Student Project Proposal

GENOA: Categorical Clustering and Regulatory Network Inference for Tumor State Discovery

Master area	AI for Drug Discovery; bioinformatics; transcriptomics; network inference; precision oncology
Project type	Computational bioinformatics project using public or tutor-provided gene expression data
Duration	2-3 months
Supervision	One tutor or co-tutor; weekly or bi-weekly progress meetings
Recommended level	Python, data analysis and basic machine learning; basic knowledge of gene expression data is useful
Main output	Documented workflow/notebook, short technical report and final presentation
Possible data	One TCGA tumor cohort and, when appropriate, matched or reference healthy expression data such as GTEx

Short description. The project asks the student to implement a simplified GENOA pipeline that transforms continuous gene expression into categorical states, clusters patients in this discrete space and reconstructs cluster-specific regulatory networks.

1. Background and rationale

Tumor cohorts are highly heterogeneous. Standard clustering of continuous gene expression matrices may be sensitive to noise, technical variability and high dimensionality, and may not capture the discrete regulatory states that characterize biologically distinct tumor subgroups.

GENOA is based on the idea that gene expression can be discretized into interpretable states and that patients can be compared according to categorical activity patterns. After patient stratification, each cluster can be described by gene-gene relationships or information networks.

This project is relevant to AI for drug discovery because patient stratification, biomarker discovery and network-based interpretation are central to precision oncology and target identification.

2. Project objective

The objective of the project is to build a reproducible proof-of-concept pipeline for categorical clustering of tumor gene expression data and cluster-specific regulatory network inference.

The project should focus on one cancer type and a manageable set of genes. The goal is not to create a clinical classifier, but to demonstrate how categorical state-space analysis can reveal interpretable tumor subgroups and network reprogramming.

3. Specific aims

- Preprocess and normalize a gene expression matrix from a selected tumor cohort.
- Filter genes and convert continuous expression values into discrete categorical states.
- Cluster patients using categorical distances or recurrent gene-state patterns.
- Reconstruct a gene interaction or information network within each cluster.
- Compare clusters to identify gene pairs, hubs or network structures that differ across tumor states.
- Produce visualizations that support biological interpretation of the inferred subgroups.

4. Proposed methodology

4.1 Data preprocessing

The student will select one cancer dataset, remove genes with low information content, harmonize the matrix and apply Z-score or comparable normalization across patients.

4.2 Categorical discretization

Continuous expression values will be transformed into a small number of states, for example low, intermediate, homeostatic and high expression. The chosen thresholds should be clearly justified.

4.3 Categorical clustering

Patients will be clustered using a method suitable for discrete data, such as Hamming distance, K-modes-like clustering or a simplified recurrent gene-pair state strategy.

4.4 Network inference

Within each cluster, the student will reconstruct a network using dominant categorical co-states or mutual information between discretized gene variables.

4.5 Cluster comparison

Networks and categorical profiles will be compared across clusters to identify differential hubs, rewired gene pairs and candidate biological signatures.

5. Indicative work plan

Period	Main task	Expected output
Weeks 1-2	Dataset selection and preprocessing	Clean expression matrix and gene filtering report.
Weeks 3-4	Discretization strategy	Categorical patient-gene matrix and diagnostic plots.
Weeks 5-6	Patient clustering	Initial cluster assignments and internal validation metrics.
Weeks 7-8	Network reconstruction	Cluster-specific networks based on co-states or mutual information.
Weeks 9-10	Comparative interpretation	List of differential gene pairs, hubs and cluster-level signatures.
Weeks 11-12	Final integration and reporting	Final notebook, technical report and presentation.

6. Expected deliverables

- A documented Python notebook or code repository implementing the GENOA prototype.
- A processed categorical expression matrix for one tumor cohort.
- Patient clusters with basic validation and visualization.
- Cluster-specific gene networks and a comparison of relevant hubs or gene pairs.
- A final technical report and a 10-15 minute presentation.

7. Skills acquired by the student

- Preprocessing of transcriptomic datasets.
- Discretization of continuous biological data into interpretable states.
- Clustering with categorical representations.
- Network reconstruction from gene expression data.
- Biological interpretation of tumor heterogeneity and candidate biomarkers.

8. Suggested software and resources

The project can be implemented with pandas, NumPy, scikit-learn, scipy, NetworkX and matplotlib. Optional extensions may use scanpy, seaborn alternatives or specialized information-theoretic packages.

The tutor should define the tumor cohort and provide a small gene list or selection criterion to keep the project feasible within 2-3 months.

9. Optional extensions

Possible extensions include testing alternative discretization strategies, adding pathway-level annotation, comparing mutual information with correlation networks, or validating clusters against known clinical or molecular subtypes.

10. Evaluation criteria

Criterion	What should be assessed
Scientific clarity	The scientific question, assumptions and expected use case are clearly stated.
Technical implementation	The code is readable, documented and reproducible on the selected data.
AI relevance	The project produces representations, models or workflows that are suitable for AI-based analysis.
Interpretability	The student explains what the results mean in a drug discovery or molecular science context.
Critical discussion	The final report discusses limits, uncertainty and possible future extensions.

Note for tutors. The minimal successful version should focus on one tumor type, one discretization strategy, one clustering method and one network inference approach. Larger cross-cancer analyses should be considered optional.