Integrating Pathway Commons and cBioPortal for Patient Survival Prediction using PyTorch Geometric to Build a Graph Neural Network Model

Favour James [1] Yoshitaka Inoue [2] Augustin Luna[3]

[1]Obafemi Awolowo University [2]University of Minnesota [3]Harvard Medical School

Abstract

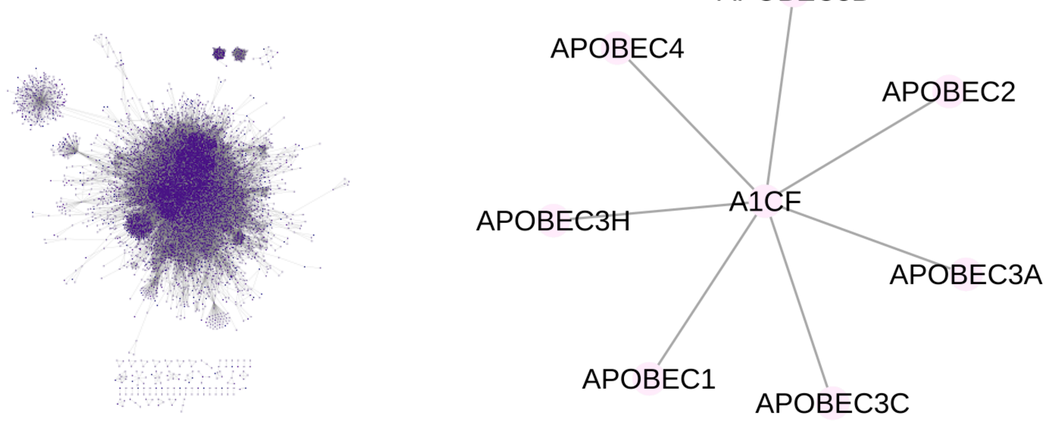
Millions of people each year die from cancer worldwide. Understanding the underlying biological mechanisms that result in some individuals outliving others can help us produce better treatment strategies. Machine learning, specially neural network-based, models can help identify otherwise hard-to-identify patterns in patient data. Graph Neural Networks (GNNs) allow the development of predictive models based on known biological network data.

To aid in the development of such models, this project aims to generate an example dataset that integrates Pathway Commons (biological network) and cBioPortal (cancer patient) data for use with the popular PyTorch Geometric GNN library for the prediction of cancer patient overall survival. We provide scripts for processing similar datasets and example code to showcase how models can be generated. We hope this work will aid researchers in unlocking valuable insights into the role of biological pathways and genetic variation in cancer to help patients.

Introduction

Graph Neural Networks (GNNs) are a recent class of deep learning methods designed to perform inference on data described by graphs. Graphs can be used to model real-world phenomena such as transportation, social, and biological networks. Graphs are heterogeneously structured whereby the number of neighbors to each node is variable (as opposed to a fixed neighborhood size of images). GNNs are used to learn node embeddings through an aggregation of information from connected neighbors of a node through a process of message passing between nodes.

In the present work, we use data from Pathway Commons, an aggregated database of millions of molecular interactions across 20 source databases. cBioPortal for Cancer Genomics is a database of multidimensional cancer genomics data collected from 200 studies. Each patient from our cBioPortal dataset are represented by a graph where the nodes are genes and the edges represent the connection between genes. Input node values are provided by genomics data from cBioPortal while the edges are obtained from the Pathway Commons (using the Reactome subset) database.



(a) The Pathway Commons graph (b) APOBEC gene subgraph of DNA repair

Figure 1. The Pathway Commons data

Research Objectives

The present study investigates the following objectives:

- **Objective 1:** Create a dataset for use with PyTorch Geometric that includes cancer patient genomics and survival data from cBioPortal and network data to describe the graph structure for the genomics data from Pathway Commons.
- **Objective 2:** Develop example models using the dataset from Objective 1.

Study methodology

The study adopted the following workflow to achieve the research objectives.

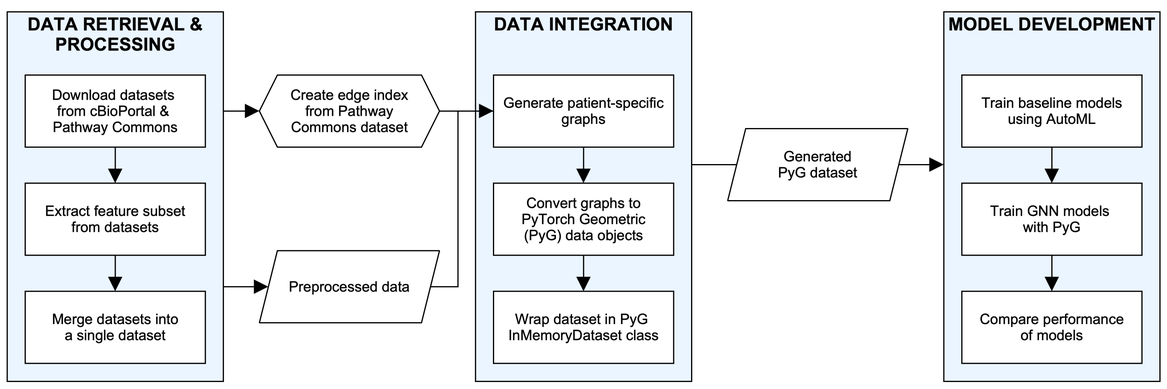


Figure 2. Workflow diagram

First, the acc_tcga_2018 (ACC: Adrenocortical Carcinoma) and brca_tcga_2018 (BRCA: Breast invasive carcinoma) datasets collected by The Cancer Genome Atlas project were selected from cBioPortal representing small (N=78) and large (N=1082) collections of cancer patient data, respectively.

Gene expression features (N=9288) were extracted that overlapped with biological network data from Pathway Commons. Additionally, overall survival time in months was extracted as the value to be predicted..

Next, an edge index (N=271771 edges) was generated using Pathway Commons v12 data in a tabular format. Integrating these two data types resulted in patient-specific graphs which were then converted into PyG data objects. These graphs are wrapped using the InMemoryDataset class for use with PyG.

Lastly, for model development, models were generated for graph-based and baseline techniques for comparison. GNN models for each cancer type (i.e., ACC and BRCA) are developed employing Graph Convolutional Networks (GCNs); model architecture shown in Figure 3. Separately, a model is created using the Automated Machine Learning (AutoML) library: FLAML Fast Library for Automated Machine Learning (FLAML).

Results and Discussion

Table 1 describes the resulting datasets used for model development after conducting data preprocessing and converting the datasets into PyG data objects:

Table 1. Dataset Statistics

Dataset Name	# Graphs	# Nodes	# Edges	# Features
acc_tcga_2018	78	9288	271771	78
brca_tcga_2018	1082	9288	271771	1082

In order to ensure reproducible notebooks, the final processed PyG datasets, as well as unprocessed data, were uploaded to Zenodo, a general-purpose data repository meant for long-term storage of academic datasets.

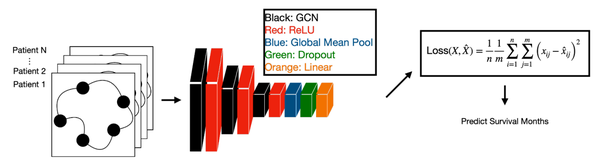
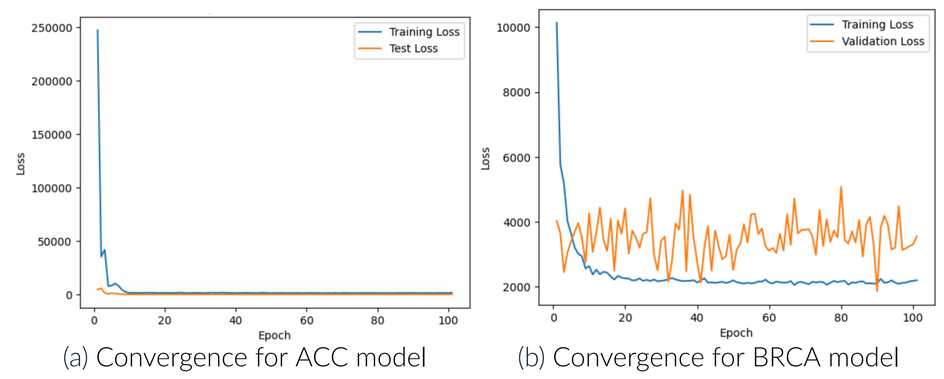


Figure 3. Graph Neural Network Overview

Cancer Type	Sample Count	Model	MSE	Train,Test,Validation	Train Time(s)
Adrenocortical carcinoma (ACC)	78	FLAML	694.12	70%; 30%	301.3
Adrenocortical carcinoma (ACC)	78	GCNs	665.97	70%; 30%; 0%	186.3
Breast invasive carcinoma (BRCA)	1082	FLAML	1182.29	60%; 40%	303.6
Breast invasive carcinoma (BRCA)	1082	GCNs	957.83	60%; 20% 20%	2644.0

Table 2. Modelling Statistics

Preliminary results in Table 2 demonstrate that the GNN model performs similarly to the FLAML-based model for the smaller ACC model in predicting overall survival time. By contrast, the BRCA GNN model outperforms the BRCA FLAML model. We believe this observation between the models is based on the larger BRCA dataset. Note that given the small size of the ACC data, it was split to only train and test sets. However, for the BRCA data, a validation split was included. During the development of the baseline models, only train and test splits were used. This was done to fit with the requirements of FLAML.



(a) Convergence for ACC model (b) Convergence for BRCA model

Figure 4. Training/Validation Loss Graphs

Figure 4 shows train and validation losses for the GNN model. Both training and validation losses converge quickly for the ACC model, while for the BRCA model, the training loss converges while the validation loss does not. The lack of convergence in the BRCA validation loss graph suggests the model is overfitting; this requires further effort to address.

Conclusions/Future Work

GNNs have only in the recent few years been applied to the challenge of understanding the survival of cancer patients. To aid in this area of research, we have produced example material for dataset generation and analysis utilizing large, public collections of data using both GNN and AutoML techniques.

The preliminary models we have produced help us begin to understand issues of necessary sample size and overfitting to be addressed with future work. We believe GNNs can help researchers to understand complex relationships observed in cancer biology and be of use in personalized medicine research. Further exploration is also planned with Graph Attention Networks (GATs) techniques to aid in model interpretability.

Data Links

- BRCA Dataset Link: `https://zenodo.org/record/8212008`

References

[pyg] Pytorch geometric. `https://pytorch-geometric.readthedocs.io`. Accessed: Summer 2023.

[pyt] Pytorch. `https://pytorch.org`. Accessed: Summer 2023.

[cbi] cbioportal. `https://www.cbioportal.org`. Accessed: Summer 2023.

[pc] Pathway commons. `https://www.pathwaycommons.org/`. Accessed: Summer 2023.

Acknowledgements

We would like to thank the Google Summer of Code program for funding and the National Resource for Network Biology (nrnb.org) for overall coordination.

Project Repository

Scan QR code