Survey of Public Assay Data: Opportunities and Challenges to Understanding Antimicrobial Resistance

Akshay Agarwal  
IBM Almaden Research Center  
San Jose, California, USA  
Akshay.Agarwal1@ibm.com

Gowri Nayar  
IBM Almaden Research Center  
San Jose, California, USA  
Gowri.Nayar@ibm.com

James Kaufman  
IBM Almaden Research Center  
San Jose, California, USA  
jhkaufl@us.ibm.com

ABSTRACT

Learning methods allow researchers to make predictions, draw inferences, and automate generation of mathematical models. These models are crucial to solving real world problems including, for example, antimicrobial resistance (AR). Machine learning (ML) methods depend upon ground truth data to achieve specificity and sensitivity. Since AR assay data is limited, it is of paramount importance to understand the distribution of ground truth data, the analyses it is suited for, and any limitations that bias downstream methods. In this paper, we report an analysis of bacterial biochemical assay data associated with whole genome sequencing (WGS) and discuss important implications when making use of assay data in combination with genetic features as training data for ML models.

CSC CONCEPTS

• Applied computing → Computational biology; Computational genomics; Bioinformatics;

KEYWORDS

Machine Learning, MIC prediction, Assay, Biosample

ACM Reference Format:

Akshay Agarwal, Gowri Nayar, and James Kaufman. 2020. Survey of Public Assay Data: Opportunities and Challenges to Understanding Antimicrobial Resistance. In The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic. ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3341105.3374087

1 INTRODUCTION

Introduction of ML and other computational techniques to biology has revolutionized research in these disciplines[1]. Due to ML, researchers are now able to leverage the power of data to identify patterns that can help solve important problems such as AR [2].

Achieving reliable predictions requires a considerable amount of curated ground truth data [3]. Collectively, AR genes are described as part of a “resistome” [4]. A number of issues confound the ability to collect reliable data for AR prediction. There are multiple mechanisms of resistance to antimicrobial compounds and these mechanisms vary by compound. Observation of a particular resistome, while correlated with resistance to a compound, is not always a predictor of resistance for that compound[5]. Furthermore the phenotypes, Resistant(R) or Susceptible(S), are not boolean properties. Instead, AR should be understood as a stress response that results in non-discrete phenotypes. Antimicrobial compounds subject microorganisms to a stress, the nature and magnitude of which depends on the dosage and compound. Microorganisms respond to this stress by mechanisms which include, but are not limited to: enzymatic deactivation of the drug, alteration of the drug target or protein binding site, changes to metabolic pathways, reduced cell wall permeability to reduce absorption of the compound, or activation of efflux to pump the drug out of the cell[6]. Current approaches for resistance prediction perform phenotype classification using gene counting. For this above approach to be successful, the list of known R genes must be continuously updated and understanding co-occurrence requirements may be neglected. As new R genes are discovered, the above approach may not scale.

AR response varies based on an organism’s particular resistome. As a part of stress response, microbes can acquire new genes through transfer of plasmids, integrative conjugative elements (ICE), or other horizontal gene transfer (HGT) mechanisms [7]. As new genes are acquired, the organism may survive at increased concentration of the antimicrobial compound. So the organism’s ability to ‘resist’ treatment is a function of concentration. A clinician prescribes antimicrobial medication based on an optimal safe dosage. Clinical labs use several tests to measure resistance. One widely used test provides an estimate of the minimum inhibitory concentration or MIC[8]. Data from tests like MIC are used to select secondary treatments for problematic infections. Any ML or AI approach to predicting resistance should, therefore, focus on predicting MIC. The phenotype of an organism’s response to antimicrobial treatment can’t be decoupled from the prediction of MIC.

In this paper, we describe the challenges and opportunities of using biological assay data available in public repositories (e.g., NCBI’s BioSample) [9, 10]. We describe an approach to clean assay data and to identify and correct conflicts for identifying ground truth MIC.

2 DATA

From NCBI BioSample, we got bacterial assay data, resulting in 78000 assays. XML headers were mined to extract compound id, SRA id, Biosample accession, phenotype, measurement type, sign, units, typing method, platform, vendor, version, and testing standard.

This raw data, referred to as Data Stage 1, included 99 antibiotics associated with 4962 SRA_IDs (each SRA_ID indicates a bacterial WGS) and 5173 BioSample accessions (Table 1). “Phenotype description”(PD) relays the observable physical trait resulting from the
antibiotic assay which is comprised of 6 potential values: susceptible, intermediate, resistant (SIR) and not defined, susceptible-dose dependent, and non-susceptible. There are 5 typing methods, 15 typing platforms, 12 vendors, 16 LTMV and 6 testing standards such as MIC, CLSI, and agar dilution. Thus, the amount of variations by assay greatly reduces the number of reference data points per class.

To predict MIC values for a given antibiotic, we process and label data according to the following stages [data dist in Table 1]:

1. raw data (Data Stage 1)
2. removal of all rows missing values for the headers shown above (Data Stage 2).
3. For self consistency, we use data only from high quality, complete genomes. For these data sets, we downloaded the sequence data, assembled and annotated them in order to identify all relevant nucleotide and amino acid sequences (Data Stage 3).

### Table 1: Assay Distribution - Data Processing

| Data Stage     | #Antibiotics | #SRA_ID | #Genomes | #assays |
|----------------|--------------|---------|----------|---------|
| Data Stage 1   | 99           | 4963    | 5173     | 77424   |
| Data Stage 2   | 74           | 4962    | 4962     | 73600   |
| Data Stage 3   | 50           | 1399    | 1399     | 30076   |

Note: the compounds associated with only a single assay were removed as there was insufficient data. The final distribution of ground truth data after conflict resolution is listed in Table 2 and Table 3. We only show entries where we found more than 1000 data points. Table 1 shows approximately 40% of assays were kept after curation. Table 2 refers to all genome accessions independent of microbial genus. Table 3 shows how the curated measurements are distributed across genera with more than 1000 data points.

### Table 2: Data Point Distribution by Antibiotic - Data Stage 3

| Antibiotic                              | # of MIC Data Points |
|-----------------------------------------|----------------------|
| trimethoprim-sulfamethoxazole           | 1336                 |
| ciprofloxacin                           | 1314                 |
| tetracycline                            | 1296                 |
| ceftriaxone                             | 1240                 |
| gentamicin                              | 1234                 |
| levofloxacin                            | 1204                 |
| amikacin                                | 1147                 |
| imipenem                                | 1125                 |
| tobramycin                              | 1111                 |
| cefazidime                              | 1107                 |
| ampicillin                              | 1105                 |
| cefotaxime                              | 1067                 |
| aztreonam                               | 1026                 |
| cefazolin                               | 1011                 |

### Table 3: MIC Data Point Dist by Genus - Data Stage 3

| Genus          | #   |
|----------------|-----|
| Acinetobacter  | 13992|
| Klebsiella     | 4423 |
| Streptococcus  | 1567 |
| Escherichia    | 1456 |
| Salmonella     | 1109 |

or organization that conducted the experiment, GIS data, source of the isolate, as well as the Antibiogram data. All of the software and process for Genome Assembly, Annotation, Curation, and Selection are described in detail in [11].

### 3.2 Extracting Features from Assays

Biochemical assays include measurements for a particular antibiotic at various concentrations tested with a specific cultured organism. Resistance at a particular concentration does not necessarily imply the organism is resistant to the antibiotic at all concentrations. Here, susceptible implies the organism does not grow (or, for some protocols, that it dies) at a particular concentration. At very low concentrations, resistance is the expected outcome for any compound. To classify the organism’s susceptibility one needs to know the minimum inhibitory concentration (MIC), defined by the maximum concentration at which the organism is resistant and/or the minimum concentration at which it is susceptible. It is then necessary to compare the critical MIC obtained from the assay to the maximum safe does for the patient, which varies between humans and livestock. Different laboratories use different protocols in conducting assays. Some labs may stop measurements if, for example, the organisms is found to be susceptible at low concentration or resistant above a defined concentration.

In this study, we first categorized all assays by specific antibiotic and then by genome accession. For each antibiotic-accession pair we then created a resistant values list, a susceptible values list and an intermediate values list. Each list contained only two entries, min concentration and max concentration for each category. Thus, for each antibiotic-accession pair, we extracted from the assay unique the resistant_min, resistant_max, susceptible_min, susceptible_max, intermediate_min, and intermediate_max.

In order to do the above, we parse each assay and identify the antibiotic-accession pair. Then we look at the phenotype description - SIR. This dictates which list, resistant, intermediate or susceptible, to update. We then look at the measurement and the measurement sign and update the min or max values of the relevant list.

The data from some assays may contain only a range of resistant or a range of susceptible concentrations, thus a switch from resistant to susceptible is not observed. Therefore, for each assay we extract the two important values, i.e. min susceptible and/or max resistant concentration, either of which can define the MIC value described above. However, it is still necessary to resolve experimental error, noise, and associated conflict within this data. For further data cleaning, we define and describe the steps in the next two sections.
3.3 Conflict Identification

One of the most important factors to consider while training data is noise. Noisy data can often produce spurious results and degrade performance, even if the evaluation scores are high.

In assay data, conflicts are a major problem due to the scarcity of the data. We take a subset of fields from Antibiogram, SRA_ID (genome accession), phenotype_description, measurement, measurement_sign and compound. From this data we identify two types of conflicts: 1. Direct Conflict: all fields but phenotype description identical, 2. Range Conflict: where SIR ranges overlap Fig 2.

Up to 10% of measurements per genus were identified as conflicts, with the largest number of conflicts, 1441 values, within Acinetobacter. Up to 35% of measurements for some antibiotics were conflicting, with ceftriaxone containing the most number of conflicts, 434 values. [Note: For sake of brevity, we haven’t included full table on number of conflicting entries.]

There can be multiple reasons for conflicts in phenotypes, including differences in testing standards or testing equipment. However, irrespective of these, it is necessary to clean the data and ensure that confounding conflicts are not passed on to the ML model. It is also important to note that we measure conflicts for a particular antibiotic-accession pair, thus evolution of genomes need not be considered here as they will be captured by different accessions.

We have thus come up with a novel method to transform the data such that these conflicts are resolved (calculate probable MIC) and do not impede model’s learning rate.

3.4 Conflict Resolution

To curate the assay data we developed approaches to resolve both the range conflicts and direct conflicts discussed above. These methods can potentially be improved through future work by looking at additional experiment fields like typing_method etc, or by looking at BioSample metadata for the genomes or experiment.

3.4.1 Range Conflicts. Figure 2 illustrates a range conflict. Over a range of concentration, multiple data points are associated with conflicting outcomes (some concentrations, \(C\), are labeled both resistant and intermediate or vice-versa). These conflicts can be conservatively resolved by adopting the higher \(C_S\) and lower \(C_R\) to avoid erroneously predicting the drug is effective when there is a discrepancy. Range conflicts were resolved by the following:

- Calculate concentration ranges for each SIR phenotype
- Calculate the fraction of entries (conf. score) in each range
- Determine min and max concentrations for SIR phenotypes
- Keep entries within each range (discarding the rest)

In order to decide which range to resize, we use the confidence value. The aim is to ensure that the MIC value is correctly identified. MIC value is defined by the higher value between maximum resistant value and the minimum susceptible value. Thus, we resolve each range using:

\[
\text{MIC} = \max\{C_R(\text{min}), C_R(\text{max}), C_I(\text{min}), C_I(\text{max}), C_S(\text{min})\}
\]

We include both \(C(\text{min})\) and \(C(\text{max})\) since in most cases only one of the two values are obtained in the assay.

3.4.2 Direct Conflicts. To resolve direct conflicts, we check if the measurement value and phenotype agree with the new ranges. If there is agreement, we keep the assay and increase the confidence value, else we discard it.

Figure 1 shows the list of antibiotic-accession pairs and the associated min/max values in the resistant, susceptible and intermediate categories. For each data point we plot \(C(\text{min})\) and \(C(\text{max})\), colored based on phenotype.
4 DISCUSSION

4.1 Ground Truth Data Availability and Bias
In Data Stage 2 we removed all assays with null SRA_ID and testing standards other than MIC (see Section 2 Data). To analyze this data, we plotted the distribution of min susceptible and max resistant MIC values for each antibiotic after sorting them by phenotype. We placed the assays with R phenotype first, followed by I then S. We did not consider ambiguous MIC values listed as ‘not defined’, ‘susceptible-dose dependent’ or ‘non-susceptible’. The number of assays with such phenotype descriptions was minimal in the complete set of 73600. The MIC distribution for Tetracycline is shown in Figure 1.

In Figure 1A, the x-axis is the genome accessions and y-axis is the concentration value. Resistant values occur at higher concentrations and susceptible values are reported most often at lower concentrations, which may seem contrary to expected observations. This may reflect underlying lab procedures. E.g. if an assay reveals resistance at concentrations higher than approved max dose, the lab may not test for susceptibility at higher concentrations since the drug can not be prescribed. Conversely, isolates found to be susceptible at concentrations at or below approved dosages need not be tested for resistance at much lower dosages. The divide between red and green reflects a concentration range determined by approved clinical practice. From a ML perspective, depending on the model and use case, it can be important to pass ranges instead of singular values. We can use such data for a phenotype classifier, with input features as representation of the genomes, such as component genes, along the x-axis. From the distribution of cleaned training data, we expect the classifier to perform reasonable well. However, using the same data to perform a regression task poses problems since it is difficult to fit a regression curve to the distribution. Thus, such analysis highlights both the type of model it is suited for, and the type of prediction possible (phenotype rather than MIC value).

Figure 1B, shows the distribution for Trimethoprim Sulfamethoxazole. We observe a large no.of data points at similar values. Since most assays are of Acinetobacter genomes, the high level of homogeneity between the features (gene sequences) and lack of data makes construction of a sensitive model, for both phenotype classification and MIC prediction, difficult.

4.2 Data Stage 3
From Table 2, we see that once we subset the assays to include only those with high quality and complete genomes, the data is considerably reduced per antibiotic. From Table 3, we see that the largest number of assays are present for the genus Acinetobacter and the least is for Listeria. This indicates another possible bias in the sequencing data and reflects that Acinetobacter poses a more immediate threat in areas like antimicrobial resistance.

4.3 AntiMicrobial Resistance Prediction
The final cleaned data discussed above can be used for ML tasks, specifically ones like prediction of antimicrobial resistance.

In order to perform AR prediction, it is crucial to determine which entity to predict, instead of simply predicting the resistance phenotype, resistant, susceptible or intermediate, for a genome.

As described previously, MIC is the highest concentration at which bacteria is resistant to the antibiotic or the lowest concentration at which it is susceptible. Furthermore, to predict the resistance phenotype, one can simply identify thresholds of safe concentrations of antibiotics for different target users and assign a phenotype based on the minimum safe dosage.

Though the discussion of ML model is outside the scope of this paper, however, we would like to share that using component resistant gene sequences [5] for each genome and XGBoost we built an AR MIC prediction model and get an R² value of 0.67. Also, phenotype classification accuracy was 0.94. Though we are still working on the approach, high phenotype classification accuracy is promising and more robust than gene counting.

Since different antibiotics target different cellular processes, the gene features relevant to prediction of resistance vary by compound and it is necessary to partition by compound. Drugs are often prescribed based on organism name or genus. Partitioning training data by genus may reduce conflicts, but reflects an incomplete understanding of the full resistome given the reality of HGT.

5 CONCLUSIONS
Available assay data has conflicts which must be resolved before being input to ML models. To ensure higher sensitivity and specificity, we developed a method for conflict identification and for extraction of MIC values that can be used to train AR/MIC prediction models. Despite this curation, there remains a serious lack of data per antibiotic and per genus. Even the genus with the most assays, Acinetobacter, has insufficient data to train an effective ML model. Unbalanced distribution make it harder to train regression models than classification models. Sampling bias must be considered when training any model. It is important to focus on development of robust ML models that have accurate predictions with limited data.

ACKNOWLEDGMENTS
The authors would like to thank Dr. K. Beck and Dr. V. Mukherjee for helpful discussion and improvements to the manuscript.

REFERENCES
[1] Ziad Obermeyer et al. Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine, 375(13):1216, 2016.
[2] Gustavo Arango-Argoty et al. Deeparc: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome, 6(1):23, 2018.
[3] Shun Yang et al. Data sharing and reproducible clinical genetic testing: successes and challenges. In PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017, pages 166–176. World Scientific, 2017.
[4] Vanessa M D’Costa et al. Sampling the antibiotic resistome. Science, 311(5759):374–377, 2006.
[5] Steven M Lakin et al. Megares: an antimicrobial resistance database for high throughput sequencing. Nucleic acids research, 45(D1):D574–D580, 2016.
[6] José L. Martinez, Teresa M Coque, and Fernando Baquero. What is a resistance gene? ranking risk in resistomes. Nature Reviews Microbiology, 13(2):116, 2015.
[7] Sophie Nolivos et al. Role of acраб-tolc multidrug efflux pump in drug-resistance acquisition by plasmid transfer. Science, 364(6442):778–782, 2019.
[8] Jennifer M Andrews et al. Determination of minimum inhibitory concentrations. J. Antimic. Chemother, 48:5–16, 2001.
[9] Akshay Agarwal et al. Survey of public assay data: Opportunities and challenges to understanding antimicrobial resistance. bioRxiv, 2019. doi: 10.1101/2019.12.13.874909. This is extended version of the current paper.
[10] Scott Federhen et al. Toward richer metadata for microbial sequences: replacing strain-level ncbi taxonomy taxids with bioproject, biosample and assembly records. Standards in genomic sciences, 9(3):1275, 2014.
[11] Edward E Seabolt et al. Omware, a cloud-based platform for studying microbial life at scale. arXiv preprint arXiv:1911.02095, 2019.