SMU Data Science Review

Volume 2 | Number 2 Article 12

2019

Using Machine Learning for Antimicrobial Resistant DNA Identification

Jason I. Lingle
SMU, jlingle@smu.edu

John Santerre
SMU, jsanterre@smu.edu

Using Machine Learning for Antimicrobial Resistant DNA Identification

Jason Lingle, John Santerre

Southern Methodist University
Dallas, Texas USA
{jlingle, jsanterre}@smu.edu

Abstract. In this paper, we present a machine learning-based methodology for identifying bacteria DNA sub-sequences that are associated with antimicrobial resistance. The dramatic rise in cases of antibiotic resistant bacteria being an increasing threat across the globe with existing treatments rendered ineffective in many cases due to mutations of their DNA. Neisseria Gonorrhea is one such bacteria that has shown increasing antimicrobial resistance (AMR) and has been identified as an urgent threat by the Centers for Disease Control (CDC). To help address AMR, researchers need tools to help them identify resistance quickly and accurately. Fortunately, by the use of machine learning, we are able to make a prediction of an isolate being resistant or susceptible to antibiotics given a unique, overlapping DNA substring or k-mer of length 10 with accuracy above 90%. Using this approach, researchers without resources may be able to identify the resistant isolates to determine proper course of action and prevent wide infections.

1 Introduction

Antibiotics are drugs that are used to treat and prevent bacterial infections. They have been an important tool for doctors in dealing with infections since the introduction of penicillin in 1928. While other antibiotics have been developed since then, antibiotic resistant bacteria have rapidly emerged as a threat. Almost all antibiotics have developed some level of resistance and have lead to increasing health and economic burdens to health care [1]. In the United States alone, at least 2 million people are reported each year as having an antibiotic resistant infection which leads to at least 23,000 fatalities. In 2013, the Centers for Disease Control and Prevention (CDC) published a report on the top 18 antibiotic threats in the US which highlighted the increasing threat of antimicrobial resistance to public safety. Furthermore, some projections suggest that deaths due to bacterial infections that are resistant to multiple antibiotics are projected to reach 10 million annually by 2050 which is a higher mortality rate than cancer [2]. In September 2018, the United Nations further sounded the alarm related to this threat by issuing "The AMR (Antimicrobial Resistance) Challenge". The purpose of the challenge was to seek alignment from parties

such as pharmaceutical companies, medical professionals, and governments in areas such as development of drugs and testing as well as antibiotic use.

Treatment of patients is usually time critical, but timing is even more essential when antimicrobial resistant organisms are involved. Rapid identification of antibiotic resistant bacteria greatly improves the clinical and financial outcomes for patients[3]. Additionally, proper identification of antibiotic resistant infections is one of the most effective ways to reduce antibiotic misuse. Rapid and accurate diagnosis have been cited in helping reduce the empiric use of antibiotics where selection of the proper antibiotic is based on the provider's experience or educated guess. This is important as in some cases, patients may be prescribed multiple antibiotics which are harmful to their antibiotic resistance in the long term[4].

With 78 million cases globally, Gonorrhea is among the most prevalent sexually transmitted infections (STI) even as diagnosis rates continue to increase[6]. The bacterium that causes Gonorrhea, Neisseria Gonorrhea, has been identified as a global threat and priority as it is becoming increasingly resistant to existing treatments[5]. Extended-spectrum cephalosporins (ESCs) represent the "last-line" treatment, and there are no new options available for treatment and few in development as the current. As the first-line antibiotics treatments using ciproflaxin, cefixime, and azithromycin for Neisseria Gonorrhea have experienced high levels of resistance, there is increasing concern that Neisseria Gonorrhea may become untreatable. This is highlighted by a case in 2018 where an isolate from a heterosexual male being treated for Neisseria was resistant to treatment from combined high-level azithromycin resistance and ceftriaxone.

This increasing resistance to existing treatments poses a threat to segments that are at high risk of exposure such as men who have sex with men (MSM), individuals who have sex with multiple partners, and recreational drug users [6]. The threat of this STI is increased by the lack of signs to those infected by Gonorrhea as symptoms typically develop after two weeks but might be ascribed to something else. Furthermore, Gonorrhea is spread easily with the transmission rate from a single sexual encounter is 50% to 60% from an infected man to an uninfected woman and 20% from an infected woman to an uninfected man [7].

The techniques used for identifying AMR bacteria has evolved over time from culture-based testing. Antibiotic resistance information was traditionally gathered by public health officials who analyzed bacteria isolates after they were grown in a culture [8]. Conventional approaches are known to have limitations as they are heavily targeted to testing human pathogens. Additionally, only a few bacterial microbes can be cultured. Sequencing of pure culture isolates is a newer approach to detection and identification of AMR. By using genome sequencing, it is possible to identify the AMR phenotype with high accuracy [8]. However, genome sequencing may not be the best solution as most models interpret the genomic code rather than predicting a phenotype from the genome sequence because of the complex relationship between a genotype and phenotype [9, 10]. Moreover, DNA sequencing is a complex task where there is a high risk of misinterpretation without the appropriate expertise [11].

To improve the process in identifying the genes that have mutated to make Neisseria Gonorrhea resistant to antibiotics, we identify the responsible DNA regions which can then be mapped back to genes using machine learning. We take each contig and divide them into smaller sub-sections called k-mers which can be used as a feature. We can then evaluate those features using feature importance to identify the k-mers that are most important for antimicrobial resistance. While we are comparing the DNA of susceptible and resistant isolates to identify mutations, this is by no means a simple string search. We will create a unique, overlapping matrix with sub-strings of a fixed length k. The matrix is typically sparse with millions of columns before it is compressed to half a million. The benefit of this approach is that we have successfully identified the sequencing responsible for genetic mutation that are responsible for AMR in a way that would otherwise be computationally difficult to execute given the complexity of DNA.

We train the models using Naive Bayes, Random Forest, and Support Vector Machines (SVM) classifiers for k-mers of length 5 through 10. We find that Random Forest performs consistently better than Naive Bayes and SVM until lengths of k = 10 where SVM produces an accuracy of 93.6% compared to 92% for Random Forest.

In our work, we find results that support prior evidence that this approach for using k-mers of increasing lengths can be used to identify gene regions most likely responsible for AMR. Further, we use a different dataset than used in prior research to provide further evidence of the approach. However, we also find that using k-mers with a length of 10 can still provide high accuracy results which allows for further research without optimized hardware.

In this paper, we review similar works performed by others in predicting AMR using various machine learning methods. We review the scope of their work in terms of data as well as the algorithms used. We next provide a short tutorial on some of the topics that are discussed to provide additional background. This background covers items related to technical and subject matter. We also provide an overview of the data set used in our analysis including the source and how the data was gathered. Next, we present our results and analysis as well as ethical considerations related to this topic as well as our solution. Finally, we present the conclusions that we have reached from this research as well as future works that we would consider for this topic.

2 Related Work

In this section we discuss related works and research trends for the use of various machine learning methods in identifying antimicrobial resistant microbes.

Davis et. al [13] provide insights into the infrastructure built at PATRIC with different sets of genomes that are binned by their AMR phenotype. Various machine learning classifiers are made available through PATRIC to detect AMR phenotype. In their study, the k-mer counting program KMC is used in their research to calculate the number of k-mer occurrences within each contig, or strand

of DNA, but the counts are suppressed in lieu of "1" to indicate presence and "0" for absence to create a binary matrix. Next, the AdaBoost classifier is used to identify the most informative k-mers which are used to predict the susceptibility or resistance to an antibiotic. Further, they used the AdaBoost classifier using 31 mers as sizes ranging from 24 and 31 provide higher levels of accuracy in similar classifications. Using the classifier, Davis et. al are able to identify resistance to various antimicrobials in Acinetobacter baumannii, Staphylococcus aureus, and resistance in Streptococcus pneumoniae with accuracy between 8899% and Mycobacterium tuberculosis with accuracy between 71-88%.

Santerre et. al [14] explore the use of k-mer space and Random Forest for a de novo identification of Single Nucleotide Protocols (SNP) or gene regions associated with resistance. Naive Bayes, AdaBoost, Lasso, and Support Vector Machines (SVM) classifiers are also used for comparison of the results. 4 different types of bacteria with different antibiotic resistance, including Acinetobacter baumannii, Staphylococcus aureus, Mycobacterium tuberculosis, and Klebsiella pneumoniae that are loaded into a matrix before the classifiers are run on k-mers of length 10 through 20. While accuracies for Acinetobacter baumannii range from 88-94% and Staphylococcus aureus of 83-99%, accuracy for Mycobacterium tuberculosis varies significantly between resistance to different antibiotics. They theorize this is due to the pathway the antibiotic interacts with.

Brooks et. al [15] explore the use of correlation-informed incoherent framing (CIF) for dimensionality reduction and feature selection using breast cancer and cell lung cancer datasets. Using CIF, sensing matrices are constructed based on the idea of random assignment. Brooks et. al found that CIF improves classification accuracy. Further, a k length of 12 was used in the tokenization process.

Hamid et. al [16] explore the use of Stochastic Gradient Langevin Dynamics (SGLD) to classify AMR using a dataset with 14,907 sequences that are resistant to 15 antibiotics. Hamid et. al compared the accuracy of the models trained using SGLD to those with ADAM which is an adaptive learning optimization algorithm and found that ADAM produced higher accuracy and confidence scores.

Drouin et. al [17] explore the use of rules based algorithms including Classification and Regression Trees (CART) and Set Covering Machines (SCM) to predict phenotypes using AMR data from PATRIC for 12 different types of bacteria in "Interpretable genotype-to-phenotype classifiers with performance guarantees." In the study, k-mers of length 31 were used when building the models. Douin et. al found that both models produce high accuracies with 95% of the models producing results greater than 75% while also generating interpretable results.

Kavvas et al. [18] focused on Mycobacterium tuberculosis using 1,595 sequenced strains from the Pathosystems Resource Integration Center (PATRIC) database. They used Support Vector Machine (SVM) to identify AMR genes. In their research, their focus was not high accuracy but rather key insights from the data. Among their findings were a high number of genetic interactions "underlying variable AMR phenotypes." (Need to include the accuracy here).

Her et al. [19] also used the data set from the PATRIC database, focusing on Escherichia coli. They proposed a pan-genome based approach to characterize AMR strains and hypothesized that it would better define and predict AMR. Pan-genome describes shared features of all strains of a type of bacteria to understand the strain-level diversity of the species. A previous pan-genome study of Klebsiella pneumoniae found that it could be split into three distinct groups with branches that could be hyper-virulent or multi-drug resistant. SVM was used to predict resistant or susceptible genes. They reached a conclusion that a small set of accessory genes with AMR activity annotations achieved the best predictive accuracy. (Need to include the accuracy here).

Drouin et al. [20] used a total of 36 data sets of five bacterial species from PATRIC to both measure performance of their model and predict antimicrobial resistance using Set Covering Machine (SCM). The SCM algorithm generates models that are logical combinations of Boolean valued rules from the data. Drouin et al. also compared the performance of their model using SCM with other models such as linear, kernel-based SVM, and decision trees and found that their model achieved comparable and occasionally better performance. The SCM models predicted with an accuracy of 80% or better in 33 of 36 data sets. They also found that the SCM rapidly generated the results with limited computational resources.

2.1 Gonorrhea

Gonorrhea is a sexually transmitted disease (STD) caused by the Neisseria Gonorrhea bacteria with over 800,000 estimated infections in the US each year, which makes it the second most prevalent STD [21]. Some of the symptoms of gonorrhea are discharge and inflammation of the urethra, cervix, pharynx, or rectum [1]. However, gonorrhea is frequently asymptomatic, especially in women, and fewer than half of the cases are diagnosed [22]. Lack of diagnosis and treatment can lead to life-threatening ectopic pregnancies in women and infertility in both men and women as well as an increased risk of acquiring HIV [21].

The CDC has monitored the response of Neisseria Gonorrhea to antibiotics since 1986 under the Gonococcal Isolate Surveillance Project (GISP) [21]. Treatments have been available for gonorrhea for decades, Neisseria Gonorrhea has developed resistance to most antibiotics used to treat it. While penicillin was used to treat penicillin from the 1940s, increasing resistance forced higher doses until it was no longer viable in the 1970s [23]. Currently, there is only one recommended treatment option available which is the dual treatment of the antibiotics anithromycin and ceftriaxone [21]. Neisseria Gonorrhea bacteria has been identified as an urgent threat which requires additional monitoring and prevention activities [1].

2.2 Antimicrobial resistance (AMR)

Antimicrobial resistance occurs when a bacteria or microbe becomes resistant to the effects of a drug after exposure. There are three ways that a bacterium might

develop antimicrobial resistance. First, natural resistance by the bacterium. Resistance can also be developed as genes can be inherited from other bacteria through horizontal gene transfer (HGT) from a species that is resistant [1]. Finally, resistance can be developed by mutation as antibiotics remove any drug-sensitive genes but leave behind the resistant bacteria which are then able to reproduce as part of natural selection [1]. Antimicrobial resistant (AMR) organisms include strains of bacteria, fungi, parasites, and viruses. They prolong illness, increase the likelihood of fatality, and increase treatment costs of patients. Antimicrobial bacteria are classified by Phenotype which indicate whether the organism is resistant or susceptible to one or more antibiotics. A significant number of different genes may be responsible for AMR so identification of the genes is important to verify the phenotype that are resistant [24].

Germs have a number of resistance mechanisms which "protect" them from antibiotics including target alteration, impermeability, enzymatic modification or destruction, and efflux [26]. With target alteration, changes in the antibiotic target interfere with or limit antibiotic interaction which promotes resistance [26]. Using impermeability, antibiotics are unable to cross the outer membrane of bacteria [26]. With enzymatic modification or destruction, resistance develops when genes that are susceptible are destroyed or genetic elements that are resistant are acquired [26]. Finally, with efflux, bacteria develop resistance by using the efflux pump to move antibiotics outside of the cell.

The first commercialized antibiotic, penicillin, was introduced in the 1940s to treat infections [1]. Antimicrobial resistance to penicillin began shortly after its introduction, but other antibiotics in the form of beta-lactam antibiotics were discovered, developed, and deployed in the 1950s before they too began to experience resistance in the late 1950s and early 1960s. This cycle has continued as new antibiotics have been discovered, developed, and deployed by pharmaceutical companies, but they are no longer developing antibiotics at the same pace that they had through the early 1980s [1].

2.3 Antimicrobial resistance identification

Antimicrobial susceptibility testing (AST) is performed by clinical microbiology laboratories to verify that bacterial isolates are susceptible to antimicrobial agents such as antibiotics. AST is important in verifying that isolates have not acquired resistance to standard treatments. The standard process for AST begins when an individual sample or isolate is taken and placed in a petri dish where they are separated and cultured. Next, a sample is put it into what is a type of "blender." From the blender, you find an example of 4 million base pairs that you feel like describe the population. Next, a low dose of antibiotic referred to as the the minimum inhibitory concentration (MIC) is placed in one of the samples. In the other, no antibiotics are placed. The samples are left, usually overnight. There are some examples that can and cannot survive the antibiotics. The bacteria that do not survive treatment by the antibiotics are referred to as the susceptible phenotype while those that continue to survive after treatment are referred to as the resistant phenotype. One can then start to look at

what is different in the DNA of the resistant and susceptible samples using DNA sequencing.

2.4 DNA Sequencing

Deoxyribonucleic acid (DNA) sequencing is used to determine the order of "bases" which are the chemical building blocks of the DNA molecule [25]. The DNA Sequence is genetic information carried out within a DNA segment, and if understood can be used to determine which pieces of DNA that contain genes and pieces transmit instructions. Sequencing can be used to identify variations in a gene that may cause a disease. The four chemical bases that bond together to form base pairs in DNA sequences are illustrated in table 1 with their International Union of Pure and Applied Chemistry (IUPAC) codes.

Table 1: DNA Bases

IUPAC Code	Base
A	Adenine
C	Cytosine
G	Guanine
T	Thymine

K-mer analysis is one of the more widely used approaches for modeling DNA and protein sequences by counting every DNA word with length k using a sliding window [27]. This is done to "eliminate the influence of an arbitrary chosen starting point" [28].

2.5 Machine Learning

In Machine Learning, a matrix is an important and frequently used two-dimensional structure arranged into rows where m is the number of rows and n is the number of columns. There are 3 standard data structures including "tall and skinny" where $n < m$, "short and fat" where $n > m$, and "roughly square" where $n = m$ [29]. With the Neisseria Gonorrhea data used and other data made available by PATRIC, the number of features far exceed the number of isolates which are rows, resulting in a short and fat matrix. As a result, we encounter the issues associated with high dimensionality including the number of features possibly masking noisy data. Additionally, the overhead required for computations increases.

3 Data Set

Pathosystems Resource Integration Center (PATRIC) is an information system at the University of Chicago funded by the National Institute of Health (NIH)

and National Institute of Allergy and Infectious bacterium Diseases (NIAID). PATRIC provides a variety of resources to support the research community including a number of rich datasets. Each file contains short reads which are sequences of varying length of DNA fragments. Short reads provide a less expensive alternative to long read sequencing [14].

While the file contains an assembled collection of short reads called contigs, they are not necessarily in the correct sequence. Rather, they are ordered based on descending length of the contig.

The files are obtained through the PATRIC FTP site. The dataset used in this research is limited to the Neisseria Gonorrhea bacterium antimicrobial resistance (AMR) dataset for azithromycin suscptibility and resistant isolates. While the dataset includes a "features.tab" file which includes details related to each isolate, only the ".fna" files are used in our analysis. Within the ".fna" files, there are multiple sections, as illustrated in Figure 1, which separates the different contigs, that are contained within the section. For example "485.291.con.0001" indicates that we are looking at the first contig or strand for genome id 485.291. The files are anonoymized, with no specific reference to the subject from which the isolate has been taken, but additional information may be found on the PATRIC site with geographical, subject, and timeline information generally available.

```
>485.291.con.0001      [Neisseria gonorrhoeae 19095 | 485.291]
aaaaaacgttgaccggaaccggctgtccgcccggtcaaaagcgcaaaaagaaaacatcgc
ccggatgttgtcgggcgcaaaggtatcggaagacgaagccctgacgtgcggcatcatgat
gcggctggttgccgccaaagtccgccgatactgcggcgaacatccgggcgtatttgacgg
cgcggcaggaagcaggcagcttgaacggtatattaagccgtctgaatttcacgccgtgga
aatacagcccgaagcctgcaaagccttgttgcaaaactaccctgccgcggccggtaatct
cggcaggcggaaacaaaaaaagccccgagttcgtaagcgtatcagccctagccgaatggc
```

Fig. 1: Example of Neisseria Gonorrhea Isolate. An example of the first lines of an isolate from PATRIC which includes the header and part of the first contig within the file.

There are a total of 392 files available from the Neisseria Gonorrhea AMR phenotype. Each file represents the DNA of an individual case or isolates which were collected between 2002 and 2018. Of the 392 isolates, 178 of the isolates are susceptible which means that they can be successfully treated with antibiotics. The remaining 214 isolates are antibiotic resistant meaning that they are cases where the Neisseria Gonorrhea was resistant to treatment by antibiotics. Each isolate is approximately 2 million base pairs with lengths ranging from 2,069,600 and 2,245,905. Two additional isolates have lengths of 4,127,771 and 4,170,863. There is variation in length of the files within both the susceptible as well as resistant cases.

PATRIC collects genome level AMR phenotype from National Center for Biotechnology Information (NCBI) BioSample and Anti-biogram records [12]. Additional metadata is also collected from National Institute of Allergy and Infectious Diseases (NIAID) funded Genomic Centers for Infectious Diseases.

4 Approach

We build a model using machine learning to identify the gene region in the antibiotic resistant bacteria responsible for the phenotype. Contigs are divided into k-mers based on overlapping sub-sequences of length k. In the matrix representation of our problem, rows corresponds to isolates (files) and columns correspond to the unique k-mer. The value in our matrix corresponds to the total number of times the k-mer occurred in that particular isolate. As shown in Figure 2, one column of the matrix corresponds to the k-mer "ggacg". The first row of the matrix has shows a corresponding value of 11 for that column which is the number of occurrences for the k-mer "ggacg". For computational reasons, we first generate a dictionary representation, which appears in Figure 2.

```
ggacg,"{0: 11, 1: 19, 2: 16, 3: 18, 4: 12, 5: 12, 6: 11, 7: 16, 8: 9, 9: 8, 10: 14
```

Fig. 2: The Master Dictionary is created with an outer dictionary of k-mers with an inner dictionary of the isolate index and count of the k-mer.

Table 2: K-Mer by Isolate Matrix. An example of the matrix created of k-mers in the columns with the isolate index on the rows. The count of occurrences of the k-mer within each isolate are populated in the matrix.

isolate	**attat**	**ttatg**	**tggac**	**ggacg**
Isolate 1	0	1	2	0
Isolate 2	0	0	0	1
Isolate 3	0	0	0	0
Isolate 4	1	0	0	0
Isolate 3	0	0	0	0

We fit our model using the matrix given the set of features and make a prediction of an isolate being resistant or susceptible to antibiotics using Random Forest, Naive Bayes, and Support Vector Machine (SVM). We use the default hyperparameters for each classifier while setting a random seed of 75 to ensure repeatability of our model. We perform the same process for k-mer lengths of 5 through 10 and compare the accuracy of each model.

As can be seen in table, the length of k in k-mers shows that there is a rapid increase in the amount of memory required as the count increases which results in a sparse matrix. Importantly, the space required is only a fraction of potential exponential theoretical range as the number of actual k-mers in the empirical columns indicates. We see that the separation begins to occur with length of k = 10.

Table 3: Number of k-mers as k Increases. The number of theoretical and empirical k-mers as the size of k increases. The theoretical and empirical occurrences diverges significantly only when k increases to 10.

K	Theoretical	Empirical	K	Theoretical	Empirical
1	4	4	6	4,096	4,096
2	16	16	7	16,384	16,384
3	64	64	8	65,536	65,535
4	256	256	9	262,144	261,448
5	1,024	1,024	10	1,048,576	979,112

In addition to accuracy to measure the performance of our model, we further evaluate the performance of our models using the following metrics:

1. Accuracy: The ratio of correct predictions out of the total number of predictions.
$$\frac{TP + TN}{TP + TN + FP + FN}$$
2. Sensitivity (or Recall, or True Positive Rate): Sensitivity characterizes the performance the classification.
$$\frac{TP}{TP + FN}$$
3. Specificity: Helps us understand how many were correctly classified as susceptible out of the total true resistant cases.
$$\frac{TN}{TP + FN}$$
4. Precision: Helps us test the certainty of our predictions where we want to identify the susceptible cases that were truly susceptible by giving the proportion of true positives to predicted positives.
$$\frac{TP}{TP + FP}$$
5. Area Under the Curve (AUC): Summarizes the overall performance of a classifier on the training data. A larger area under the curve where the curve "hugs" the top left corner indicates a better performing model.
6. F1 Score: Provides a balance of precision and recall scores. A score of 1 indicates perfect precision and recall.
$$\frac{2 * Precision * Sensitivity}{Precision + Sensitivity}$$

We posit that our solution will not only provide a highly accurate model but one that performs well in each of these metrics. Finally, we will measure the performance in terms of time to perform the tasks to evaluate the ability to execute the task on any machine and not just a high-performance machine.

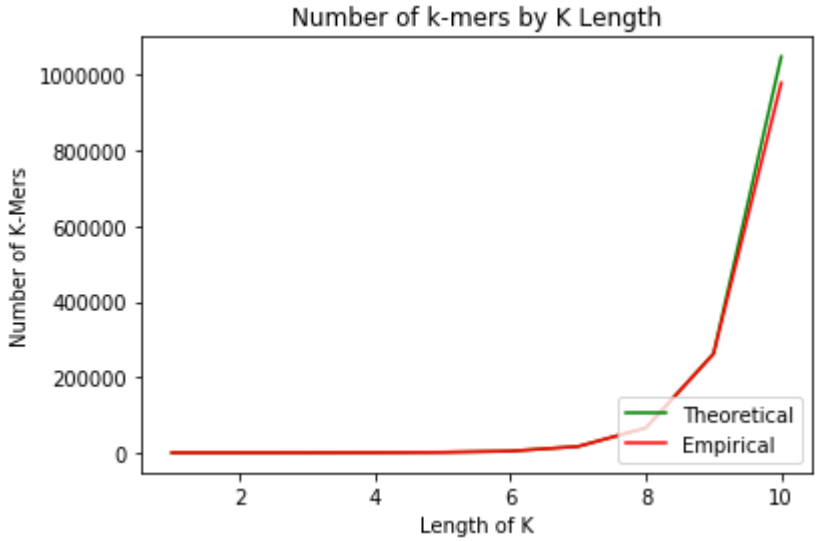


Fig. 3: The theoretical and empirical number of k-mers diverge when k is 10.

5 Results

We execute the model using the Random Forest, Naive Bayes, and Support Vector Machine (SVM) classifiers for k-mers of length 5 through 10. We find that the use of lower k-length results in considerably lower accuracy for Naive Bayes and SVM before they each improve at k of 8.

The model build time also increases significantly when the k-mer size increases to 10. Because of the significant increase in number of features in the columns of the matrix, it is not possible to increase the length of k with the hardware used for this study.

The average build time for the model as the length of k increases. While build time increases significantly with k = 10, the model is unable to build for k greater than 10 due to memory constraints as the number of features increases.

While all 3 classifiers produce models that improve while increasing length of K, Random Forest consistently generates models with a high degree of accuracy ranging from 79-92% for length of k between 5 and 10. Naive Bayes produces an accuracy ranging from 51-85% while SVM produces an accuracy ranging from 50-94% with much larger disparity in accuracy for shorter lengths of k which corresponds to previous works performed on larger k-mer sizes.

The classification report for all 3 classifiers for k = 10 provides similar results for F1, precision, and recall for SVM and Random Forest. Meanwhile, Naive Bayes provides results for all metrics below SVM and Random Forest.

The plot in figure 8 is generated using Gaussian Mixture Model (GMM) to better visualize the feature identified by Random Forest with the highest feature importance. The plot has the number of occurrences of the feature on the y axis

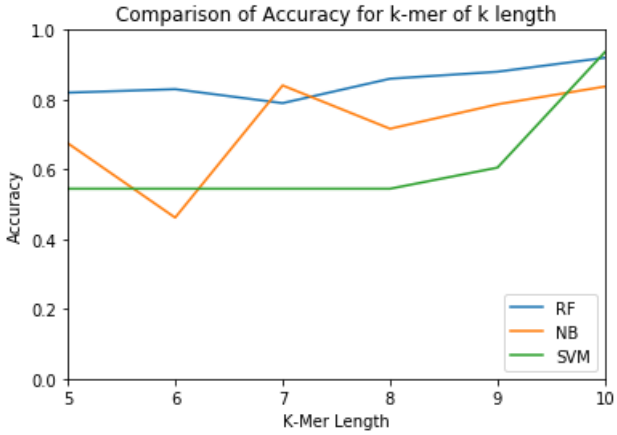


Fig. 4: Accuracy for k-mers of length 5 through 10 for Random Forest, Naive Bayes, and Support Vector Machine. Accuracy for all classifiers improves as k increases.

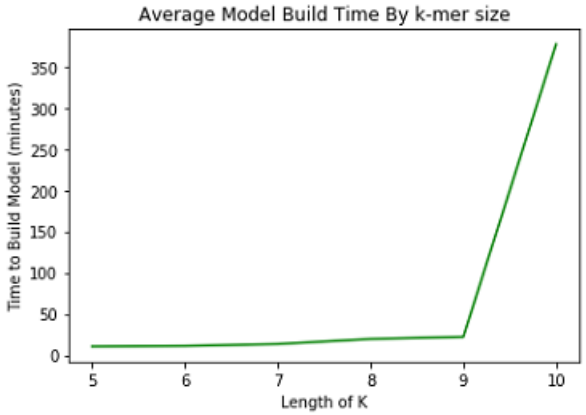


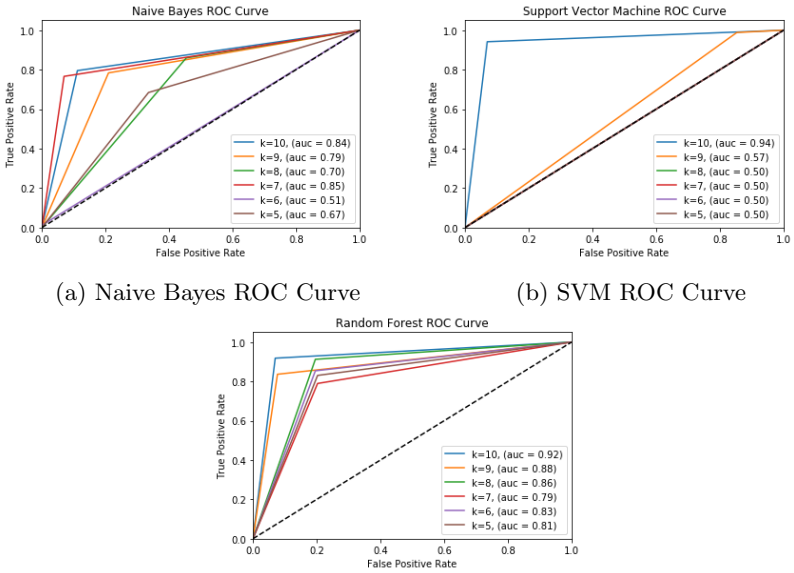
Fig. 5: Average Build Time of models by length of k. While build time remains stable for lengths of k for 5 through 9, it increases significantly with length of 10.

with the isolate number on the x axis where 1 through 178 are susceptible and 179 through 392 resistant. We can see that there is separation between the leftmost susceptible isolates and the later resistant isolates.

6 Analysis

All work on this study is conducted on an Intel Core i5 machine with 12 GB memory. We have built this model in an Python 3.7 environment. We use a

ROC Curve for NB, SVM, and RF Classifiers for k 5 through 10



(a) Naive Bayes ROC Curve

(b) SVM ROC Curve

(c) Random Forest ROC Curve

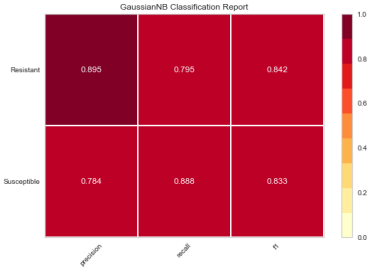
Fig. 6: The ROC Curve for NB, SVM, and RF Classifiers for k 5 through 10 show improving accuracy for each classifier as the length of k increases. While SVM provides a higher accuracy for length of k = 10, Random Forest provides the most stable accuracy for all lengths.

10-Fold stratified shuffle cross validation with an 80/20 test train split for each classifier. While SVM generates a model with an average of 94.4% with k-mer length of 10, Random Forest provides the most stable and consistent accuracy for all lengths.

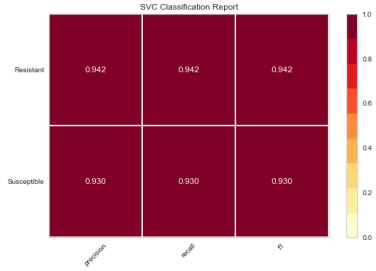
Build times increase as the length of k increases but is manageable through a length of k equal to 10. We find that the performance of the model becomes untenable on a machine without optimized hardware when a build length of k greater than 10.

Finally, as seen in the GMM plot, there are some areas that appear to show a high degree of accuracy in predicting susceptibility and resistance. The isolates with the highest frequency with counts of 6 and 7 of the feature that are in yellow classify perfectly for resistance. The same is true for the isolates with the lowest frequency with counts of 0 and 1 which classify all but two isolates correctly as being susceptible.

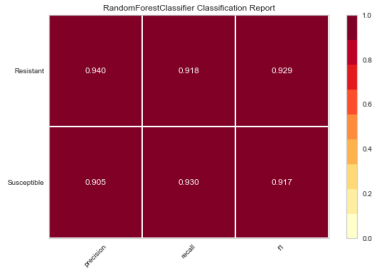
Classification Reports for Naive Bayes, SVM, and Random Forest for k =10



(a) Naive Bayes Classification Report



(b) SVM Classification Report



(c) Random Forest Classification Report

Fig. 7: The Classification Reports for Naive Bayes, Support Vector Machine, and Random Forest Classifiers where k = 10. The reports show similar metrics for both SVM and Random Forest generate similar results while Naive Bayes produces results that lag behind.

7 Ethical Considerations

Antibiotics are a shared global resource where there is a limited supply that can effectively be used for treatment. If their use is not managed properly, they can be depleted to the point where there are no treatments available for resistant infections.

While the treatment of patients who have antimicrobial resistant bacteria in any way possible seems like the logical approach that most would take, it is not the approach that doctors do or should take. Some of the causes cited for an increase in antimicrobial resistance are the misuse and overuse of antibiotics[1]. As such, a doctor may withhold antibiotics unless the patient is considered high risk. However, there is no clear definition of what qualifies as high risk. Furthermore, there is almost certainly variance in opinions of when antibiotics should be used between doctors.

Another possibility is that legislators or hospitals might take action to either reduce the amount of antibiotics given. They could place restrictions on the amount of antibiotics prescribed or make specific restrictions on when they

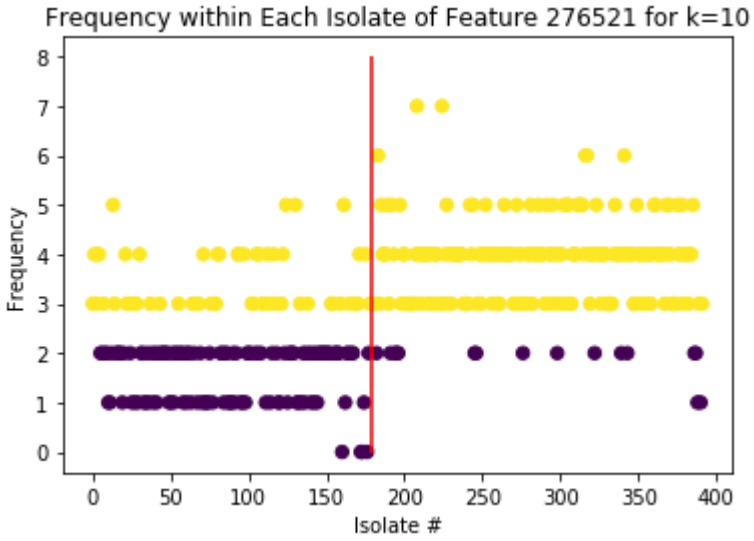


Fig. 8: Gaussian Mixture Model plot of the number of occurrences within each isolate of the most important feature. There is visible separation for isolates with both lower frequency and higher frequencies.

may be given. The Center for Disease Control (CDC) has already given a recommendation in support of antibiotic stewardship programs to offer guidance to health care practitioners and facilities to offer guidance in the use of antibiotics. [4]. While this may be a way to ensure misuse or overuse of antibiotics, this could take the decision out of the hands of doctors who would be able to make decisions on a case by case basis whether they are appropriate.

Another reason cited for the increase in antimicrobial resistance is the use of antibiotics in agriculture. The transmission of antimicrobial resistant microbes from animals to humans has been an issue known since the 1960s [30]. However antibiotics have been widely used to promote the growth of livestock and infection prevention[30]. When antibiotics are included in the feed, the livestock that are eating it constantly have their bacteria exposed to the antibiotics on a constant basis. It will kill off any bacteria that is susceptible but will leave behind resistant bacteria. Resistance to colistin, which is the antibiotic of last resort for multi-drug-resistant infections, by the mcr-1 gene has been attributed to livestock in China where colistin has been used as a therapeutic drug and food additive [31]. While the first case was identified in China in 2011, it has now been identified in multiple countries across continents. Furthermore, it has been found in a number of environmental settings, including public urban beaches in Brazil and the Haihe River in China. Although the use of colistin as a food additive for livestock has now been banned by China, the appearance in these environments illustrates the danger in overusing antibiotics.

While Europe has banned the use of antibiotics in livestock feed, the US has no such restriction. Instead, the US FDA offers guidance in the use of antimicrobial drugs in livestock feed in the veterinary feed directive (VFD) [32]. The FDA is working to phase out the use of the antimicrobials for production purposes in the promotion of growth, but they support the continued treatment of sick animals when necessary with their use. This issue relates to the humane treatment of animals as well as the ability of food producers to make a living, but it all needs to be weighed against the "greater good."

Cleaning and hygienic products that are antibacterial also contribute to the increase in antimicrobial resistance[1]. This is due in part to insufficient guidelines in the personal care industry where home hygiene products contain high concentration of antibacterial ingredients [33]. While these products such as antibacterial hand sanitizers have grown in popularity and offer a valuable product in helping people clean their hands or other surfaces, there is a trade-off to society at large as they are contributing to this issue of antimicrobial resistance. As a result, hospitals are taking different approaches to prevent the spread of AMR such as Probiotic Cleaning Hygiene System (PCHS) which is a probiotic based approach to sanitation[34].

Yet another factor is a reduction in the development of new antibiotics. This is largely due to economic and regulatory obstacles. Pharmaceutical companies see antibiotics as a poor investment as the cost to develop them is relatively high when compared to the revenue generated. This is in part because of the restraint that doctors have been advised to use in prescribing them to patients due to this same crisis. Meanwhile, the Food and Drug Administration (FDA) has changed standards for clinical trials which have made them more challenging.[1]. Do these companies have an obligation to society to dedicate resources to the development of antibiotics in spite of the lack of profitability?

Another proposal to both improve identification of AMR and reduce the improper prescription of antibiotics is collection of data from patients that have had antibiotic resistant infections[4]. While this is critical for ensuring that researchers have sufficient data for testing and developing new antibiotics, it is critical that patient information is protected.

8 Conclusions

In prior work, Dr. Santerre achieved accuracies of 93-98% with Random Forest, 51-84% with Naive Bayes, and 84-98% with SVM with a length of k = 10 on models created using different AMR phenotypes[35]. Our current research supports prior research that the approach for using k-mers to identify gene regions responsible for AMR is a viable solution. Further, by using an entirely different data set for AMR with Neisseria Gonorrhea supports the probability that this approach will work for biologists. These results democratize research into AMR as supercomputers are not required to pursue research. Moreover, this approach is one that may be used to investigate outbreaks in developing countries where computing resources may be limited but time is of the essence.

While prior works provide results suggesting k-mer length of more than 10 are needed to provide a high level of accuracy, we see that a model of length 10 can generate a model with an average accuracy of 92% . These results may provide sufficient accuracy as it offers the ability to further expand access of models investigating the AMR datasets available through PATRIC or other similar sets using short reads.

9 Future Work

While this current work has validated the approach in prior works of using k-mers of lengths and identified the regions most likely responsible for AMR, we have not completed the alignment step. Alignment against a well curated reference genome of Neisseria Gonorrhea bacteria would provide better insights into the specific gene regions that are responsible for AMR.

References

1. Ventola, C. Lee. "The antibiotic resistance crisis: part 1: causes and threats." Pharmacy and therapeutics 40.4 (2015): 277.
2. Hegreness, Matthew, et al. "Accelerated evolution of resistance in multidrug environments." Proceedings of the National Academy of Sciences 105.37 (2008): 13977-13981.
3. Perez, Katherine K., et al. "Integrating rapid diagnostics and antimicrobial stewardship improves outcomes in patients with antibiotic-resistant Gram-negative bacteremia." Journal of Infection 69.3 (2014): 216-225.
4. Ventola, C. Lee. "The antibiotic resistance crisis: part 2: management strategies and new agents." Pharmacy and Therapeutics 40.5 (2015): 344.
5. Clifton, Soazig, et al. "Prevalence of and factors associated with MDR Neisseria gonorrhoeae in England and Wales between 2004 and 2015: analysis of annual cross-sectional surveillance surveys." Journal of Antimicrobial Chemotherapy 73.4 (2018): 923-932.
6. Peters, Joanna, et al. "Whole genome sequencing of Neisseria gonorrhoeae reveals transmission clusters involving patients of mixed HIV serostatus." Sex Transm Infect 94.2 (2018): 138-143.
7. Bodie, M., et al. "Preventing the spread of extensively drug-resistant gonorrhea." (2019).
8. Forbes, Jessica D., et al. "Metagenomics: the next culture-independent game changer." Frontiers in microbiology 8 (2017): 1069.
9. Kelley, David R., et al. "Sequential regulatory activity prediction across chromosomes with convolutional neural networks." Genome research 28.5 (2018): 739-750.
10. Leung, Michael KK, et al. "Machine learning in genomic medicine: a review of computational problems and data sets." Proceedings of the IEEE 104.1 (2016): 176-197.
11. Macas-Garca, Laura, et al. "A study of the suitability of autoencoders for preprocessing data in breast cancer experimentation." Journal of biomedical informatics 72 (2017): 33-44.
12. Antonopoulos, Dionysios A., et al. "PATRIC as a unique resource for studying antimicrobial resistance." Briefings in bioinformatics (2017).

13. Davis, James J., et al. "Antimicrobial resistance prediction in PATRIC and RAST." Scientific reports 6 (2016): 27930.
14. Santerre, John W., et al. "Machine learning for antimicrobial resistance." arXiv preprint arXiv:1607.01224 (2016).Santerre, John William. Machine Learning for the Genotype-to-Phenotype Problem. 2018.
15. Brooks, Eric L., and Ryan D. Kappedal. "Compressive sampling for phenotype classification." 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2017.
16. Hamid, Md-Nafiz, and Iddo Friedberg. "Reliable uncertainty estimate for antibiotic resistance classification with Stochastic Gradient Langevin Dynamics." arXiv preprint arXiv:1811.11145 (2018).
17. Drouin, Alexandre, et al. "Interpretable genotype-to-phenotype classifiers with performance guarantees." Scientific reports 9.1 (2019): 4071.
18. Kavvas, Erol S., et al. "Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance." Nature communications 9.1 (2018): 4306.
19. Her, Hsuan-Lin, and Yu-Wei Wu. "A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains." Bioinformatics 34.13 (2018): i89-i95.
20. Drouin, Alexandre, et al. "Large scale modeling of antimicrobial resistance with interpretable classifiers." arXiv preprint arXiv:1612.01030 (2016).)
21. Arsenal, A. Shrinking. "Addressing the Threat of Drug-Resistant Gonorrhea." (2016).
22. Town, Katy, et al. "Neisseria gonorrhoeae molecular typing for understanding sexual networks and antimicrobial resistance transmission: A systematic review." Journal of Infection 76.6 (2018): 507-514.
23. Hook III, Edward W., and Robert D. Kirkcaldy. "A brief history of evolving diagnostics and therapy for gonorrhea: lessons learned." Clinical Infectious Diseases 67.8 (2018): 1294-1299.
24. Zankari, Ea, et al. "Identification of acquired antimicrobial resistance genes." Journal of antimicrobial chemotherapy 67.11 (2012): 2640-2644.
25. Mitra, M. "DNA Sequencing Basics and its Applications." SCIOL Genet Sci 1 (2018): 80-84.
26. Poole, Keith. "Mechanisms of bacterial biocide and antibiotic resistance." Journal of Applied Microbiology 92 (2002): 55S-64S.
27. Sievers, Aaron, et al. "K-mer content, correlation, and position analysis of genome DNA sequences for the identification of function and evolutionary features." Genes 8.4 (2017): 122.
28. Ghandi, Mahmoud, et al. "Enhanced regulatory sequence prediction using gapped k-mer features." PLoS computational biology 10.7 (2014): e1003711.
29. Zhang, Xian-Da. Matrix analysis and applications. Cambridge University Press, 2017.
30. Holmes, Alison H., et al. "Understanding the mechanisms and drivers of antimicrobial resistance." The Lancet 387.10014 (2016): 176-187.
31. Mediavilla, Jos R., et al. "Colistin-and carbapenem-resistant Escherichia coli harboring mcr-1 and blaNDM-5, causing a complicated urinary tract infection in a patient from the United States." MBio 7.4 (2016): e01191-16.
32. US Food and Drug Administration. "New animal drugs and new animal drug combination products administered in or on medicated feed or drinking water of food-producing animals: recommendations for drug sponsors for voluntarily aligning product use conditions with GFI #209. Guidance for Industry# 213. 2013." (2014).

33. Zaman, Sojib Bin, et al. "A review on antibiotic resistance: alarm bells are ringing." Cureus 9.6 (2017).
34. Elisabetta, Caselli, et al. "Impact of a probiotic-based hospital sanitation on antimicrobial resistance and HAI-associated antimicrobial consumption and costs: a multicenter study." (2019): 501-510.
35. Santerre, John William. Machine Learning for the Genotype-to-Phenotype Problem. 2018.

Appendix

Code to Create Master Dictionary

```
master_dict = {}
# keys are kmers
# values are a dictionary
# keys are position  of isolate in all_isolates
# Value is count value of count
#k = 7 # kmer size
k_len = 10

#Create variable with isolate name and length
isolate_info = []

for idf, a_file in enumerate(all_isolates):
  with open(a_file, 'r') as f:
    data = list(csv.reader(f))
  data = [d for d in data if d]
  print(a_file)
  contigs = []
  tmp = ''
  for x in range(len(data)):
    if data[x][0][0]=='>':
      contigs.append(tmp)
      tmp = ''
    else:
      tmp += data[x][0]
  #if kmer known
  #  if kmer already in isolate
  #  else show kmer to isloate
  #else:
  #  know kmer
  #  know kmer in isolate
  for contig in contigs:
    end_pos = len(contig)-k_len
    for x in range(end_pos):
      kmer = contig[x:x+k_len]
```

```
        if kmer in master_dict:
          if idf in master_dict[kmer]:
            master_dict[kmer][idf]+=1
          else:
            master_dict[kmer][idf] = 1
        elif 'n' not in kmer:
          master_dict[kmer] = {}
          master_dict[kmer][idf] = 1
    print(sum([len(c) for c in contigs]))
    isolate_info.append(a_file + '␣' +
        str(sum([len(c) for c in contigs])))
```

Code to Create k-mer Matrix

```
M = np.zeros([len(all_isolates), len(master_dict)],
    dtype=np.int64)
for column, kmer_string in enumerate(master_dict.keys()):
    for row in master_dict[kmer_string].keys():
        #print(master_dict[kmer_string].keys())
        M[row, column] = master_dict[kmer_string][row]
sus_files = [f for f in os.listdir(sus_dir)
    if f[-1] != 'b']
res_files = [f for f in os.listdir(res_dir)
    if f[-1] != 'b']
total_number_isolates = len(sus_files) +
    len(res_files)
all_isolates =
    ['./SUS/'+ x for x in sus_files] +
      ['./RES/'+ x for x in res_files]

L = [1   if x >= len(sus_files)
    else 0 for x in range(len(sus_files) +
        len(res_files)) ]
L = np.array(L)
L = L.astype(int)
```