Antimicrobial Resistance Prediction Using Deep Convolutional Neural Networks on Whole Genome Sequencing Data

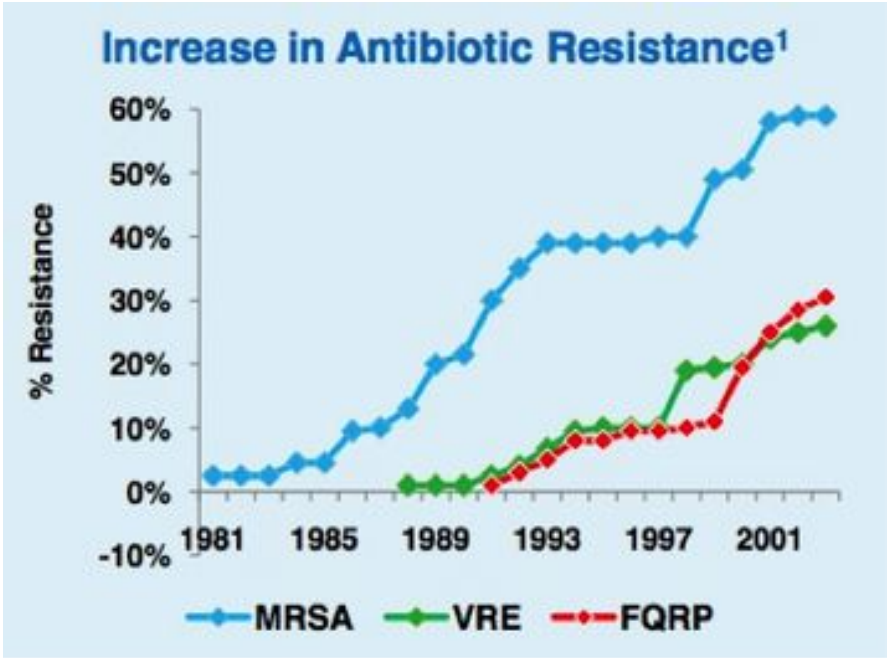
PRIMES Research Report by: Andrew Zhang
Mentor: Dr. Gil Alterovitz
PRIMES Conference October 13, 2018

Why Research on AMR?

- AMR-bacteria evolve to resist antibiotics
- Tens of thousands of deaths each year in the US alone
- Diagnosis time–two days to even a month
- Current practice–broad-spectrum antibiotic therapy, which contributes to AMR
- 10 million deaths each year from AMR bacteria over the world by 2050

Conclusion: Fast diagnosis of AMR in clinics is urgently needed to save patient lives and prevent the spread of AMR.

Some examples of AMR Increases



Methicillin-resistant *Staphylococcus aureus*, Vancomycin-resistant *Enterococci*, Fluoroquinolone-resistant *Pseudomonas aeruginosa*

Why use Machine Learning for AMR Prediction?

- ML success in many fields
  - E.g. image recognition, drug discovery
- AMR prediction with ML in literature with encouraging results
  - K-mer, Logistic Regression
  - However advanced neural networks have not yet applied
- Abundant AMR data in public databases for ML to do training and verification
- Sequencing faster and cheaper:
  - 5 minutes to sequence a bacterial genome with modern sequencing machines
  - Sequence data is more accessible now to make AMR prediction

Success of ML in Image Classification

ML is shown to be able find features in images to classify them. It is likely that this ability to classify can be applied successfully to AMR prediction

Modern Sequencing Machine

MinION

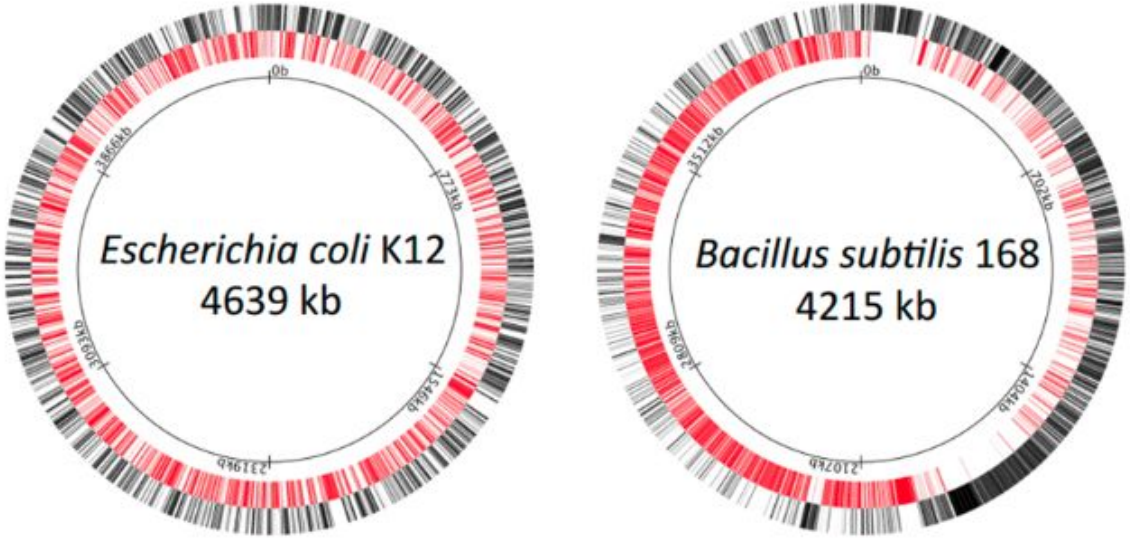
- Pocket-sized, portable device for biological analysis
- Up to 512 nanopore channels
- Simple 10-min sample prep available
- Real-time analysis for rapid, efficient workflows
- Adaptable to direct DNA or RNA sequencing

With powerful and portable sequencing machine that costs under $1000, WGS is more available now

AMR Data Acquisition

- Data needed: phenotype data (AMR data) and bacterial Whole Genome Sequence (WGS) data
- AMR phenotypes retrieved from PATRIC database.
  - PATRIC provides a large scale integration platform for researchers to share their data
  - The accumulated data provides a large repository of data for ML to use
- Corresponding WGS is retrieved from NCBI, in FASTA format
  - The FASTA files are text files representing the nucleotide sequences of a strain
  - NCBI also provides WGS of strains of the same bacteria that are not resistant

Bacterial genome structure



A few million base pairs (human has about 3 billion base pairs);

Closed circle structure (human genome has a linear structure)

Sequence Data in FASTA Format

- FASTA format file starts with a single-line description, followed by sequence data.
- The description line is distinguished from the sequence data by a ">"
- Typical bacterial FASTA file–several MBytes, few million base pairs

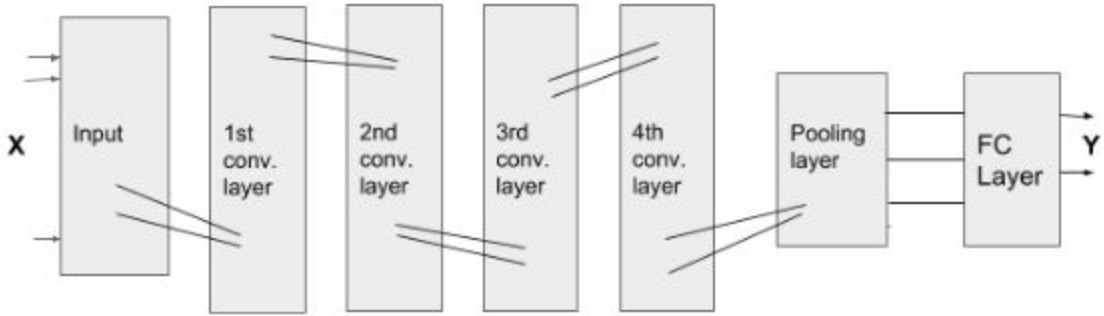
FASTA example:
>NC_016845.1 Klebsiella pneumoniae subsp. pneumoniae HS11286 chromosome, complete genome
GGTGGTCTGCCTCGCATAAAGCGGTATGAAAATGGATTGAAGCCCGGGCCG
TGGATTCTACTCAACTTTCGTCTTTCGA

Convert FASTA to Genomic Image for ML

- FASTA data–text with descriptions
- Requires processing for ML use
- Parser-remove descriptions, extract bases
- Bases are encoded to numbers
  - A flexible encoder converts bases to numbers so that the size of the resulting array is always one million
- Folded/reshaped to an 1000x1000 array
  - This is what we call “genomic image”, similar in format to a traditional image represented with pixels
  - It retains all the genomic info contained in FASTA format

Deep CNN Model for AMR Prediction

4 Conv layers, one Pooling Layer, and one Fully Connected layer



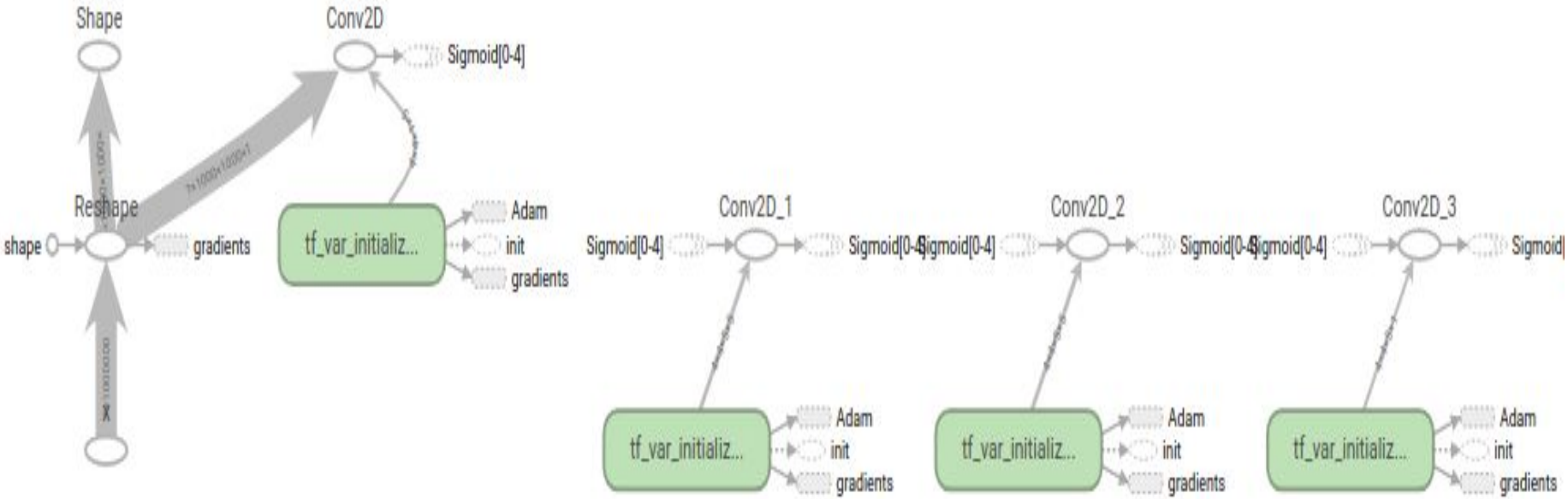
Explanation of Deep CNN Model in this Project

- Input is genomic image
- Output label – use one hot encoding, (0, 1) for resistant, and (1, 0) for non-resistant
- In the 4 convolutional layers–filter size 4x4
- 5 filters per layer
- At the last convolutional layer, a large pooling layer is used to reduce the size before a fully connected layer.
- The Model is decided after many trials and adjustments.

Implementation and Training the Model

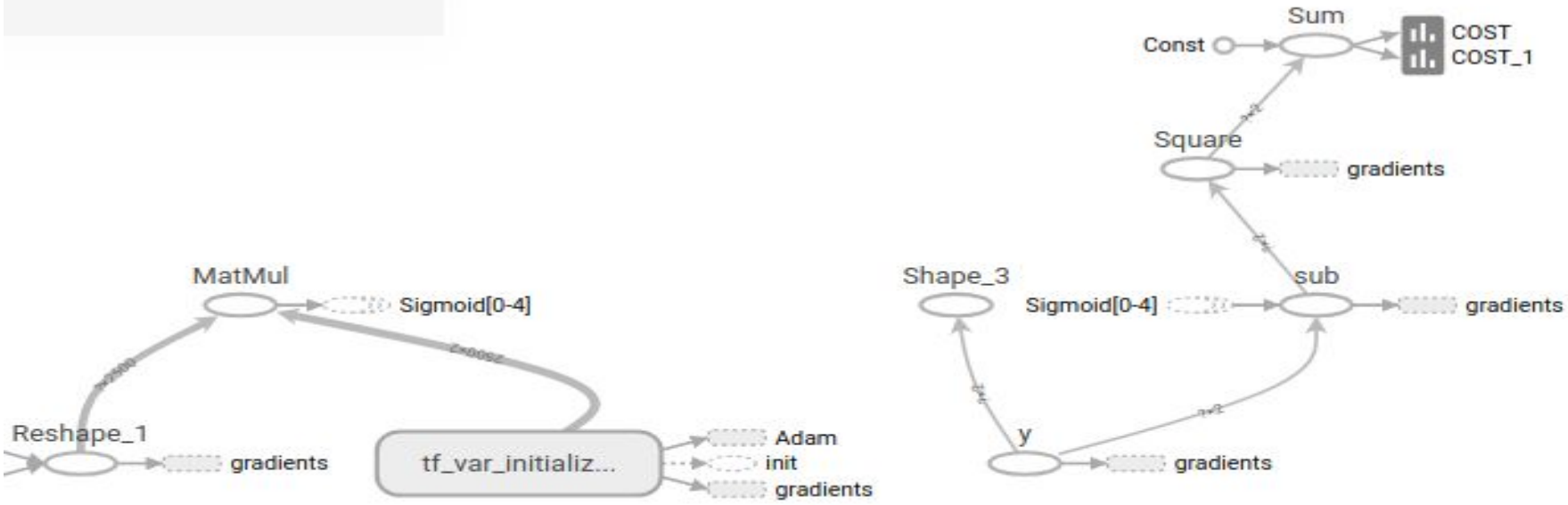
- FASTA parser implemented with Python
- Deep CNN is implemented with Python and Tensorflow, on a PC with Nvidia GPU (1080 TI, with 8G of memory)
- Data sets from two Bacteria and Antibiotics were used for training:
  - *Acinetobacter Baumanii* resistance to Carbapanems.
  - *Klebsiella Pneumoniae* resistance to Ampicillin.
  - Same Deep CNN model is used for both data sets
- Both training data sets reach 100% accuracy in about 25 minutes

Tensorboard Graph Showing Conv Layers



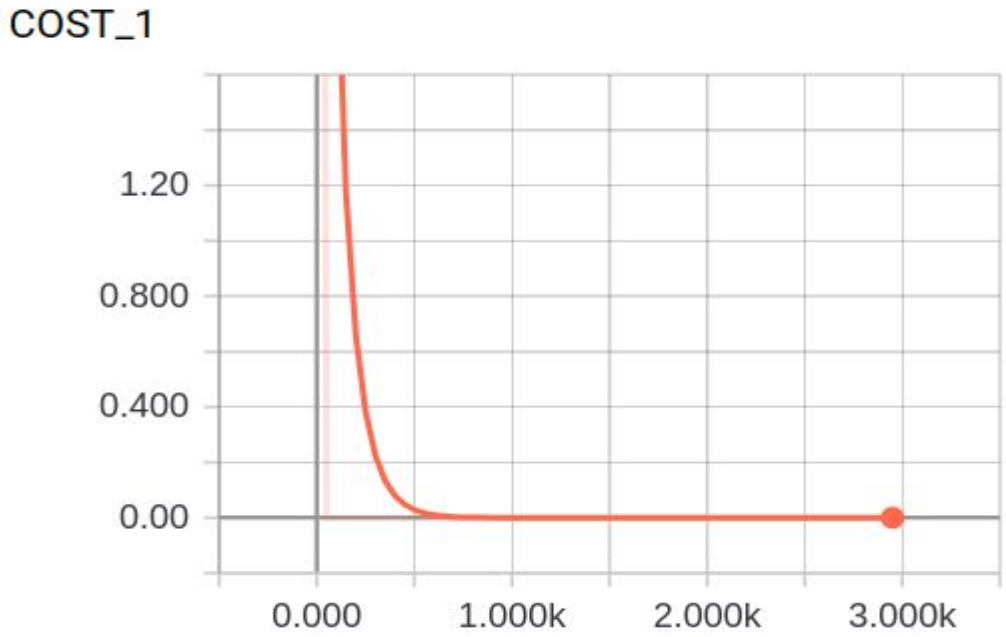
Convolutional Layer

Tensorboard Graph Showing FC Layer and Cost



Fully connected layer and cost function

TB Scalar Showing Cost reduction with Epoch



Cost function with respect to Epoch for *A. Baumannii* resistance to Carbapenem

Model Verification

- Data are divided to training/verification sets with 70/30 ratio
- Verification data are not used in training
- Verification accuracy reached 95% average.
- Prediction time is less than 1 second

Verification: *A. Baumannii* resistance to Carbapenem

	GCA 2016955	GCA 2017355	GCA 2017595	GCA 2017605	GCA 177695	GCA 177735	GCA 186565	GCA 187205
AMR based on PATRIC	Yes	Yes	Yes	Yes	No	No	No	No
AMR based on Model	Yes	Yes	Yes	Yes	No	No	No	No
Prediction Confidence	51.5%	91.8%	89.9%	99.9%	99%	97.6%	99.7%	85.8%
Prediction time	26 ms	26 ms	26 ms	26 ms	26 ms	26 ms	26 ms	26 ms

Verification: *K. Pneumoniae* resistance to Ampicillin

	GCA 900093365	GCA 000016305	GCA 900093185	GCA 900093305	GCA 900093325	GCA 000281595	GCA 000281335	GCA 000281515	GCA 000281475	GCA 000281455
AMR based on Patric	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No
AMR based on Model	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No
Prediction Confidence	96.9%	74.0%	100%	100%	99.9%	50.3%	96%	100%	100%	85.7%
Prediction time	26 ms	26 ms	26 ms	26 ms	26 ms	26 ms	26 ms	26 ms	26ms	26ms

Comparison with Literature

- Using Random Forest (RF), Santerre et al. achieved prediction accuracy of 92% on a high end machine with 32 cores and 1TB of memory
    - Though computer time is not presented in the paper, RF takes long time to compute for such large amount of data
- Using Logistic Regression (LR), Pesesky et al. achieved prediction accuracy ranging from 57.7% to 94.9% depending on the Resistance Gene Database (RGDB) being used.
    - RGDB is used to first identify resistance genes from genomic data
    - Also requires an expert user to annotate WGS data first, takes a long time
    - Different DB produced different prediction accuracy
- Deep CNN model I developed reaches verification accuracy of 95% average.
    - No need for user to use any DB to annotate Genome data, just plug in WGS data
    - Prediction takes less than a second
    - In total, less than half an hour to go from getting bacterial sample to prediction

Conclusion

- AMR is world health crisis, causing tens of thousands of deaths in US alone each year
- Quickly identifying AMR is important to ensure good patient outcomes, and to prevent the spread of AMR and development of more AMR bacterial strains
- Developed a Deep CNN model that can predict AMR under a second, with an accuracy of 95%.
    - Used with the sequencing machine on the market today that can produce sequence data of a bacteria in 5 minutes with 10 minute prep time, the model can make AMR predictions over the course of a patient’s visit.

References

1. CDC, (2018), https://www.cdc.gov/drugresistance/index.html, Centers for Disease Control and Prevention
2. Didelot X., Bowden R., et al, (2012), Transforming clinical microbiology with bacterial genome sequencing, PMC, available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5049685/.
3. Nanopore Technologies, (2018), https://nanoporetech.com/products
4. Peseky, M., Hussain, T., et al, (2016), Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome Sequence Data, Frontiers in Microbiology
5. Rafael Gómez-Bombarelli, et al."Automatic chemical design using a data-driven continuous representation of molecules", arXiv, Cornell University Library, 2017, https://arxiv.org/pdf/1610.02415.pdf
6. Santerre, J. W., Davis, J. J., Xia, F. & Stevens, R. Machine learning for antimicrobial resistance. arXiv preprint arXiv:1607.01224 (2016)
7. Gillespie, J., Wattam, R., et al. Patric: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infection and immunity, 79(11): 4286–4298, 2011.
8. CIFAR10, https://en.wikipedia.org/wiki/CIFAR-10, 2018
9. Fishbain J., Peleg A, Treatment of *Acinetobacter* Infections, *Clinical Infectious Diseases*, Volume 51, Issue 1, 1 July 2010.
10. Everyday Health, https://www.everydayhealth.com/klebsiella-pneumoniae/guide/, 2018
11. The Economist, "Antibiotic use is rapidly increasing in developing countries", https://www.economist.com/graphic-detail/2018/04/02/antibiotic-use-is-rapidly-increasing-in-developing-countries, 2018

Acknowledgement

Many thanks to Dr. Gil Alterovitz and Dr. Fan Lin for their guidance on this project, and to the PRIMES program for giving me this opportunity.

Questions and Answers