Graph Mining of Multi-Omics Networks in Parkinson's Disease Across the Brain Using BioNeuralNet

Elyas Larfi
University of Colorado Denver

1. Overview and Motivation

Parkinson's Disease (PD) is a complex neurodegenerative disorder driven by multi-layered molecular disruptions spanning transcriptional, epigenetic, and proteomic levels. While much research has focused on the substantia nigra, emerging multi-omics datasets reveal widespread alterations across cortical and subcortical regions, including the motor, prefrontal, and temporal cortices. Understanding these molecular relationships across the brain requires tools capable of uncovering *network-level* patterns, rather than isolated gene lists.

This project aims to apply **BioNeuralNet**, a Python-based Graph Neural Network (GNN) framework for multi-omics analysis, to construct and analyze large-scale brain-wide molecular graphs in PD. Through **graph mining**, we will identify key communities, subgraphs, and motifs that correspond to biological pathways, regulatory mechanisms, or conserved dysregulated modules across brain regions. By linking these graph-derived motifs to known pathways (KEGG, Reactome), the project bridges raw multi-omic data with interpretable systems-level insights.

2. Data Sources and Requirements

2.1. Datasets

We will integrate publicly available, multi-omic datasets that represent different PD-affected brain regions:

- **AMP-PD (Accelerating Medicines Partnership in PD):** RNA-seq, DNA methylation, and proteomics from postmortem PD and control brains, including multiple cortical and subcortical regions.
- **Synapse PsychENCODE:** Multi-omics datasets from cortical and limbic regions, providing transcriptomic and epigenomic layers.
- **GEO Series (e.g., GSE49036, GSE49037):** Regional transcriptomic and methylomic data for validation.
- **STRING / BioGRID:** Protein-protein interaction databases used for graph prior construction.
- **KEGG / Reactome:** Biological pathway repositories for motif and module validation.

2.2. Data Scale

Each omic layer typically contains 10^4–10^5 molecular entities (genes, CpGs, proteins) with 100–300 samples per region and condition. After preprocessing and integration, the unified graph is expected to have $\sim 10^4$ nodes and $\sim 10^7$ weighted edges per brain region. Graphs will be represented in sparse tensor format, suitable for GPU computation (RTX 3090).

3. Problem Definition as a Graph Analytical Task

We define the multi-omic biological system as a multi-layered graph $G = (V, E, A)$, where:

- V: Genes, proteins, and regulatory elements across omic layers.
- E: Weighted edges representing correlations, co-expression, or prior biological interactions.
- A: Node attributes representing omic-specific features (e.g., expression, methylation, proteomic intensity).

The problem is framed as discovering the *hidden structure* of G through:

1. **Graph Construction:** Build omic-specific graphs (e.g., correlation, RBF, mutual information) and integrate them using BioNeuralNet's SmCCNet-based correlation fusion.
2. **Graph Embedding:** Learn node embeddings using GNN architectures (GCN, GAT, GraphSAGE) that capture cross-omic relationships.
3. **Graph Mining:** Identify communities, frequent motifs, and differential modules using Leiden community detection, gSpan motif discovery, and PageRank-based stability analysis.
4. **Validation:** Compare discovered motifs and modules with KEGG/Reactome pathways and test for reproducibility across cohorts and brain regions.

4. Research Roadmap

4.1. Phase 1: Data Integration and Graph Construction

- Download and preprocess multi-omic datasets (normalization, batch correction, feature alignment).
- Construct omic-specific correlation and similarity graphs.
- Integrate omics layers into unified multi-layer graphs per brain region.

4.2. Phase 2: Graph Embedding via BioNeuralNet

- Implement GCN, GAT and GraphSAGE embedding pipelines.
- Optimize embeddings via hyperparameter tuning (Ray Tune) for disease classification and feature representation.

4.3. Phase 3: Graph Mining and Motif Discovery

- Apply Leiden algorithm for community detection.
- Perform motif enumeration using gSpan and measure motif frequencies across brain regions.
- Conduct differential module analysis between PD and control networks.

4.4. Phase 4: Biological Interpretation and Validation

- Match motifs and modules to KEGG/Reactome pathways.
- Assess motif reproducibility across datasets and regions.
- Interpret conserved or region-specific molecular modules as potential PD biomarkers.

5. Methods and Tools

5.1. Software Stack

- **BioNeuralNet:** GNN-based multi-omics integration framework.
- **NetworkX / PyTorch Geometric:** For graph operations and GPU-based embedding computation.
- **Leidenalg & gSpan:** For community detection and motif discovery.
- **GSEApy / Enrichr:** Pathway enrichment and validation tools.
- **Cytoscape / Plotly:** Visualization of modules, motifs, and embeddings.

6. Expected Outcomes

- A unified graph-mining pipeline for multi-omic PD analysis implemented within BioNeuralNet.
- Identification of key molecular modules and motifs conserved across PD-affected brain regions.
- Pathway-level interpretation of discovered motifs, linking graph patterns to biological mechanisms.
- A reproducible benchmark for integrating graph mining into multi-omics neuroscience research.