*, representing objects of interest, linked by *edges*, representing specific relationships between nodes.
Node	Vertex	Element of interest in a network and linked to other nodes. For example: people, cells, proteins or genes. Nodes can have several properties called *attributes* like cell type or position.
Edge	Link	The relationship between 2 nodes in a network. For example: friendship in social networks, cells in contact in a spatial network, or gene-gene interactions in a gene regulatory network.

References