Preparation for the Bioinformatics Exam

International Biology Olympiad Training

May 19, 2025

Course Overview

- Single-cell transcriptomics
- Oligodendrocyte precursor cells (OPCs)
- Transcriptomic clustering
- Gene and protein databases
- Mutation analysis
- Sequence alignments
- Practical browser-based tasks

Single-Cell Transcriptomics

- Measures gene expression in individual cells
- Reveals cell-type heterogeneity
- Common methods: Drop-seq, 10X Genomics

Why Single-Cell Data?

- Traditional RNA-seq averages expression over many cells
- Single-cell methods reveal hidden cell subtypes
- Used to study brain, cancer, development

Oligodendrocyte Precursor Cells (OPCs)

- Found in the central nervous system
- Differentiate into myelinating oligodendrocytes
- Retain proliferative potential in adulthood

Transcriptomic Clustering

- Cells grouped by gene expression patterns
- Tools: Scanpy
- Often visualized via UMAP or t-SNE

UMAP: Uniform Manifold Approximation and Projection

- A nonlinear dimensionality reduction technique
- Preserves both local and global structure of high-dimensional data
- Commonly used to visualize cell clusters in transcriptomics
- Faster than t-SNE on large datasets
- Axes are arbitrary; focus is on the relative position of clusters

t-SNE: t-Distributed Stochastic Neighbor Embedding

- Projects high-dimensional data into 2D or 3D space
- Emphasizes local relationships — similar cells appear closer
- Often used in single-cell RNA-seq to reveal cluster structure
- Sensitive to parameters (perplexity, iterations)
- Can distort global structure compared to UMAP

Marker Genes

- Specific genes highly expressed in certain cell types
- Used to identify and validate clusters
- Example: PDGFRa (OPCs), MBP (oligodendrocytes)

Gene Databases

- **NCBI Gene:** gene info, sequences
- **UniProt:** protein function, structure
- **Ensembl:** genome-wide annotations

Hands-on: Find GFAP on NCBI

- Go to `https://www.ncbi.nlm.nih.gov`
- Search for human `GFAP` gene
- Task: Identify gene ID, tissue expression, linked disorders

Protein Structure Analysis

- Tools: AlphaFold, RCSB PDB
- Visualize secondary structures
- Check for helices, sheets, ligands

Hands-on: Explore Structure

- Visit: `https://www.rcsb.org/`
- Search "GFAP" or "APP" (amyloid precursor protein)
- Task: Count helices, find model method

Protein Domains

- Functional units in proteins
- Identified via conserved sequences
- Tools: NCBI CDD, Pfam

Mutation Analysis

- Amino acid changes can affect function
- Tools: PolyPhen-2, SIFT
- Predict impact of missense mutations

Hands-on: Predict Mutation Impact

- Visit `https://www.mutationtaster.org/`
- Input: p.Arg239His in human GFAP
- Task: Interpret the output

Sequence Alignments

- Compare DNA or protein sequences
- Tools: MUSCLE, Clustal Omega
- Measure evolutionary conservation

Hands-on: MUSCLE Alignment

- Go to: `https://www.ebi.ac.uk/Tools/msa/muscle/`
- Align human and mouse GFAP or NEFL genes
- Task: Compare gene vs. protein identity

Gene Expression Patterns

- Different tissues express genes differently
- Use GTEx (human) or Mouse ENCODE
- Example: GFAP in brain, spinal cord; APP in cortex, hippocampus

Practical Tip: Use Bioinformatics Portals

- NCBI: Sequences, structures, variation
- EMBL-EBI: Alignments, functional annotation
- UCSC Genome Browser: Genomic context

Summary

- Understand transcriptomics & OPC biology
- Use online tools for data exploration
- Practice with mutation, alignment, and annotation

Q&A and Final Practice

- Find GFAP domains on NCBI CDD
- Try PolyPhen for a R239H mutation in GFAP