AntiBERTotics

Scott Auerbach, Craig Corsi, Hatice Mutlu, Samuel Ogunfuye

The Erdős Institute
Deep Learning Boot Camp

Goals

- Construct a model based on structural correlations that can predict whether or not known pathogens are resistant to an antibiotic

- Combine optimized large language models intended for small-molecule drugs with models that parse DNA and other genetic information

KPIs

- Accuracy = (# True Positives + # True Negatives) / (# Predictions)
- F1 Score = 2 * Precision * Recall / (Precision + Recall)

  Precision = # True Positives / (# True Positives + # False Positives)

  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

- Public data from NIH’s National Library of Medicine
  - Genetic sequence data from NCBI (National Center for Biotechnology Information)'s Pathogen Detection project
    - Ex. atcgatctcgacatatacatatacca
  - Antibiotic structural data (SMILES) from PubChem
    - Ex. C1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)CC
- Three types of DNA sequences:
  - AMR (antimicrobial resistance)
  - VIRULENCE
  - STRESS
- Our focus for this project: *Escherichia coli* and *Salmonella enterica*

Data Preprocessing

- Accessed sequence data through MicroBIGG-E (Microbial Browser for Genetic and Genomic Elements)
  - Fetched full sequence data for each start/stop using Bio.Entrez
- Augmented sequence data with k-mer shifts and random mutations
- Mapped ‘Class’ labels in each row to SMILES
- Small random sample due to computational limits

Models and Training

- Initial model: classify sequences of type AMR (DNABERT)
  - Pretrained BERT model (Bidirectional Encoded Representations of Transformers)
  - Labels encoded with One Hot Encoding
  - DNA sequences embedded using DNABERT's tokenizer

- Expanded model (DNABERT and ChemBERTa)
  - Sequence embeddings concatenated with SMILES and encoded label tensors
  - SMILES embedded using ChemBERTa, with raw embeddings converted to logits
  - Labels encoded using sklearn's LabelEncoder
  - Two fully connected layers with sigmoid activation

Results

	Accuracy (DNABERT)	F1 Score (DNABERT)	Accuracy (DNABERT + ChemBERTa)	F1 Score (DNABERT + ChemBERTa)
E. coli (4,500 test samples)	54.6%	0.39	53.8%	0.21
Salmonella enterica (1,500 test samples)	79.3%	0.70	81.4%	0.49

Future Directions

- Refine hyper-parameters and increase accuracy statistics
- Resolve RAM limitations and train on a greater subset of the data
- 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*)

Acknowledgements

We would like to thank everyone at the Erdős Institute for this opportunity.

Special thanks to our instructor, Lindsay Warrenburg, and our TA, Marcos Ortiz. THANK YOU!!!