Erdős Deep Learning Bootcamp Summer 2024 Executive Summary

Team AntiBERTotics: Scott Auerbach, Craig Corsi, Hatice Mutlu, Samuel Ogunfuye

Github: https://github.com/auerbachsd/antiBERTotics

Overview: Given the rise in bacterial pathogens that are resistant to current antibiotics due to misuse, this has the potential to escalate into a health catastrophe. The main idea is to use optimized large language models intended for small-molecule drugs as well as those for parsing DNA and other genetic information to construct a model based on structural correlations that can predict whether or not known pathogens are resistant to a given antibiotic.

KPI (Key Performance Indicator):

- Accuracy = (# True Positives + # True Negatives) / (# Predictions)
- F1-score = 2*Precision*Recall / (Precision + Recall)
  Note: Precision = # True Positives / (# True Positives + # False Positives), and
  Recall = # True Positives / (# True Positives + # False Negatives)

Stakeholders:

- Pre-clinical genetics research teams
- Clinical research teams
- Pharmaceutical companies and medical centers (and their clients)

Data:

- All data studied is part of NIH's National Library of Medicine. This includes DNA sequence data from the National Center for Technology Information's Pathogen Detection project, accessed using MicroBIGG-E (Microbial Browser for Genetic and Genomic Elements). We also use antibiotic structural data (SMILES format) from PubChem.
- There were 32,000,000 different genes spread across various pathogens. Due to the sheer size of the data, the web download limit is 100,000 rows (genes). However, the full dataset can be accessed from the NCBI Entrez search tool with the contiguous accession number for each gene using the *biopython* package.
- We used a similar method using the *requests* package to add the SMILES data for the antibiotics.
- During both of these processes, some of the DNA sequences and SMILES information were irretrievable for whatever reason, so the actual training dataset was somewhat reduced.
- We focused on two common gut pathogens for this project: *Escherichia coli* and *Salmonella enterica*.
- The data's features include the bacterial species, start and end positions for each sequence, and the gene type (AMR, virulence, or stress).
- Salmonella data size: (48576, 19). We shuffled it and sampled. It has STRESS:0.487154, AMR:0.286726, VIRULENCE: 0.22612

Models:

- *Binary Classifier for AMR Genes:* Initially, we focused on developing a quick binary classifier to identify AMR genes versus non-AMR (for example, virulence or bacterial stress) - one-hot encoding was used to further differentiate the gene types (AMR with a value of 1 with everything else being 0). We used a pre-trained BERT (Bidirectional Encoded Representations of Transformers) model from Hugging Face, DNA-BERT-6. Due to RAM issues on Google Colab, we had to only use some of the dataset because of the large length of the sequences (>1000 nucleotides). Once the sequences were tokenized, we created a class called SimpleDataset to convert it back to a format suitable for PyTorch's DataLoader package to test the DNABERT model.

- *Identifying AMR based on SMILES and sequence data:* This approach used DNABERT, but also another pre-trained BERT model on top of that from DeepChem, ChemBERTa-77M. Rather than one-hot encoding for the type of genes, we used the LabelEncoder method from sklearn.preprocessing. Another distinction is that we used the logits, or classification scores for embedded SMILES since there were only ~80 different antimicrobial molecules in the dataset shared across genes rather than the raw embeddings themselves. For the model, the encoded label tensors, sequence embeddings, and SMILES logits were concatenated and processed with a class ResistancePredictionModel, which processes this concatenated data through two fully connected layers with a sigmoid activation to estimate the probability of a DNA sequence coding for a gene conferring resistance to a given antibiotic given its SMILES information

Results:

For the binary AMR classifier using the pretrained DNABERT model, we got these results for the following species:

- *E. coli*: 54.6% accuracy, F1: 0.39 (4,500 samples)
- *Salmonella enterica*: 79.3% accuracy, F1: 0.70 (1,500 samples)

Pivoting towards incorporating SMILES information to predict antibiotic resistance, we achieved the following results:

- *E. coli*: 53.8% accuracy, F1: 0.21
- *Salmonella enterica*: 81.4% accuracy, F1: 0.49

Further Directions:

We would like to develop an application that can take the SMILES input of a given antibiotic and then predict the likelihood of generating resistance for multiple common microbes (including but not limited to *E. coli*, *Salmonella*, *Listeria*). However, before doing that, we need to refine our hyper-parameters further and increase our accuracy statistics, and this may be done through acquiring more of the original Microbigge dataset on Google Cloud or finding an alternative, more comprehensive dataset.