Identification of antibiotic resistance in tuberculosis with whole genome sequencing and neural networks

Michael L. Chen[1,2], Chang Ho Yoon[1], Anna G. Green[1], Isaac S. Kohane[1], Andrew Beam[1,3*], and Maha Farhat[1,4*]

[1]Department of Biomedical Informatics, Harvard Medical School, [2]Stanford University School of Medicine, [3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, [4]Division of Pulmonary & Critical Care, Massachusetts General Hospital. *Equal contribution.

Introduction

- Tuberculosis (TB) is a global health threat with an estimated 558,000 new cases of multidrug and rifampicin resistant TB in 2017 *(1)*.
- 29% of cases were detected and reported, indicating the need for timely and accessible diagnostic tools *(1)*.
- Current molecular tests probe few genetic loci *(2)*.
- Advances in whole genome sequencing show promise as a conduit for quick antibiotic phenotyping.

Objectives

Building upon our prior work using a multitask wide and deep neural network (WDNN) for predicting resistance to 11 anti-tuberculosis drugs *(3)*, we propose a new convolutional model on a larger set of WGS isolates.

Methods

- Previously, we trained the WDNN on a public data set of 3,601 MTB strains that underwent targeted or whole genome sequencing and conventional drug resistance phenotyping.
- We propose a convolutional neural network (CNN) approach on a larger set of 10,198 isolates.
- The CNN directly analyzes the genetic input sequences and aims to capture genetic motifs that contribute to antibiotic resistance.
- We trained the CNN on 1-2 specific genes and flanking regions previously known to be important to resistance for each of four drugs (rifampicin, isoniazid, pyrazinamide, and ethambutol).
- We used a gradient-based saliency map to visualize the most important genetic regions for resistance.

Results

- Our prior work demonstrated that the multitask WDNN and regularized logistic regression classifiers have the highest performance.
- On the expanded set of 10,198 isolates, the WDNN demonstrates high performance: AUC of 0.959 for first-line drugs and 0.911 for second-line drugs.
- The CNN shows promising but slightly lower predictive performance than the WDNN for the four drugs tested.
- A visualization of the important regions of the genes with respect to the predictions of the CNN are shown in a saliency map. We see well-defined regions for resistance in all drugs but pyrazinamide.

Conclusions

- We show deep learning algorithms have high predictive performance across 11 drugs.
- We believe that the high performance of the CNN in a limited genomic scope indicates promise for future performance gains with incorporation of more genomic regions.

	AUC (95% Confidence Interval)											
	1st line Drugs					**2nd Line Drugs**						
Algorithm	RIF	INH	PZA	EMB	*Average*	STR	CAP	AMK	MOXI	OFLX	KAN	*Average*
Logistic Regression	0.994 (0.993 - 0.995)	0.989 (0.987 - 0.991)	0.959 (0.955 - 0.963)	0.977 (0.975 - 0.979)	*0.980 (0.975 - 0.984)*	0.939 (0.934 - 0.943)	0.953 (0.948 - 0.958)	0.944 (0.933 - 0.954)	0.905 (0.895 - 0.915)	0.921 (0.902 - 0.941)	0.91 (0.901 - 0.919)	*0.928 (0.916 - 0.941)*
Random Forest	0.986 (0.985 - 0.988)	0.982 (0.98 - 0.985)	0.954 (0.949 - 0.958)	0.966 (0.964 - 0.969)	*0.972 (0.967 - 0.977)*	0.924 (0.92 - 0.929)	0.966 (0.962 - 0.97)	0.962 (0.956 - 0.969)	0.921 (0.914 - 0.929)	0.93 (0.914 - 0.946)	0.92 (0.911 - 0.928)	*0.937 (0.927 - 0.948)*
Deep MLP	0.994 (0.993 - 0.995)	0.988 (0.987 - 0.99)	0.96 (0.957 - 0.964)	0.973 (0.97 - 0.975)	*0.979 (0.975 - 0.983)*	0.934 (0.929 - 0.938)	0.962 (0.956 - 0.967)	0.953 (0.944 - 0.963)	0.914 (0.904 - 0.924)	0.935 (0.924 - 0.946)	0.909 (0.898 - 0.92)	*0.934 (0.924 - 0.945)*
SD-WDNN	0.994 (0.993 - 0.995)	0.987 (0.985 - 0.989)	0.959 (0.955 - 0.963)	0.971 (0.968 - 0.973)	*0.978 (0.973 - 0.982)*	0.936 (0.932 - 0.941)	0.962 (0.958 - 0.966)	0.944 (0.934 - 0.954)	0.909 (0.9 - 0.918)	0.918 (0.902 - 0.933)	0.896 (0.886 - 0.907)	*0.928 (0.916 - 0.939)*
MD-WDNN	0.994 (0.994 - 0.995)	0.988 (0.987 - 0.99)	0.961 (0.958 - 0.964)	0.973 (0.971 - 0.975)	*0.979 (0.975 - 0.983)*	0.935 (0.93 - 0.94)	0.963 (0.958 - 0.968)	0.952 (0.943 - 0.962)	0.914 (0.905 - 0.924)	0.941 (0.931 - 0.952)	0.913 (0.904 - 0.923)	*0.937 (0.926 - 0.947)*

Table 1. Tuberculosis drug resistance prediction AUROC performance of the models examined using cross-validation on the set of 3,601 isolates. The cells are colored by rank of the model for each drug.

	AUC (95% Confidence Interval)											
	1st line Drugs					**2nd Line Drugs**						
Algorithm	RIF	INH	PZA	EMB	*Average*	STR	CAP	AMK	MOXI	OFLX	KAN	*Average*
MD-WDNN	0.979	0.974	0.948	0.937	*0.959*	0.936	0.889	0.906	0.888	0.906	0.938	*0.911*
CNN	0.970		0.927		*0.948*		0.850				0.924	*0.887*

Table 2. Tuberculosis drug resistance prediction AUROC performance of the models examined using cross-validation on the set of 10,198 isolates. The cells are colored by rank of the model for each drug.

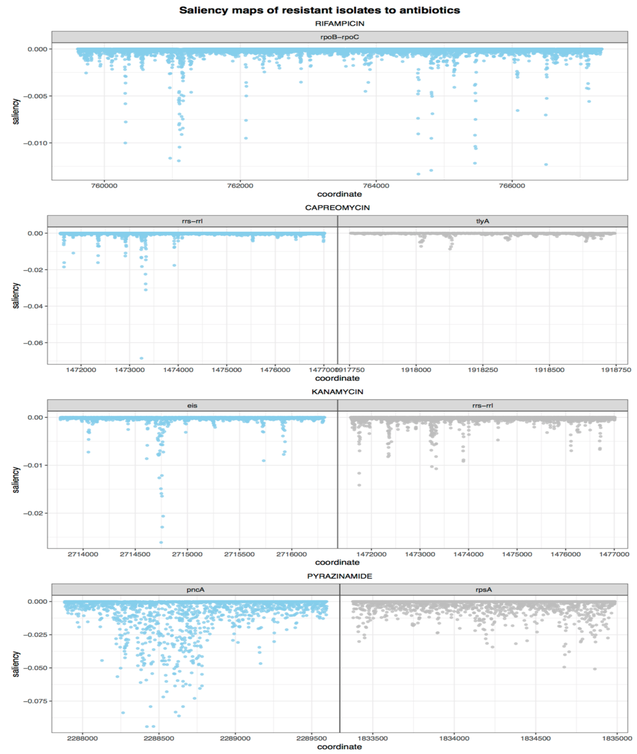

Figure 1. Relative importance of genomic coordinates to resistance phenotype within each drug for the CNN.

References

1. WHO. Global Tuberculosis Report 2018.
2. CDC. Report of Expert Consultations on Rapid Molecular Testing to Detect Drug-Resistant Tuberculosis in the United States. 2009.
3. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, Kohane IS, Beam A, Farhat M. "Beyond multidrug resistance: leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction." EBioMedicine. 2019;43:356-369.