Advanced Topics in Bioinformatics

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2024
Daifeng Wang
daifeng.wang@wisc.edu
• More machine learning applications

• Machine learning challenges in bioinformatics

• Spatial transcriptomics

• Imaging genetics/genomics

• Artificial intelligence in drug discovery
Training and Testing

Generative vs. Discriminative models

• Generative approaches model the joint probability $p(x,y)$ for generating data

• Discriminative approaches directly model $p(y|x)$ for classification

https://medium.com/@jordi299/about-generative-and-discriminative-models-d8958b67ad32
Predicting TF binding via Generative vs. Discriminative models

Semi-supervised learning (e.g., gene finding)

- Train a model with known gene sequences
- Predict labels for many unknown sequences
- Refine the model with known and predicted sequences

Data heterogeneity

Biological regularization of machine learning for phenotype prediction and interpretation

Gene regulatory networks (GRNs)

2.5M eQTLs (expression Quantitative Trait Loci) linking 1.3M SNPs to 33k coding and non-coding genes in human brain

Wang et al., Science, 2018

Input data

Genotype + Gene expression

Functional genomics

GRNs eQTLs

Regularization $\Omega(\cdot)$

Machine learning algorithm

$f^*(\cdot)$

Predictive Model

New Knowledge

Interpretability
DeepGAMI: Deep biologically guided learning for cross-modal imputation and phenotype prediction

Chandrashekar et al., Genome Medicine, 2023
DeepGAMI architecture

• Biological Drop-Connect layer

• Auxiliary learning layer
  o To estimate latent representations of one modality from another

\[ C_{GEX}^* = f_{\theta}^{aux}(C_{SNP}) = (\alpha \times C_{SNP}) + \beta \]

\[ \mathcal{L}^{aux}(C_{GEX}, C_{SNP}) = \frac{1}{N} \sum_{i=1}^{N} (C_{GEX}^i - C_{GEX}^i)^2 \]

Chandrashekar et al., Genome Medicine, 2023
Training of DeepGAMI

- Data split into training (80%) and testing (20%)
- 5-fold Cross validation
- Balanced Accuracy and AUC scores as performance metric.
- Minimizing loss function

\[
\arg\min_\theta \left( \mathcal{L}^{pri} + \lambda \times \mathcal{L}^{aux} \right)
\]

\[
\mathcal{L}^{pri}(y, \hat{y}) = L(f^{pri}_\theta(X_{SNP}, X_{GEX}), y) = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad \text{Classification loss}
\]

\[
\mathcal{L}^{aux}(C_{GEX}, C_{SNP}) = \frac{1}{N} \sum_{i=1}^{N} \left( C^i_{GEX} - f^{aux}_\theta(C^i_{SNP}) \right)^2 \quad \text{Imputation loss}
\]
Population multi-omics data in Schizophrenia and Alzheimer's disease (AD)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Modality 1 (Genotype)</th>
<th>Modality 2 (Gene expression)</th>
<th>Tissue/Cell-type</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>PsychENCODE</td>
<td>2080 SNPs</td>
<td>126 genes</td>
<td>Bulk (Prefrontal cortex)</td>
<td></td>
</tr>
<tr>
<td>CommonMind</td>
<td>339 SNPs</td>
<td>247 genes</td>
<td>Oligodendrocytes</td>
<td>Schizophrenia</td>
</tr>
<tr>
<td></td>
<td>231 SNPs</td>
<td>552 genes</td>
<td>Microglia</td>
<td>Vs Control</td>
</tr>
<tr>
<td></td>
<td>206 SNPs</td>
<td>465 genes</td>
<td>Inhibitory neurons</td>
<td></td>
</tr>
<tr>
<td></td>
<td>198 SNPs</td>
<td>414 genes</td>
<td>Excitatory neurons</td>
<td></td>
</tr>
<tr>
<td>ROSMAP</td>
<td>273 SNPs</td>
<td>102 genes</td>
<td>Prefrontal cortex</td>
<td>Cognitive Diagnosis (No CI vs Mild CI vs CI(AD/dementia))</td>
</tr>
<tr>
<td></td>
<td>467 SNPs</td>
<td>98 genes</td>
<td>Prefrontal cortex</td>
<td>CERAD score (No AD vs AD probable vs AD definite)</td>
</tr>
<tr>
<td></td>
<td>544 SNPs</td>
<td>114 genes</td>
<td>Prefrontal cortex</td>
<td>BRAAK staging (Early vs late stage)</td>
</tr>
</tbody>
</table>
Classification of schizophrenia by SNPs and cell-type gene expression

Chandrashekar et al., Genome Medicine, 2023
Prioritization of cell-type gene regulatory networks for Schizophrenia via Integrated Gradients
Predicting cognitive impairment (CI) in AD and prioritizing gene regulatory networks

Chandrashekar et al., Genome Medicine, 2023
Missing data and imputation

• Remember Netflix Problem?

• Biological data has many missing values
  – e.g., single cell dropouts

https://krishnaswamylab.github.io/tutorial/imputation_and_netflix/
Cross-Modality Imputation by Machine Learning

- Learn joint latent space of multimodal data
- Train model to reveal cross-modality relationships
- Apply pre-trained model to impute missing modality
JAMIE: Joint Variational Autoencoders for Multimodal Imputation and Embedding

Cohen-Kalafut, Huang, Wang, Nature Machine Intelligence, 2023
JAMIE Imputation of Electrophysiological Features from Gene Expression

- Joint Autoencoders for Multimodal Imputation and Embedding (JAMIE)
  - Electrophysiological features from gene expression
  - Gene expression from chromatin accessibility
- Mouse Visual Cortex
  - 3,654 cells (6 interneuron types)
  - 1,302 genes and 34 electrophysiological features

Cohen-Kalafut, Huang, Wang, Nature Machine Intelligence, 2023
Prioritization of Genes for Electrophysiological Features Using Machine Learning (SHAP)

Spatial Transcriptomics

- Assessment of gene expression profiles and spatial organization for interrogation of complex, heterogeneous tissues

Bulk RNA-seq
- Average gene expression level, lacks cellular or spatial resolution

Single cell RNA-seq
- Identify distinct cell types and their gene expression profiles

Spatial Transcriptomics
- Visualize gene expression profiles with tissue context

Functional tissue
- Spatial organization of distinct cell types and their gene expression profiles

Genomics Visium platform

- Visium Targeted Gene Expression combines crucial spatial insights with the ease and breadth of targeted panels
- Accelerate the understanding of human health and disease with a more refined picture of the biology captured on a tissue slide

**Figure 7. Targeted Spatial Gene Expression workflow with Visium Spatial solutions.** Targeted Spatial Gene Expression enables the enrichment and analysis of a targeted set of mRNAs prepared from tissue sections. Starting with a final, barcoded 10x Genomics library, the workflow allows whole transcriptome and targeted gene expression on the same samples, while simultaneously examining morphology or co-detecting proteins.

**Figure 8. Spatially resolved targeted gene expression profiling with Visium Spatial solutions.** A human cerebellum tissue section was H&E stained and processed using the Visium Spatial Gene Expression workflow, then enriched using Targeted Gene Expression with the Human Neuroscience Panel. Shown are the H&E image (A), H&E image overlaid with total UMI counts for 36 locomotory behavior genes from the neuroscience panel (B), and H&E image overlaid with SOD1 expression level (C).

https://pages.10xgenomics.com/rs/446-PBO-704/images/10x_BR060_Inside_Visium_Spatial_Technology.pdf
Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex (DLPFC)

- Localize spatial gene expression in the human brain at cellular resolution will be critical to gain further insight into disease mechanisms

Spatial transcriptomics in DLPFC using Visium

Further applications of spatial gene expression

- Extensive layer-enriched expression signatures and refined associations to previous laminar markers

- Differential layer-enriched expression of genes associated with schizophrenia disorder (SCZD) and autism spectrum disorder (ASD), highlighting the clinical relevance of spatially defined expression signatures

- Novel cortical layer-enriched genes

Further applications of spatial gene expression in DLPFC

- A data-driven framework to define unsupervised clusters in spatial transcriptomics data, which can be applied to other tissues or brain regions in which morphological architecture is not as well defined as cortical laminae.

Supervised annotation of DLPFC layers

Schematic illustrating the data-driven clustering pipeline

Evaluation of clustering performance

Comparison of gene-wise test statistics

Expression patterns for selected laminar and nonlaminar genes
Imaging genetics

- GWAS for imaging phenotypes
  - Subcortical region volumes


UK BioBank

$n = 8,428$ subjects for 3,144 functional and structural brain imaging phenotypes
Linking genes to brain phenotypes

• IMAGEN cohort

Artificial Intelligence vs. Machine learning

• Artificial Intelligence (AI)
  – Broad concept using machines to do human-intelligence tasks
  – Visual perception, speech recognition, etc.

• Machine learning (ML)
  – An AI subarea making machines (e.g., computers) automatically learn how to finish tasks

https://commonfund.nih.gov/bridge2ai/faqs
AI/ML in drug discovery

• Previously computer-aided drug design (CADD)
  – Molecular structures
  – Low successful rate (6.2%)

• Increasing data enables AI/ML application
  – Omics
  – Imaging
  – Diagnosis
  – Behaviors
Drug discovery pipeline

Successful applications in drug discovery

- Target identification and prioritization based on gene–disease associations
- Target druggability predictions
- Identification of alternative targets (splice variants)
- Compound design with desirable properties
- Compound synthesis reaction plans
- Ligand-based compound screening
- Tissue-specific biomarker identification
- Classification of cancer drug–response signatures
- Prediction of biomarkers of clinical end points
- Determination of drug response by cellular phenotyping in oncology
- Precise measurements of the tumour microenvironment in immuno-oncology

Required data characteristics

- Current data are highly heterogeneous: need standardized high-dimensional target–disease–drug association data sets
- Comprehensive omics data from disease and normal states
- High-confidence associations from the literature
- Metadata from successful and failed clinical trials
- Large amounts of training data needed
- Models for compound reaction space and rules
- Gold standard ADME data
- Numerous protein structures
- Biomarkers: reproducibility of models based on gene expression data
- Dimension reduction of single-cell data for cell type and biomarker identification
- Proteomic and transcriptomic data of high quality and quantity
- Pathology: well-curated expert annotations for broad-use cases (cancer versus normal cells)
- Gold standard data sets to improve interpretability and transparency of models
- Sample size: high number of images per clinical trial

ML approaches for drug discovery

Coregulated gene modules predict drug candidates

- Network proximity approach
- 2,891 U.S. FDA-approved or clinically investigational drugs screened; 34 candidates prioritized

- Everolimus, an mTOR inhibitor, was top predicted drug from M1

- MTOR is directly connected with multiple key AD pathology regulators

- Rifampcian, an antibiotic, according to module M4; enriched with immune responses

- Donepezil, Sildenafil

Gupta et al., PLoS Comp. Bio., 2022
Deep learning applications in pathology

Final exam review

- Deep Learning Applications
- Learning Motif Models
- Protein Structure and Alphafold
- Genotype Analysis