ML4G review

Papers:

2022 exam paper: https://pmc.ncbi.nlm.nih.gov/articles/PMC8490152/

AL questions:

Describe the method

- key aspects
- explain the objective function (/loss)
- what parameters were used, how were they chosen?
- prior belief – what does it model? why? (for Bayesian models)

Evaluation metrics:

- which of the used is the best?
- why evaluated this way?
- other way to benchmark (alternative metrics)
- describe the experimental setup (datasets, how evaluated etc.)

Advantages/disadvantages compared to other methods

Terms:

- nucleotides (A, T, G, C)
- Gene Exons Introns 5'-UTR (near gene start) 3'-UTR (near gene end) TSS TES Promoter Enhancer
- Nucleosome: the bead made of histones
- Chromatin: DNA wrapped around nucleosomes
- Epigenomics: Analysis of chromatin (e.g. histone marks etc., accessibility, looping)
- Histone: grouped form beads (nucleosomes)
- Histone modification: (acetylation, phosphorylation, methylation) indirectly causes e.g. chromatin opening. Methylation decreases transcription.
- Chromosome: unit of genome, continuous chromatin strand
- Genome: all chromosomes (all DNA)
- Transcriptome: all RNA from the genome (including mRNA and non-coding RNA)
- Central dogma: DNA (transcription) -> RNA (translation) -> protein

- RNA polymerase [-ase: catalyzing]: does transcription
- Ribosome: does translation
- Promoter: regulatory sequence immediately upstream of TSS (binding site for RNA polymerase and TFs that initiate transcription); starts with TATA block
- Enhancer: regulatory sequence increasing gene expression, can be far away [1Mbp] (DNA looping)
- Transcription factor: bind at promoters/enhancers, help dispose nucleosomes
- Binding motif: where transcription factors bind
- DNA methylation (one of histone modifications)
- polymer (poly[many]-mer[parts]): DNA/RNA
- oligomer (oligos[few]-mer[parts]): short DNA/RNA sequence
- DNA looping: enables promoter-enhancer interactions (even if very distant)
- CTCF: ▲ transcription factor which determines TAD (topologically-associated domains, where prom. and enh. are more likely to interact)
- YY1: protein which gives promoter-enhanced interactions (similar to CTCF)
- In vitro, In vivo, In silico: (outside living, inside living, computer)
- SNP (single-nucleotide-polymorphism -common in a lot of people) vs Mutation
- locus (plural: loci): region of genome; e.g. where given gene variants reside in chromosome
- allele: one version of a gene in particular locus
- phenotype: observable characteristics of organisms (e.g. eye color, rat tail wrinkles)
- germline (all cells of children) vs somatic mutations (local change)
- DNase I, DNase II: enzymes that help cleave (cut) DNA to dispose of it (e.g. dying cells)
- CpG: "In humans, about 70% of promoters located near the transcription start site of a gene (proximal promoters) contain a CpG island." – can be used as a feature for genome methods
- cis-regulatory vs trans-regulatory loci (direct [e.g. promoters, enhancers] vs activing trhough a proxy)
- lysis: breaking up nuclei to release its contents (for experiments)
- ligation: join DNA pieces using DNA ligase (for experiments)
- sonication: shear DNA into smaller pieces using sound waves (for experiments),
- MSA (multi-sequence alignment),
- poly-A tail: AAAAA (multiple) tail present in mRNA (coding) [most of them]
- library size: (in scRNA-seq) total number of read counts per cell
- FPKM: fragments per kilobase of transcript, per million mapped reads

$$FPKM = \frac{\left(\frac{read\ count_{gene}}{\frac{total\ exon\ length_{geme}}{10^3}}\right)}{\frac{read\ count\ (total)}{10^6}} = \frac{read\ count_{gene}}{total\ exon\ length_{geme} \cdot read\ count\ (total)} \cdot 10^9.$$

- RPKM: reads per kilobase of transcript, per million mapped reads (similar, but for single-ended reads)
- UMI: unique molecular identifier – single read in scRNA-seq
- barcode: bead id
- DGEX: differential gene expression analysis
- marker gene: gene that is siginificantly differently expressed for two cell types
- DEG: differentially expressed gene
- GSEA: gene set expression analysis – find links between a set of DEG and known pathways

- H&E: hematoxylin and eosin: used to stain histological images (highlight nucleus and cytoplasm)

Experiments:

Identifying loci (differentially-expressed for trait):
- GWAS: genome-wide association study: Shows SNP positions correlated with trait. Thousand individuals (genomes) grouped based on phenotype/disease (cases/controls).
- eQTLs: expression quantitative trait loci: locations where mutations cause change in expression

Protein binding detection:
- SELEX (small sequences in vitro): [10-14bp] in-vitro protein binding detection
- PBM (polymer binding microarray): [<=10bp all-kmer arrays] in-vitro protein binding detection
- ChIP-seq: (chromatin-immunoprecipitation-sequencing) genome-wide in-vivo protein binding detection
- CUT&RUN

Chromatin opening
- DNase-seq
- ATAC-seq

Chromosome conformation capture (3C) [looping]
- 3C, 4C, 5C, Hi-C, HiChIP.
- output: pairwise interaction scores (1Kb resolution) (within chromosome)

Bulk gene expression:
- bulk RNA-seq
- output: gene expression vector

Single cell gene expression:
- scRNA-seq
- output: cell × gene matrix of expression

Cell type annotation:

Flow cytometry

Spatial:
- multi cell (spot-based)
  - Visium 10x
- single cell
  - CODEX (paint with channels)
  - Visium HD
- sub-cell
  - Xenium
  - MERFISH

Lectures:

Genome-related methods

1. Lecture1_Introductory.pdf: Genome basics, Genes, Central dogma, regulation, DNA looping

Genome basics:
- DNA: exons + promoters/enhancers/silencers + introns (regulatory) + more
- nucleotides forming pairs A-T, G-C
- [3.1 * 10^9 x 2 copies in human genome]

Genes:
- gene:

```
[promoter] <--~100bp---> [TSS within 5' UTR]----------[exon]--------------[exon]------------[3'UTR]
```

- gene sequence wrapped around nucleosomes (forming beads)

```
[140-147bp around nucleosome][20-90bp link][140-147bp around nucleosome][20-90bp link]...
```

- genes:
  - protein-coding (20k in human genome)
  - long non-coding RNA (regulatory, 25k)
  - micro-RNA (regulatory, 2k)
- histone modifications are associated with e.g. active enhancers (serve as markers)

Central dogma: (DNA (transcription) -> RNA (translation) -> protein)
- transcription: DNA sequence has introns and exons. Exons are spliced-out with some probability (alternative splicings).
- translation (in ribosome): each RNA triplet codes some amino-acid / STOP sequence

DNA looping could be caused by a mutation, which then enables interactions causing tumors

Lecture 5: Chromosome regions are grouped in compartments (ch19 has 6), which closely interact, and have different histone marks.

2. Lecture2_WorkingWithDNA.pdf: Binding site prediction

Distinction:
- - SNP vs Mutation
- - Germline vs somatic mutations
- - Coding vs non-coding mutations

- GWAS
- eQTLs (expression-Quantative-Trait-Loci); eQTL is a locus which is associated with a

measurable gene expression trait. Variations in eQTLs cause changes to expression.

Binding site prediction using:

- PWM: get fixed-length sequences & create a matrix of position×(A,C,T,G) scores (weights)
- 1D-CNN:
  - use nucleotide (A,T,G,C) sequence as 4 channels (1-hot encoded) [nicely interpretable kernels]
  - can use PWM to init kernels
  - can use dilated convolutions to increase receptive field

Ground truth data:

SELEX (small sequences in vitro)
10-14bp
PBM (polymer binding microarrays):
position sequences on an array
bind tagged TFs
use fluorescent stuff that binds to tags
inspect which positions glow
ChIP-seq
1. Let proteins bind to CHromatin
2. Break the strands
3. filter specific protein (IP – immuno-precipitation)
4. read sequences linked to proteins
5. map sequences back to the geneome

3. Lecture3_WorkingWithDNA.pdf

Paper 3

Improvements:

1. Can use k-mer encoding (by itself)
2. Can use k-mer encoding + word embeddings
3. Can apply RNN (recurrent) or biLSTM/GRU (several layers) [not as easily interpretable, computationally intensive]
   - can stack LSTM on top of CNN to capture global dependencies between "regulatory grammar"
4. Can use attention.

AL paper:

Q1: Compared to SpliceFinder
(https://link.springer.com/article/10.1186/s12859-019-3306-3),
what is the major weakness of SpliceRover?
Q2: Which of the evaluation metrics used (Table 1) is most
suitable for this task and why?
Q3: How do the authors solve the problem of “black box” deep

neural networks? Please, describe in detail.

4. Lecture4_WorkingWithChromatin.pdf

Note: DNase-seq etc. are peformed per cell-type (due to differences in chromatin accessibility).

DNase-seq (chromatin accessibility):

1. Let DNase eat DNA where chromatin is exposed
2. Mark the eaten ends & connect beads
3. Cut long strands away from the beads
4. Mark the other ends.
5. Amplify with PCR.
6. Map reads back to genome.

ATAC-seq (Assay for Transposase-Accessible Chromatin):

1. Let transpostase with markers bind where chromatic is exposed.

?. Cut?

2. Amplify with PCR.
6. Map reads back to genome.

5. Lecture5_Folding.pdf; Prediction od 3D structure of (chromatin) & (proteins)

Paper 3

Mutations in CTCF binding sites can influence chromatin folding.

Can influence promoter/enhancer interactions.

Can influence gene expression.

Predicting 3d structire helps answer the question:

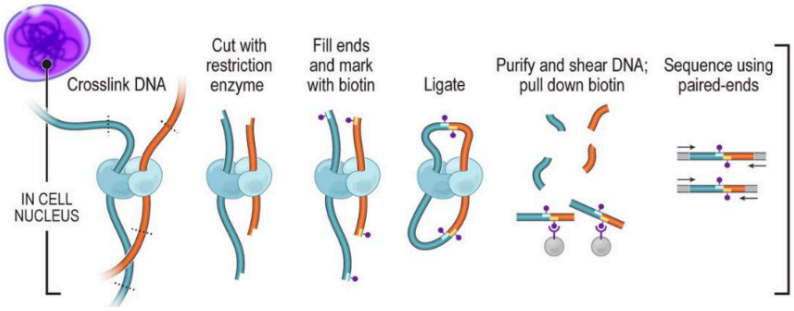
Which genes could be affected by mutations in cis regulatory elements?

Chromosome conformation capture (3C) techinques:

sequence ligated (glued) strips of DNA which interacted with each other

more or less:

1. let protein attach [in vivo, nucleus] and trigger interaction
2. lysis (break cell nucleus), sonication (shear strands), ligation (join those close [co-interacting])
3. [optional Hi-ChIP] ChIP-like filtering (pull-down) of proteins of interest (e.g. particular TFs)
4. PCR and map-back.



Histone marks help predict looping.

DeepC (CNN):

- predict Hi-C from DNA sequences,
  - process Hi-C by bucketizing scores based on quantiles
  - **makes the model cell-type specific!**
- pretrain on chromatin data (DNase-seq, CTCF, TFs and other histone marks),
  - weight transfer

Note on cell types:

- the same DNA,
- different transcription factors expressed,
- different regions are open,
- different folding.

Less than 50% of genes are expressed for a given cell type.
Expression vectors vary between cell types.

DeepTACT:

- also uses chromatin accessibility data,
- broken model.

TransEPI:

- CNN, max-pool, attention, FC.

3D protein folding:

- MSA (multiple species alignment) – use aligned information across species (to see analogous proteins and which parts are important during evolution),
- Models:
  - RaptorX,
  - AlphaFold.

DNA 3D interactions:

- open chromatin regions interact with open chromatin regions
- regions with similar histone marks (active or repressive) interact with each other

Protein 3D interactions:

- multiple species alignment (MSA) provides information about co-mutated and thus possibly interacting residues

C. Origami, Transformer-based approach for chromatin folding prediction. Input: DNA sequence + CTCF binding + chromatin accessibility https://www.nature.com/articles/s41587-022-01612-8 - can be an exam paper

---

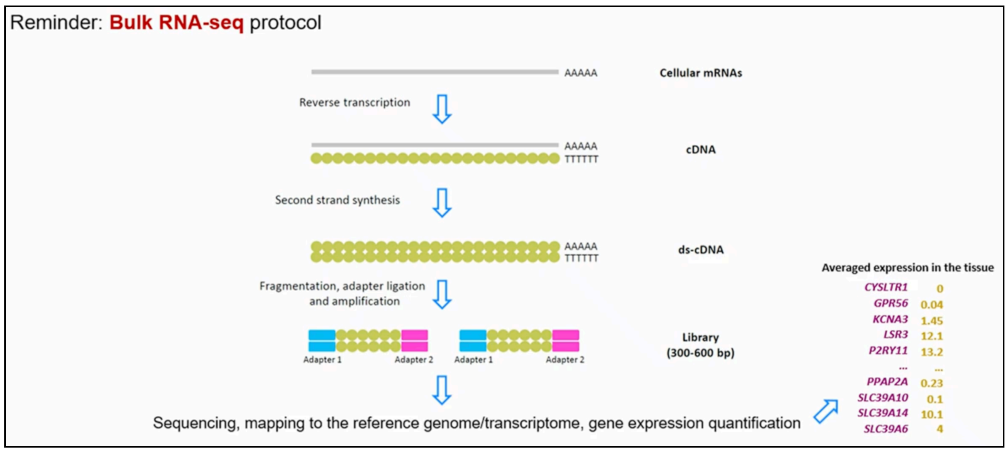
Bulk transcriptomics

6. Lecture6_Deconvolution.pdf

Protein expression is difficult to perform. We mesure RNA expression instead.
DNA (inside cell nucleus) → RNA (outside, interact with ribosomes) → Proteins.

Why deconvolve?

Bulk data is cheaper.
Bulk works better on frozen samples.
Can provide information on cancer composition.
(E.g. some drug could not work for a cell type, better to know this cell type is there)
(Easier to make decisions on treatment selection → e.g. activate cancer-fighting T-cells)



RNA-seq:

- Extract RNA from cells.
- Filter RNA
  - (e.g. by removing ribosomal rRNA).
  - e.g. by doing poly-A selection -- most coding RNA has poly-A tail
- Convert to DNA (cDNA using reverse transcription [from viruses]
  - more stable than RNA,

  - amplification and sequencing tools use DNA.
- Cut into similarly sized DNA.
- Amplify, map back to genome.
- Normalize for library size (total number of reads) and gene length.

Deconv – 2 paradigms:

- blind,
- reference-based (supervised) = blind + scRNA-seq.

Blind:

- FastICA (independent component analysis):
  - $X \approx A \cdot S$ (mixing times source),
  - assumption: sources independent,
    * minimize mutual information (≡ maximize non-gaussianity),
    * maximize "distance to normality" $J(X) = H(\mathcal{N}(\mu_X, \sigma_X)) - H(X) \geq 0$,
      · (use approximation)
  - preprocess X by applying whitening (decorelate)
- **Questions:**
  - how many sources K?
    * test empirically how well given K allows to classify patients (e.g. gender, status of a tumor)
- NWF (negative matrix factorization)
  - $\min_{W,H} \frac{1}{2}\|X - WH\|_F^2$, X, W, H non-negative (W – sources, H – mixing),
  - better applicability (models biologically well – non-negative values),
  - convex NMF makes proportions sum to 1.

Reference-based:

- provide additional data
  - marker gene expression matrix (CIBERSORT - support vector regression) ,
  - scRNA-seq (MuSiC)
  - set of marker genes

Evaluate deconvolution by computing:

pearson correlation,
RMSE (together with corr),
MAD (mean absolute deviation).

7. Lecture7_SingleCell1_IntroNdropouts.pdf – scRNA-seq data

scRNA-seq:

1. Beads are prepared.
   (a) bead = (PCR handle, barcode, UMIs, oligo-dT)
      i. PCR handle (for later PCR amplification)
      ii. barcode = bead id = cell id

iii. UMI (unique molecular identifier) = read id [for filtering after PCR, for computing actual counts]
iv. oligo-dT = primer used to capture mRNA by binding specifically to the poly(A) tail of gene-coding mRNA

2. In each oil droplet, there is a **bead** with a **single cell** (outliers filtered out).
3. In each captured droplet, **cell membrane disrupted**, so that **cell's RNA transcript connect to the bead** – poly-A tail connects to oligo-dT.
   (a) Many transcripts per bead.
   (b) Roughly 90% transcripts **get lost → dropout effect**
      i. high chance of dropping low-probability transcripts
4. Reverse transcription results in (PCR handle, barcode, UMI, 30xT, cDNA, PCR handle).
5. Bead gets disolved.
6. PCR (polymerase chain reaction) is used to **amplify (create multiple copies)** the resulting DNA fragments. **TO MAKE THEM DETECTABLE DURING SEQUENCING.**
   (a) Nowadays, all transcripts are roughly equally amplified :) [no need for corrections]
7. After detection, duplicates are removed to form **reads** (or **UMIs**).
8. Map unique reads (i.e., the cDNA fragments) to a reference genome.
9. Reads measure expression for a given gene region and are associated back to actual cells using barcode ids.

Low-dim representation:

- PCA → can inspect technical/biological biases
- t-SNE
- UMAP
- Autoencoders – scVI (use bottleneck as low-dim repres.)
  - goal: minimize $\|x - f(x)\|_2^2$

Resolving dropout effect:

Autoencoder imputation (scVI):

- ZINB Variational autoencoders
  - zero-inflated negative binomial models scRNA-seq read counts well
    * ZINB-loss instead of squared errors
  - Conditional autoencoders
    * can correct for batch effects (add additional on-hot encoded categorical variables to both the input and bottleneck, so that:
      · encoder can differentiate and account for this
      · bottleneck does not have to include this information
      · [this approach needs additional penalty]
  - output: mean, dispersion, dropout (ZINB parameters),
    * mean vector is the imputed transcript count!
- evaluation: bulkify imputed scRNA-seq data, compare against bulk RNA-seq, see if improves compared to no imputation

MAGIC (markov affinity-based graph imputation of cells)

Main idea: smooth out similar (neighboring) cells, in a fancy way

- normalize read counts (CPM),
- apply PCA → 20-100 dims,
- compute pairwise distances in PCA-space,
- compute affinities $A(i, j)$, matrix M_{ij}
  - apply gaussian kernel,
    - σ_i = distance to the n-th neighbor [n = hyperparam]
  - sparsify data (keep only 3n largest values non-zero)
  - row-normalize to create transition probability matrix M
- raise M to power t (diffusion time)
  - small t: oversmoothing
  - good t: noise removal
  - large t: might lead to signal removal
  - **how to choose?**
    - pick t such that there is no significant increase in R squared score: $R^2\left(M^t X, M^{t-1} X\right) \geq 0.95$
- rescale final values (column-normalize + rescale to 0.99th percentile of original column values)
  - because imputation can change the scale

SAVER (extra AL paper)

8. Lecture8_SingleCell2_BatchGEXAnnot_JY.pdf

Paper 1

scRNA-seq analysis methods:

- Batch correction
- clustering,
- differential gene expression
- cell type annotation

Why?

understand heterogeneity of cells

Batch correction

Idea: correct for any non-biological variability between different datasets.

Linear:

...

Dimension-reduction based:

- PCA

- CCA

NB, bayesian framework:
- ComBat

Harmony:
Shift batches in PCA space based on regularized soft clustering.

Deep learning based:
- Conditional autoencoders,
- transformer-based.

Clustering

Very high dimensionality, traditional methods need dim. reduction.

- K-means based
- Hierarchical
- **Community detection based (graph):**
  Louvain
  Leiden (extension of Louvain)

Cell type annotation

1. Cluster cells.
2. Find significant genes (e.g. Wilcoxon test / t-test / fold change)

Gene set enrichment analysis (GSEA)

Now that I have marker genes;
- what do they mean?
- how do they interact together (in pathways)?
- how can I compare them across datasets?

→ Computes "enrichment score" saying how much given "gene set" is correlated with a phenotype.

It does so by ranking DEGs and performing a walk...

Cell type annotation cont.

Marker-gene or reference based.

Marker based:
1. Cluster cells.
2. Find DEGs.
3. Assign cells to types based on DEGs (marker data).

GSVA: Gene set variation analysis.
GSEA but without phenotypes (ranks computed cleverly in additional steps).

deepMNN: Deep Learning-Based Single-Cell RNA Sequencing Data Batch Correction Using Mutual Nearest Neighbors

Questions:
1. What is the loss used to train the residual neural network? Describe the different components of the loss and their function.
2. How is the performance of the methodology computed? Describe the experimental setup used for evaluation.
3. Is the use of residual blocks justified in the paper? If so, explain. If not, explain how it can be verified if the residual block indeed plays a role.

9. Lecture9_Trajectories.pdf

Learn intermediate states between cell types Not only that, also actual paths visualized.

Cell differentiation is governed by TFs.
- developmental differentiation,
- transdifferentiation (imposed artificially by overexpressing some TFs).

t-SNE/UMAP retrieve gradual change nicely, but they fail to separate the individual paths (all paths are merged and smoothed out).

PHATE
1. Assume scRNA-seq data.
2. Normalize for library size.
3. Compute pairwise distances, map them through gaussian kernel (like t-SNE).
4. Use P like markov chain transition probabilities, raise to power t.
5. Create information-based distances:
   (a) $dist_{ij} = \sqrt{\left\| \log P_i^t - \log P_j^t \right\|^2}$ square root of norm between two rows,
      i. two cells are close if they are likely to go to the cells
      ii. intition: if they are on different paths/in differnt part of the same path, they should be distant
6. Optimize representations to match low-D and high-D distances.
   (a) They use MDS (predecessor to t-SNE)
      i. Matches distances well.
7. Visualize!

Question: Why use $dist_{ij}$ instead of original pairwise distances?

Answer: Experimentally, this doesn't work, no clear paths.

Monocle

1. Reduce data using ICA.
    (a) ICA yields pairwise diffenrent signals, so the inituition is they should be good to differentiate between paths (need a lot of them).
2. Compute pairwise euclidean distances as edge weights (clique).
3. Compute MST.
4. Define "backbone" (begining → end of pseudotime) as the tree diameter (longest path).

Then Monocle 2

Reduce dim with PCA.
Repeatedly cluster points and move points closer to centroids.
Map points back to high dim space (they somehow maintain reverse mapping).
Visualize.

Problem: always gives trajectories, even tough they're not there (assumes the dataset is a tree).

Monocle 3...

scGPT

Input:
cell-by-gene matrix X

word → text
gene → cell

Tokens:
each gene assigned a different token
two special tokens:
- class repr.
- pad token

Can be applied to multiple different downstream tasks.

10. Lecture10_Spatial.pdf

Paper 2

General idea:
Use (X, Y) from spatial data, along with some feature-dimension Z based on expression

SpaGCN:
Initialization:

1. Create Z dimension based on histological rgb colors,
    (a) Reflect histological similarity.
2. Compute pairwise xyz distances & pass through gaussian kernel (t-SNE like).
    (a) → Affinities.
3. Use affinities as weights in a clique graph.

Training:

1. Train GCN using the graph weights to update gene expression embeddings.

Can be used in e.g. spatial clustering.

Problem with cell-level ST

→ Very expensive.

Solution: Use spot-level + scRNA-seq to "map" scRNA-seq to space (within spots).

scDOT (Distance learning, Optimal Transport):

1. Deconvolve spots into **cell types** (using scRNA-seq as help)
    (a) How exactly? IDK. Probably first cell-annotate scRNA-seq.
2. Use optimal transport to map scRNA-seq to spots.
3. Next step: map individual cells within spot to spatial locations (using some low-level space).

11. Lecture11_singleCellIntegration.pdf

Integrating multiple modalities

Most methods can be used for batch correction too.

Problem:

A lot of methods are destructive.
Thankfully, some (expensive) methods support multiple modalities.

CITE-seq: gene expression (RNA) + protein expression + ... + ...
Multi-omics (general term, not sure about it): gene expression (RNA) + Chromatin accessbility

Transcriptomics
Epigenomics
Proteomics

Two main approaches:

- matched (barcoded cells for both modalities),
    - limited to existing technologies
- unmatched (split tissue in half, run two -omics tasks)

  - problem: different methods introduce different biases to the distribution of cells
    * assumption on analogous distr. is not entirely true

CITE-seq:

- like scRNA-seq, but there are additional poly(A)-tailed antibodies introduced which pull epitopes (protein identifiers), so then poly(A) tail binds to a bead just like mRNA from a given cell

This can help with splitting t-cells into subgroups (based on some protein expression)

Types of multi-omics:

- early
  - concat features (only matched)
- middle
  - use model to integrate features to perform specific task
- late
  - perform separate analyses, combine results (somehow, e.g. voting)

Seurat-Matched (WNN weighted nearest neighbors):

graph + graph → weighted graph

Compute **combined cell similarities** based on similarities in modality A and B:

$$\theta(c_i, c_j) = w_A(c_i)\theta_A(a_{c_i}, a_{c_i}) + w_B(c_i)\theta_B(b_{c_i}, b_{c_i}).$$

→ Weighted sum.

Output: Edge weights of a graph.

→ Can be used as input to e.g. UMAP!

How?

1. Compute kNN graphs for A and B.
   (a) k=20 or 30 (for large datasets)
2. Predict each node using its neighbors in A and in B (separately).
   (a) Note that given node has two "worlds" of edges A and B (nodes identified due to barcoding)
   (b) for example a in A has: $a_{A\to A}, a_{B\to A}$
3. Compute similarities somehow.
   (a) Gaussian kernel (t-SNE like), ratio, softmax across A and B → weights.

Seurat-unmatched (MNN mutual nearest neighbors):

CCA – Canonical correlation analysis.

Requirement:

Feature length in A and B must be the same.

E.g. for chromatin accessibility take value for promoter of a gene (for each gene)

Then, both have dimension g ($\#genes$)

Can be also used for batch correction.

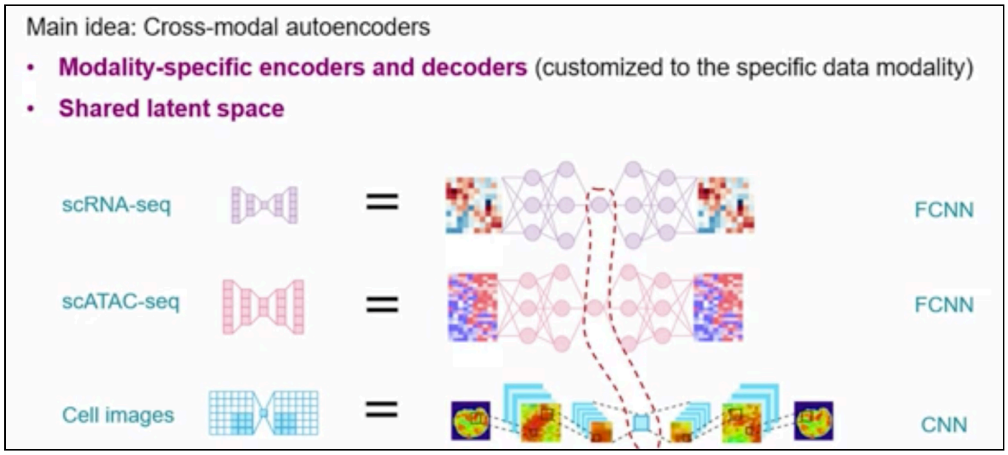
How?

1. Project with CCA to maximize correlation.
   (a) Find orthonormal projections U, V that map A and B to some latent space where entries are correlated $\rightarrow (XU)^T(YV)$ maximized.
2. Normalize results (brings alignments closer).
3. Find **anchors** by computing kNN between modalities.
   (a) MNN (mappings must be mutual)
4. Compute (somehow) compatibility score between cells and anchors.
   (a) output kinda like π in optimal transport
5. Shift modalities based on anchors and compatibilities
   (a) can be used for batch effect removal

Autoencoder-based:

Need more flexibile method if dimensions don't align.

Idea: train autoencoders to align latent spaces.



Can use latent space for downstream tasks.

How?

Use shared latent space, apply encoder and decoder to go to and from it.

""Add additional incentive to make distributions similar to the original data.""

...and more.....

Autoencoders allow imputation of unknown "?"

For example, gene expression & location of some gene!

Similarly, read count imputation works well:)

12. Lecture12_SurvivalAnalysis.pdf

- Molecular data (transcriptomics etc)
- Imaging data (H&E)
- Clinical data (age, stage, etc.)

How to predict disease stage:

Classification models

How to predict age:

Regression

How to predict survival?

Censored data (no event):

- person withdraws form the study,
- person didn't have even during entire duration of participation

Example events:

- childbirth,
- death,
- failure of a system (engineering),
- etc.

Model:

$S(t) = P(survive\ until\ t) = P(T > t)$

Kaplan-Meier curve:

- $\widehat{S}(t_i) = \widehat{S}(t_{i-1}) \cdot \left(1 - \frac{d_i}{n_i}\right)$,
  - prob of survival = 1 - prob of death
    - prob of death = d_i / n_i
- $\widehat{S}(0) = 1$.

Log rank test (finding p-value to compare different groups of patients)

- can give p-value
- .....

Cox proportial hazard model:

$S(t) = P(T > t) = 1 - F(t)$

$F'(t) = f(t)$

Hazard:

- $h(t) = \lim_{\Delta t \to 0} \frac{P(t<T\leq t+\Delta t|T>t)}{\Delta t} = \frac{f(t)}{S(t)}$
  - probability of death within the next Δt,
    - normalized, instantaneous

- $h(t|X) = h_0(t)\exp(b_1x_1 + ... + b_nx_n)$,
  - b
- $\log h(t|X) = \log h_0(t) + b_1x_1 + ... + b_nx_n$
- hazard ratio (between patients)
  - $ratio = \frac{h_0(t)\exp(x_1b)}{h_0(t)\exp(x_2b)} = \exp((x_1 - x_2)b)$

How to train?

Loss derived from likelihood of death

$$L(\beta) = \prod_{i \in Death} \frac{h(T_{death,i} \mid X_i)}{\sum_{\substack{j \in Risk \\ j:T_{event,j} > T_{death,i}}} h(T_{death,i} \mid X_j)}.$$

How to eval?

Concordance index:

$$\frac{\#concordant\ pairs = y_i < y_j \wedge \widehat{y}_i < \widehat{y}_j}{\#comparable\ pairs}$$

Pair is comparable if y_i had an event before y_j (was censored/had their event).

DeepSurv (DL method):

- replace Xb with $g(X)$ (NN)
- better c-indices

CoxTime model:

- $g(X) \to g(X, t)$
- Allows $S(t \mid X)$ to overlap.
- better c-indices

DL methods less interpretable...

We can include multi-omics data

By having multiple modality-specific encoders

And concatenating features