A Network Medicine Approach to Predict Antibiotical Resistance

By

Bruna Fernanda Fistarol

Fundação Getulio Vargas
Escola de Matemática Aplicada

A thesis submitted in fulfillment of the requirements for
the degree of M.Sc.

Supervisor: Alberto Paccanaro

May 2023

Abstract

Here goes the abstract ...

Acknowledgements

Here goes the dedication...

Table of Contents

List of Figures

FIGURE **Page**

List of Tables

TABLE **Page**

1

Introduction

The discovery of penicillin in 1928 by the Scottish scientist Alexander Fleming started a new era in medicine, saving millions of lives and making possible to treat several infectious diseases through antibiotics [1]. However, these drugs have their efficiency jeopardized by the development of resistant bacteria. When he won the Nobel Prize in Medicine for his discovery, the scientist himself warned in his speech about the possibility of the emergence of resistant bacteria [2].

According to World Health Organization (WHO), antibiotic resistance is one of the biggest threats to global health. It leads to higher medical costs, prolonged hospital stays, and increased mortality. Any person with any age can be affected. It is estimated that, in 2019, about 4.95 million people died as a direct or indirect result of antibiotic resistance [3].

Clinically, antimicrobial resistance (AMR) in pathogenic bacteria is the ability of this microorganism to resist the effect of a medication previously effective to treat the infection. This phenomenon is caused by genetic mutations or horizontal transfer of AMR genes between bacteria, both stimulated by selective pressure due the use of antibiotics [4].

Although resistance to antibiotics occurs spontaneously as a result of natural selection, this process is accelerated due to the indiscriminate and excessive use of antibiotics

in the population, hospitals and in the animal sector. In addition to factors that influence infectious diseases, like the lack of clean water and sanitation and inadequate infection prevention, there is a contribution to the development and spread of resistant bacteria due to antibiotics [5, 6]. Furthermore, there is a severe shortage in the discovery of new antibiotics. The reasons for this range from financial and regulatory obstacles to the scientific difficulty in developing a drug of this class [7, 8]. Once there is no straightforward solution for this problem, it is necessary to understand antimicrobial resistance by studying related mechanisms in order to create tools that can help to make decisions in the treatment of an infection.

The most effective way to identify the organism causing an infection and its antibiotic resistance profile is through the Antimicrobial Susceptibility Test (AST). Through this method, patient samples are cultured *in vitro* in the presence of the antibiotics of interest in order to identify the susceptibility or resistance of the microorganism by the formation of bacterial colonies. However, the availability of test results may take a few days. This time is required to have sufficient microorganisms to determine the minimum antibiotic concentration necessary to inhibit the growth of colonies, besides identify the microorganism. In urgent cases, it is not always possible to wait for these results to start treatment. Furthermore, the process of cultivating microorganisms *in vitro* is not always simple or feasible. Thus, the choice of medications used often depends on the doctor's judgment. An inadequate choice, both of the drug and the dose, can lead to the development of resistant strains of bacteria.

Strategies to identify the antimicrobial resistance profile without the need for bacterial culture can help to obtain a faster diagnosis. With the current technology associated to DNA sequencing, computational methods can be developed to analyze bacterial genomes, generating conclusions that can support the choice of antibiotics used in a given clinical case.

To sequence bacterial genomes without the need to culture bacteria, it is necessary to resort to Molecular Biology protocols that are unlikely to provide the complete sequence of the microorganism, especially when the etiological agent is not abundant in the sample used in the sequencing, since it contains a great diversity of genetic material, either from the patient himself or from other microorganisms different from the infectious agent. Furthermore, the nature of resistance genes can also make their identification difficult in this type of protocol, once they are common associated to plasmids and are not conserved. Highly conserved genes are more trivial to identify through phylogenetic analysis [9].

Although there are numerous computational studies aiming to develop algorithms

that recognize patterns in bacterial DNA sequences determining aspects that characterize the resistance phenotype, most of these studies are based on the complete sequence of the microorganism [10–16]. The existence of impasses related to the complete coverage of the bacterial genome sequence when the sequencing protocol does not require additional experiments motivates the investigation of the predictive potential of partial sequences of bacterial DNA. Recently, it was shown that conserved DNA sequences can provide a prediction regarding the susceptible or resistant phenotype of bacteria to a given antibiotic [17]. Thus, the present study proposes to investigate whether the predictive power of conserved genes is related to the proximity of interaction with genes related to bacterial resistance.

The central idea of the method is to diffuse antimicrobial resistance genes in the protein interaction network in order to map relevant conserved genes with respect to diffusion. This idea is based on the hypothesis that proteins that participate in complexes with other proteins from genes that suffered mutation and conferred antimicrobial resistance to bacteria are more likely to fix mutations related to antimicrobial resistance.

In Chapter 2, we will see how a previous study used conserved bacterial genome sequences to make predictions of the resistance phenotype. In Chapter 3, two ideas will be introduced in order to select conserved genes to repeat the experiments described in Chapter 2. The results will be presented in Chapter 4 and discussed in Chapter 5. The conclusion of the study is made in Chapter 6.

2

Prediction of antimicrobial resistance using conserved genes

Recently, Nguyen et. al. built a machine learning model using the complete sequence of *Klebsiella pneumoniae* strains to predict the minimum inhibitory concentration of 20 antibiotics in antimicrobial susceptibility tests [18]. In order to approach the case as a regression or multi-class classification problem, several popular machine learning algorithms were tested using their default parameters and oligonucleotide k-mers as features. Based on the best accuracy obtained and computational resources required, the authors decided to use XGBoost, a scalable machine learning system based on the Gradient Boosting Decision Tree structure.

In that study, Nguyen et. al. noted that the accuracy of the predictions remained practically identical when the genes known to be related to antibacterial resistance were removed from the analyzed sequence. Thus, the study suggested that partial bacterial genome sequences with no established association with bacterial resistance also had high predictive power.

Motivated by this discovery, the group developed a study of antimicrobial resistance prediction using common conserved genes among members of the same species and unrelated to the antibacterial resistance phenotype [17]. The model was constructed in order to classify the phenotype of a specific bacterium as susceptible or resistant given

its conserved genome sequence. The same methodology of the previous study was used to build the model. The next section briefly explains concepts involved behind the developed model.

2.1 Model Background

2.1.1 Decision Tree for Classification

A decision tree for classification is a sequential model which logically combines a sequence of simple tests. Each test compares a numeric attribute against a threshold value against a set of possible values. The goal of this process is to create a model that can accurately predict the class or category of a given input [19].

The decision tree works by repeatedly splitting the dataset into subsets based on the value of a chosen feature. Each split results in a new node in the tree, and the process continues until a stopping criterion is met. Each leaf node of the tree represents a class, and the path from the root node to a leaf node represents the decision rules used to classify an input [20].

In order to minimize data classification error, the model's training set is used to define how the region of the prediction space is divided. As an example, figure 2.1 shows X_1 and X_2 as features and t_i, with $i \in \{1,2,3,4\}$, is a threshold value splitting the prediction space in 5 regions [19].

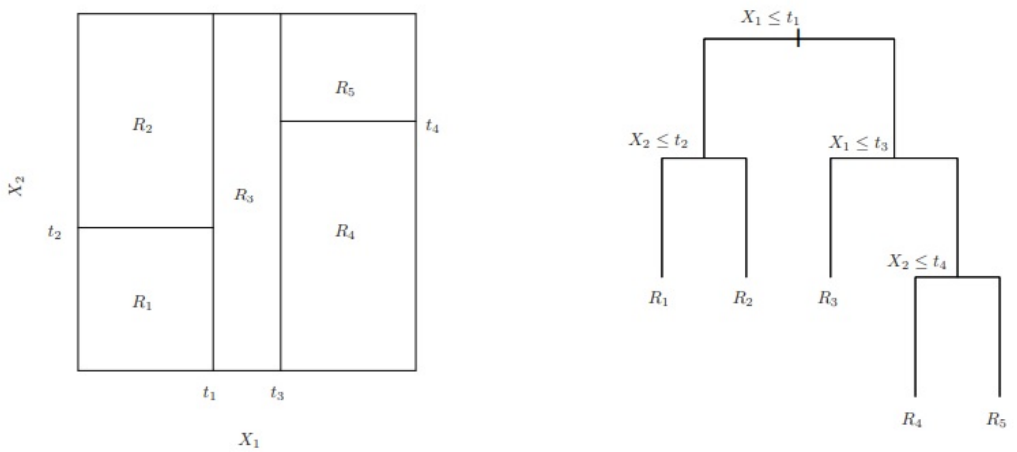

Figure 2.1: The output of recursive binary splitting on a two-dimensional example and a tree corresponding to this partition, respectively.

Despite being quite interpretative models, decision trees by themselves do not perform as well as other supervised learning methods in terms of accuracy and suffer from high variance. For this reason, they are integrated with techniques capable of improving their performance.

2.1.2 Gradient Boosting

Boosting is a general approach that can be applied to many statistical learning methods for reducing the variance. It is an ensemble technique, which means it uses multiple learn algorithms (weak learners) to obtain a better predictive performance (strong learner). In this case, the idea is to combine a large number of decision trees where each tree is dependent on prior trees. Given the current model, a decision tree is constructed using the residuals from the model, hence, each tree attempts to minimize the errors of previous tree. Every time a tree is added to the previous one, they are weighted in order to give a higher weight to misclassified input data. Thus, future weak learners focus more on the examples that previous weak learners misclassified.

This method is called **gradient boosting** when the algorithm uses a gradient descent optimization process to update the weights of the models in the ensemble, in order to minimize the overall prediction error. At each iteration, the algorithm calculates the negative gradient of the loss function with respect to the predicted values, and trains a new model to predict the residuals. The residuals are then added to the previous predictions, and this process is repeated until convergence.

2.1.3 XGBoost

XGBoost (eXtreme Gradient Boosting) is an open-source software library that provides an efficient and effective implementation of the gradient boosting framework. It is specifically designed to optimize large-scale, performance-critical machine learning problems. It can be used for a variety of tasks such as classification, regression, and ranking, and it is particularly effective when dealing with large datasets.

XGBoost is built on top of the gradient boosting algorithm and is an extension of the traditional gradient boosting method. The key features of XGBoost that make it stand out from other gradient boosting libraries include: a regularization term to prevent overfitting; a parallel construction of trees using all of the cores; capability of handling missing values; handling categorical variables as input without one-hot encoding; and a

technique to handle large dataset and large number of trees called Weighted Quantile sketch [21].

2.1.4 Oligonucleotide k-mers

Oligonucleotide k-mers of DNA sequences are sub-sequences containing k oligonucleotides. For example, taking a step size equals to one nucleotide, the sequence ATGC has the 3-mers ATG and TGC. This structure is used to describe a specific bacterial strain as a vector of features containing a count of different k-mers given a specific set of genes.

Figure 2.2: 15-mer oligonucleotide framework.

The big advantage of k-mer-based methods compared to alignment-based methods is the better scaling of computation times with sequence length [22].

2.2 Model Dataset

Genome and laboratory-derived antimicrobial susceptibility test were downloaded from PATRIC (PathoSystems Resource Integration Center) database. PATRIC is a publicly available resource that provides integrated information on bacterial pathogens, including genome sequences, functional annotation, and comparative analysis tools. The database is designed to support the research community in the study of pathogenic bacteria, and can be used for a wide range of applications, including the identification of drug targets, the development of diagnostic tests, and the understanding of bacterial evolution and pathogenesis. The database is maintained by the National Institute of Allergy and Infectious Diseases (NIAID) and is freely accessible to the public [23].

2.2.1 Protein families (PLFam)

The analyses were based on the protein-encoding genes that are shared among members of the same species. Different subspecies of a bacteria (also referred here as "strains") share an essentially analogous genome with respect to the proteome. PATRIC database relates each protein to a family of proteins, called PATRIC Local Family (PLFam), which are similar regarding their functions in the same taxonomic genus [24]. All genomes were reannotated so that they had the same set of protein family calls. Also, all families with a PATRIC annotation associated with AMR were removed from the study.

2.2.2 Conserved genes

Two criteria were used by the authors to define core gene sets. First, for each family, the average nucleotide length was computed for the corresponding genes. Any family member that had a total nucleotide length that was less than half of the average length, or that was 50% longer than the average length, was excluded. This helped to eliminate duplicate genes, partial genes, and mixtures of genes encoding single and multi-subunit proteins. Next, any family whose members represented less than 99% of the genomes of the set was excluded.

2.2.3 Antimicrobial resistance phenotype

Each strain is associated to an antimicrobial resistance phenotype. This data is available on PATRIC collection as laboratory derived values such as a minimum inhibitory concentration, or being susceptible, intermediate, or resistant determinations. Based on break-point values of minimum inhibitory concentration from the Clinical and Laboratory Standards Institute (CLSI) and the European Committee on Antimicrobial Susceptibility Testing (EUCAST), AMR phenotypes published as minimum inhibitory concentrations were converted to susceptible/resistant determinations. Classification was not performed on intermediate phenotypes because they are underrepresented.

2.3 Model Design

The model designed by Nguyen et. al. was constructed in order to classify the phenotype of a specific strain as susceptible or resistant given a set of conserved genes. Given some species of bacteria, which has several strains, the authors choose at random a group of

conserved genes in terms of protein families. Every strain is described with relation to the same protein families. Figure 2.3 shows how a species of bacteria has several strains associated to a susceptible or resistant phenotype (S or R) with relation to an antibiotic, each one having a set of chosen conserved genes.

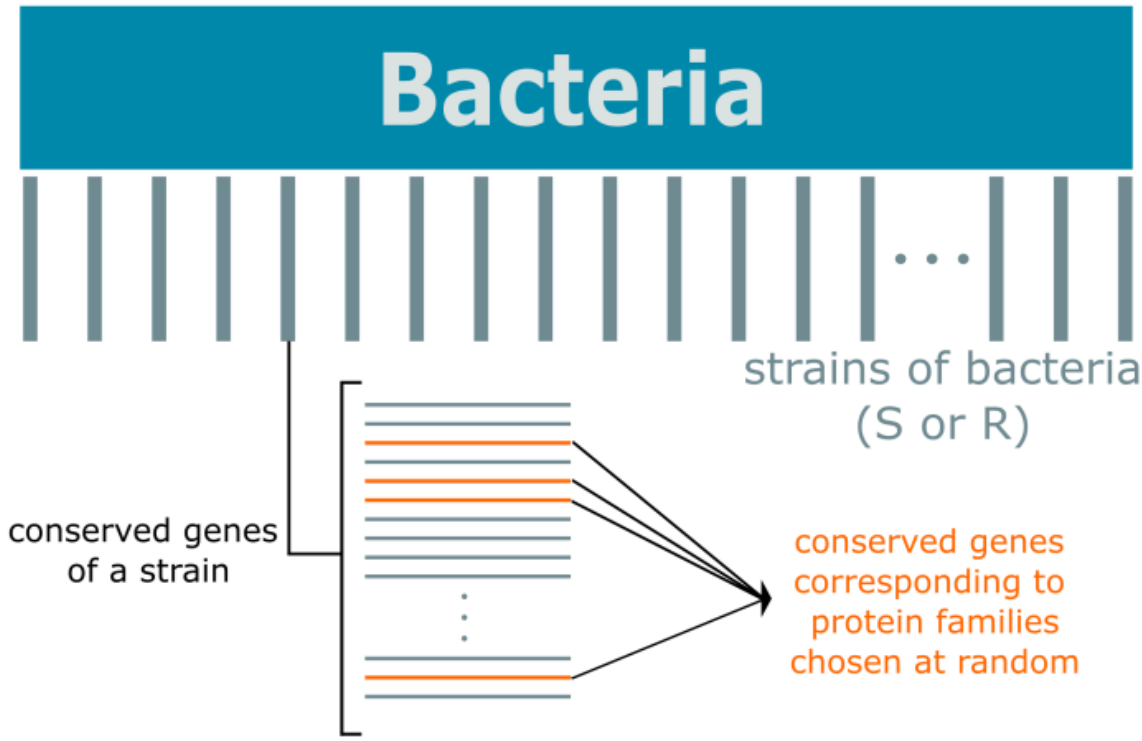

Figure 2.3: Framework of the data used by Nguyen et. al.

The authors described a strain as a vector counting, in lexicographic order, all 15-mers with relation to the selected genes, i.e., 15-mer frequencies are features of the model. Once each strain has a class associated (susceptible or resistant), this data structure is used as model input, as it is illustrated by figure 2.4.

The authors made several experiments using sets with 25, 50 100, 250 and 500 conserved genes to run the model. Four species of bacteria were chosen: *Klebsiella pneumoniae, Mycobacterium tuberculosis, Salmonela enterica* and *Staphylococcus aureus*. For all four species, models built from 25 genes range in their average F1 scores from 0.75 [0.73–0.77, 95% confidence interval] for *S. enterica* to 0.80 [0.78–0.81, 95% confidence interval] for *K. pneumoniae* (Fig 1). The F1 scores increase as the set size increases, with the models built from 500 genes having F1 scores ranging from 0.84 [0.81–0.86, 95%

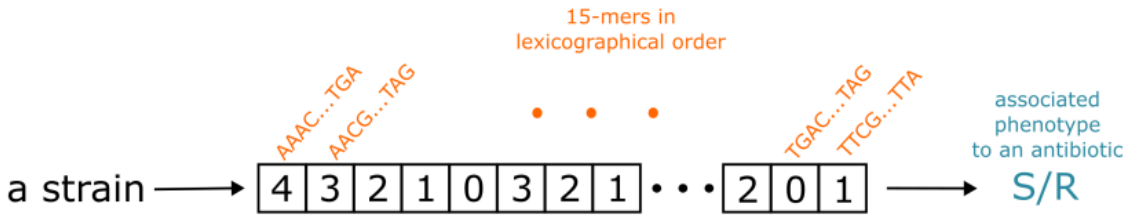

Figure 2.4: Structure of data used as features by the model

confidence interval] for *M. tuberculosis* to 0.89 [0.86–0.90, 95% confidence interval] for *S. aureus*.

Besides to conclude that AMR phenotypes can be predicted from sets of core genes, the authors also made additional experiments to emphasize that high accuracies do not appear to be the result of overfitting, memorization, strain-specific SNPs, or imbalances in sampling or phylogeny.

However, there are no criteria to create the sets of conserved genes used to run the experiments - they are constructed at random from the complete set of conserved genes, which brings the idea to study the possibility to select these genes in a better way.

3

SELECTING CONSERVED GENES BASED ON INFORMATION FROM PROTEIN INTERACTION NETWORK

Given the fact that conserved genes are chosen at random as it was explained on the previous chapter, we would like to find if there is way to choose these genes in order to improve the prediction compared to a random selection. In this sense, two methods are proposed based on ideas from network medicine.

3.1 Network Medicine

Network medicine is an emerging interdisciplinary field that aims to understand and analyze the complexity of diseases from a system perspective. It involves integrating multiple sources of biological and clinical data to build comprehensive models to identify novel therapeutics [25].

One of key tools of network medicine is to analyze complex networks of interactions between different biological components. Graphs, one of the most general representations of discrete metric spaces, are natural data structures to model such components, where nodes are objects and the relation between them are edges [26].

Network medicine relies in the idea that a disease is rarely a consequence of an abnormality on a single gene, but reflects the perturbations of the complex interactome network, which is the whole set of molecular physical interactions between biological entities in cells and organisms [25].

Regarding that the previous study was based on protein-encoding genes, a protein interaction network will be used to analyze distance between conserved genes and AMR genes as a graph where proteins are nodes and an edges is a physical interaction between two proteins.

3.2 Idea

This study proposes ways to select conserved genes using information from bacterial protein interaction network. The objective is to verify whether incorporating this information in the choice of conserved genes can provide a better selection than a random choice. This is the first approach using Network Medicine for this particular problem.

The approach is based on the hypothesis that conserved genes closer to genes previously associated with antimicrobial resistance may suffer mutations resulting from the incorporated resistance mechanism and, therefore, provide features capable of characterizing the susceptibility phenotype more precisely. For example, the highlighted conserved genes represented in figure 3.1 could have a better predictive power than choose a conserved gene at random.

The methods in this study were tested using data from *Salmonella enterica*.

3.3 Model Background

3.3.1 Kernel Methods on Graphs

Kernel-based methods are a set of techniques for analyzing complex data sets, including those arising in network science and computational biology. They offer a natural framework to study similarity between two nodes in a graph. In particular, they can be used to analyze and model the relationships between different components of biological networks - in this case, proteins.

The definitions below are based on [27], [28] and [29].

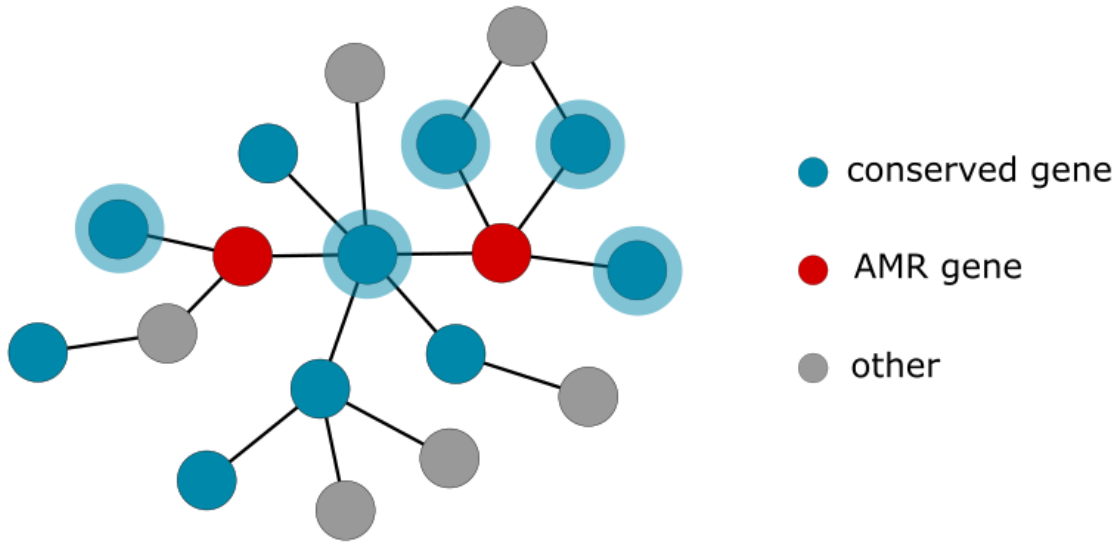

Figure 3.1: Illustrative example of a bacterial protein interaction network where conserved genes interacting with AMR genes are highlighted.

A kernel function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a measure of similarity between objects. It implicitly construct a mapping $\phi : \mathcal{X} \mapsto \mathcal{H}_k$ to a Hilbert space $\mathcal{H}_k$ in which the kernel appears as the inner product

$$k(x, x') = \langle \phi(x), \phi(x') \rangle .$$

The function k must satisfy two mathematical requirements: it must be symmetric, that is, $k(x, x') = k(x', x)$, and positive semi-definite. For finite graphs, the kernel can equivalently be specified by an $n \times n$ matrix K, with $K_{x_i, x_j} = k(x_i, x_j)$

The function of the kernel is to provide a global similarity metric, whereas graphs incorporate information on local similarity. It must be able to express the degree of similarity between any two examples x, $x' \in \mathcal{X}$ with fine distinctions in the degree to which x and x' are distant from each other in the graph. Therefore, the challenge is to define a kernel that captures the semantics inherent in the graph structure but at the same time is reasonably efficient to evaluate.

3.3.1.1 Laplace Operators

An undirected unweighted graph G consists of a set of vertices V numbered 1 to n, and a set of edges E (i.e., pairs (i, j) where $i, j \in V$ and $(i, j) \in E \iff (j, i) \in E$). The adjacency

matrix of G is an $n \times n$ real matrix W, with $W_{ij} = 1$ if i and j are neighbors (which is denoted as $i \sim j$), and 0 otherwise (by construction, W is symmetric and its diagonal entries are zero).

The adjacency matrix is not the only matrix associated with undirected unweighted graphs. Let D be an $n \times n$ diagonal matrix with $D_{ii} = \sum_j W_{ij}$. The **Laplacian** of G is defined as $L := D - W$ and **Normalized Laplacian** is

$$\widetilde{L} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}.$$

It is known from spectral graph theory that $\widetilde{L}$ is symmetric and positive semi-definite.

L and $\widetilde{L}$ can be regarded as linear operators on functions $\boldsymbol{f} : V \mapsto \mathbb{R}$, or, equivalently, on vectors $\boldsymbol{f} = (f_1, f_2, \cdots, f_n)^\top$. L could be also defined as

$$\langle \boldsymbol{f}, L\boldsymbol{f} \rangle = \boldsymbol{f}^\top L \boldsymbol{f} = -\frac{1}{2} \sum_{i \sim j} (f_i - f_j)^2 \quad \text{for all } \boldsymbol{f} \in \mathbb{R}^n.$$

3.3.1.2 Regularization via the Graph Laplacian

The fact that L induces a semi-norm on $\boldsymbol{f}$ which penalizes the changes between adjacent vertices indicates that it may serve as a tool to design regularization operators.

A class of regularization is defined as

$$\langle \boldsymbol{f}, P\boldsymbol{f} \rangle := \langle \boldsymbol{f}, r(\widetilde{L})\boldsymbol{f} \rangle$$

where $r(\widetilde{L})$ is understood as applying the scalar valued function $r(\lambda)$ to the eigenvalues of $\widetilde{L}$, that is,

$$r(\widetilde{L}) := \sum_{i=1}^{m} r(\lambda_i) \boldsymbol{v}_i \boldsymbol{v}_i^\top,$$

where $(\lambda_i, \boldsymbol{v}_i)$ constitute the eigensystem of $\widetilde{L}$.

3.3.1.3 Kernels

Giving a regularization matrix $P = r(\widetilde{L})$, the corresponding kernel is given by

$$K = r^{-1}(\widetilde{L}),$$

where the pseudo-inverse is taken wherever necessary. More specifically, if $(\lambda_i, \boldsymbol{v}_i)$ constitute the eigensystem of $\widetilde{L}$, we have

$$K = \sum_{i=1}^{m} r^{-1}(\lambda_i) \boldsymbol{v}_i \boldsymbol{v}_i^\top \quad \text{where we define } 0^{-1} \equiv 0.$$

3.3.1.4 *p*-Step Random Walk

The following choice of $r(\lambda)$ to the kernel function,

$$r(\lambda) = (aI - \lambda)^p \quad \text{with } a \geq 2,$$

provides the kernel

$$K = (aI - \widetilde{L})^p \quad \text{with } a \geq 2,$$

where a acts as a regularization term.

The value of K_{ij} is proportional of the probability of arrive i j from i in a random walk after p steps

This matrix is similar to the diffusion kernel. However, the fact that K involves only a finite number of products of matrices makes it much more attractive for practical purposes. In particular, entries in K_{ij} can be computed cheaply using the fact that $\widetilde{L}$ is a sparse matrix.

3.4 Proposed Methods

The following proposed approaches are based on information from a protein interaction network associated to a bacteria species, which is, in this case, an binary matrix of adjacency. All codes to construct gene sets are referenced in appendix A.

3.4.1 Naive selection of conserved genes

A very simple way to separate conserved genes in different groups is based on the shortest path to any AMR gene.

As an example, if conserved gene has a direct interaction with an AMR gene X (a path of length 1) and the shortest path to an AMR Y in the graph is 2, this conserved gene will be part of a group with label 1.

3.4.2 Selecting conserved genes based on kernel scores from *p*-Step Random Walk

The idea of this approach is to measure the additive effect of AMR genes through the diffusion of these genes in the interaction network between proteins.

Algorithm 1: Naive selection of conserved genes

```
Data: conserved_nodes, AMR_ nodes
Result: sets of conserved genes according to the shortest path to an AMR gene
for i in conserved_nodes do
    distance = inf;
    for j in AMR_nodes do
        if has a path between i and j then
            distance = min(shortest_path(i, j), distance)
        end
    end
    put i in a group corresponding to the variable distance
end
```

Applying the definition of normalized Laplacian on the kernel given by p-step random walk, we have

$$K = ((a-1)I + D^{-\frac{1}{2}} W D^{-\frac{1}{2}})^p,$$

where we set $a = 2$.

The matrix K provides a diffusion score from any to any node in the network. To obtain a vector with scores of each conserved gene, this matrix is multiplied by a binary vector where 1 indicates whether a gene is AMR related, which means AMR genes are the "source" of diffusion.

Conserved genes are sorted by score in decreasing order and a number of genes at the top of the list is selected to run the model.

We considered the values of $p = 2$ and $p = 1$. When we consider $p = 1$, we obtain a score proportional to the number of AMR genes that interact with conserved genes.

3.5 Model Dataset

3.5.1 *Salmonella enterica*

Antibiotic resistance in *Salmonella enterica* is a growing concern in the field of public health. *Salmonella enterica* is a gram-negative bacteria that can cause food poisoning and other infections in humans and animals. It is commonly found in raw or undercooked meats, eggs, and unpasteurized dairy products [30].

Although the previous model has been evaluated for four different species of bacteria, we choose to do this study with *Salmonella enterica* because its group of 500 conserved

genes has 11 genes annotated as AMR genes, which was founded after the publication of [17]. The number of strains for each antibiotic and susceptibility profile is given by 3.1.

In order to have the same labels for proteins in all databases, we chose to look a reference sequence of *Salmonella enterica*, the subspecies *Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*.

Antibiotic	Abbreviation	Susceptible	Resistant
Amoxicillin/Clavulanate	AUG	1009	310
Ampicillin	AMP	874	521
Cefoxitin	FOX	1087	267
Ceftiofur	TIO	1091	302
Ceftriaxone	AXO	1090	305
Chloramphenicol	CHL	1302	61
Gentamicin	GEN	1143	230
Kanamycin	KAN	201	33
Streptomycin	STR	256	543
Sulfisoxazole	FIS	712	594
Tetracycline	TET	581	810

Table 3.1: Counts of susceptible and resistant genomes used for *Salmonella enterica*.

3.5.2 Conserved Genes

The group of conserved genes used to repeat the experiments is a group of 500 genes given by the previous study (see appendix A). At the moment that this data was collected (December 1, 2018), conserved genes were chosen in a way where none of this genes was annotated as AMR related.

3.5.3 Antimicrobial Resistance Genes

The information about relation of a specific gene with a AMR phenotype was given by PATRIC database, searching for the specific subspecies of *Salmonella enterica* and accessing the specialty genes table, which contains antimicrobial resistance genes. This is necessary to localize AMR genes on the PPI.

3.5.4 Protein Interaction Networks

We are assuming that, although the conserved genes vary to the point of providing a high-performance prediction for the antimicrobial resistance phenotype, these changes are

subtle enough to assume that different strains share the same protein-protein interaction (PPI) network. That is, proteins associated with the same family are represented by the same node in the interaction network.

3.5.4.1 Direct Physical Interactions Between Proteins

To construct a physical protein interaction network, we used the prediction method described in [31], once *Salmonella enterica* is not well-studied in terms of protein physical interactions. The method works transferring useful experimental information from well-studied organisms though gene ontology. The dataset used was obtained from National Center for Biotechnology Information (NCBI) database, searching for protein sequences of the specific subspecies of *Salmonella enterica* (see appendix A).

3.5.4.2 Co-occurrence of Proteins in Physical Complexes

The organization of protein provides particularly strong evidence for their biological relationship. In this sense, STRING database assigns scores to pairs of proteins protein if the proteins show evidence of co-occurring in a complex, which means they can be directly or indirectly interacting [32]. These scores are calculated for selected evidence channels (text-mining) and aggregated into a combined physical interaction score and can be interpreted as the probability of two proteins being together in a gold standard complex.

The PPI can be found searching for *Salmonella enterica* specific subspecies at STRING database web page. On the download page, the PPI is referred as protein network data with physical links (physical subnetwork, scored links between proteins).

4

RESULTS

4.1 Prediction based on distance

Proceeding as it is described on algorithm 1 using the given set of 500 genes and the PPI network described on section 3.5.4.1, we get the following number of genes in each group according to the shortest path to any AMR gene:

Length of shortest path to AMR	Count of conserved genes
0	9
1	83
2	219
3	38

Table 4.1: Counts of conserved genes based on the shortest path to any AMR gene in the PPI network

There are 9 conserved genes with distance zero from an AMR gene, which means these genes were annotated as AMR genes after the data collection from the previous study. We will not consider these 9 genes to run experiments for conserved genes close to AMR genes. Instead, we will use these 9 genes as a separated batch to run the model.

The bar charts on figure 4.1 show the model performances in terms of F1 score (y axis), a metric which take into account not only the number of prediction errors that the model makes, but that also look at the type of errors that are made. This metric is scaled from 0 to 1, there 1 is the best score.

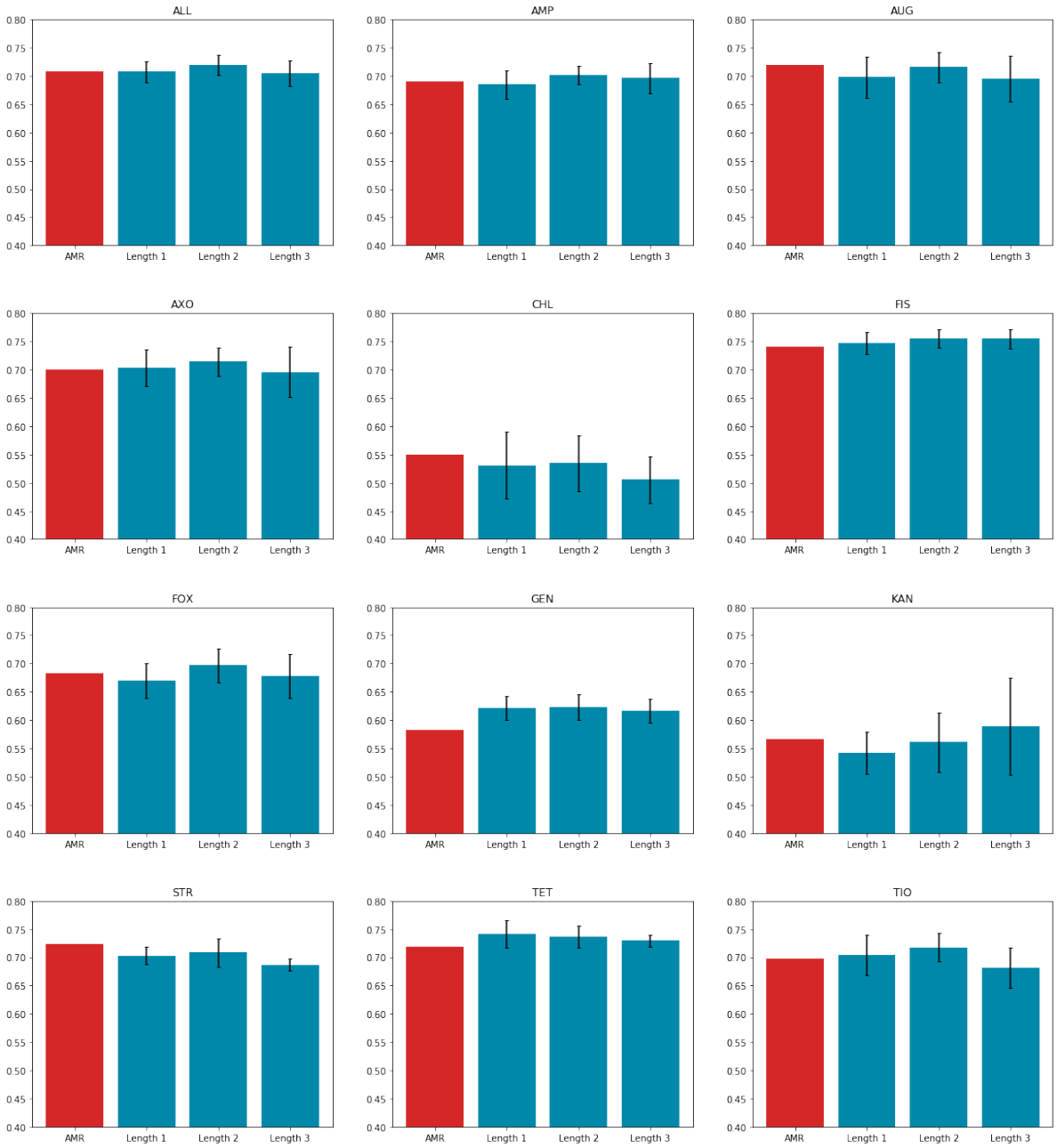

Figure 4.1: F1 scores obtained selecting genes based on distance from an AMR gene to a conserved gene.

The red columns indicates the F1 score obtained from 9 conserved genes annotated as AMR. The blue columns are the mean of several experiments. Basically, there are 83 genes with the length of the shortest path to any AMR gene equals to 1, which makes possible to do 9 non overlapping groups of 9 genes. Hence, the blue bars labeled as "Length 1" are the mean F1 score of 9 experiments. Bars labeled as "Length 2" and "Length 3" are the mean F1 score of 24 and 4 experiments, respectively.

According to the assumption made on this study, we expect a decreasing score as the conserved genes move away from AMR genes. However, this pattern is not observed in the charts.

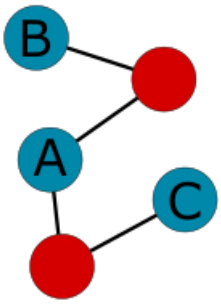

Figure 4.2: Interactions between protein showing a possible reason to not have a good result with a naive selection.

A possible reason is illustrated as an example on figure 4.2. The conserved genes A, B and C have length 1 from the AMR genes in red. However, while genes B and C interacts with one AMR gene, gene A is interacting with more than one AMR gene. This suggests that gene A could be a better choice for the prediction batch if the assumption is correct. This motivates the next approach, which makes a diffusion along the PPI using AMR genes as source nodes.

4.2 Prediction based on scores from kernel method

Kernel methods provide a way to summarize the contribution of all AMR genes. Using the method described on section 3.4.2 it is possible to get scores for each conserved gene in the PPI network in terms of a diffusion along the graph. The idea is to sort conserved genes in decreasing order based on these scores and use conserved genes on the top of the list to run the experiment.

In order to compare this result with a distribution given by a set of randomized experiments, we used several sets of randomized source nodes instead of AMR genes (which can include any node in the PPI network) having the same number of genes as there are AMR genes in the PPI network. The same methodology was used to run the experiments with randomized source nodes and obtained F1 scores were used to approximate normal distributions. All histograms in this section have 10 bins from the lowest to the highest F1 score (x axis). Additionally, the first plot on all figures (titled as "ALL") is a mean result of all antibiotics. The Gaussian curves are plotted with three standard deviation for each side around the mean.

4.2.1 Prediction using 2-step Random Walk in a network with direct physical protein interactions

Figure 4.3 shows histograms obtained from 30 randomized experiments using 20 conserved genes. From these values, normal distributions were approximated and plotted on figure 4.4. The vertical green line on figure 4.4 is the F1 score obtained using AMR genes as sources to calculate scores from 2-step kernel method.

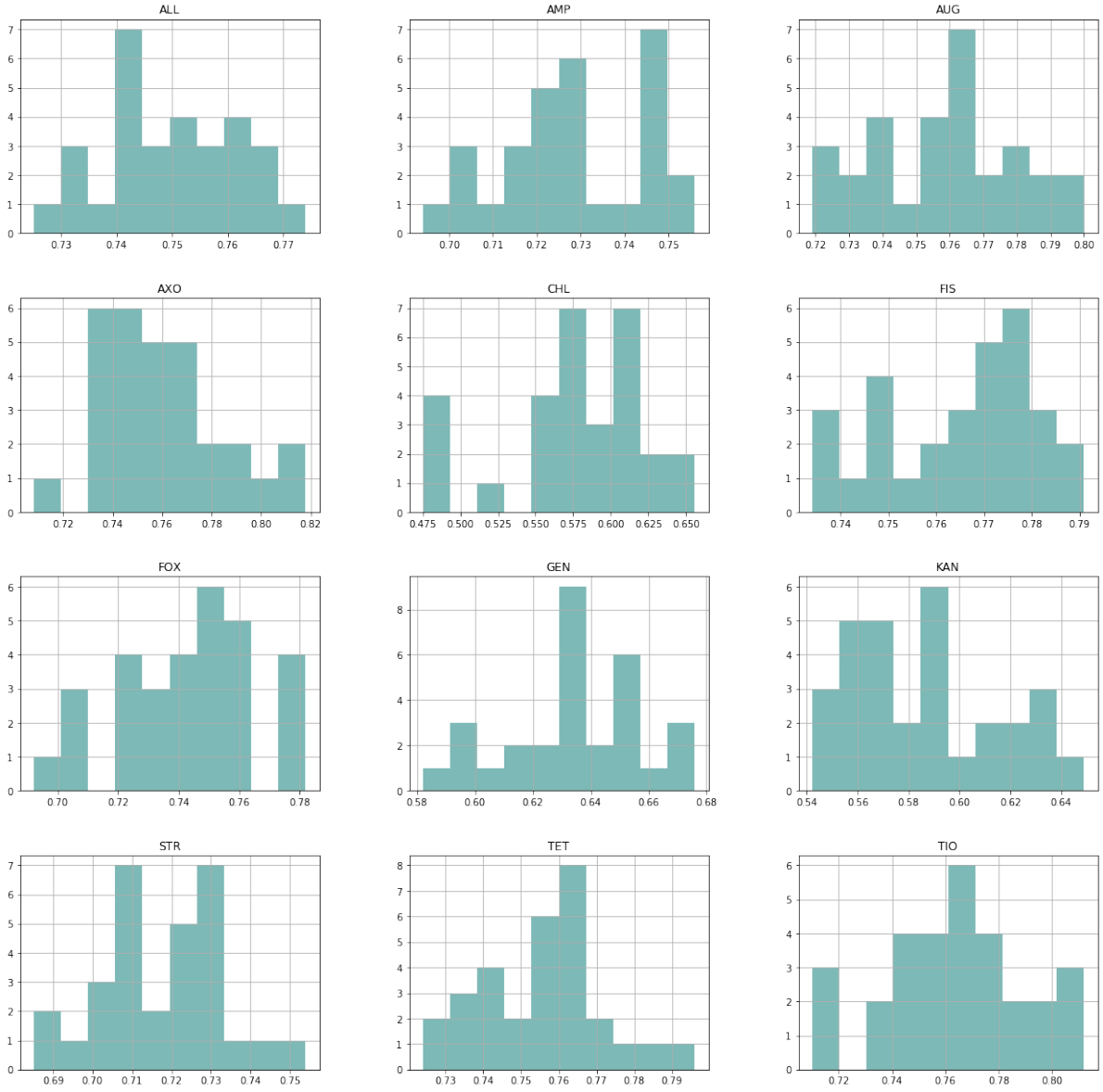

Figure 4.3: Histograms for F1 scores from 30 experiments using sets with 20 genes and a randomized group of genes as sources.

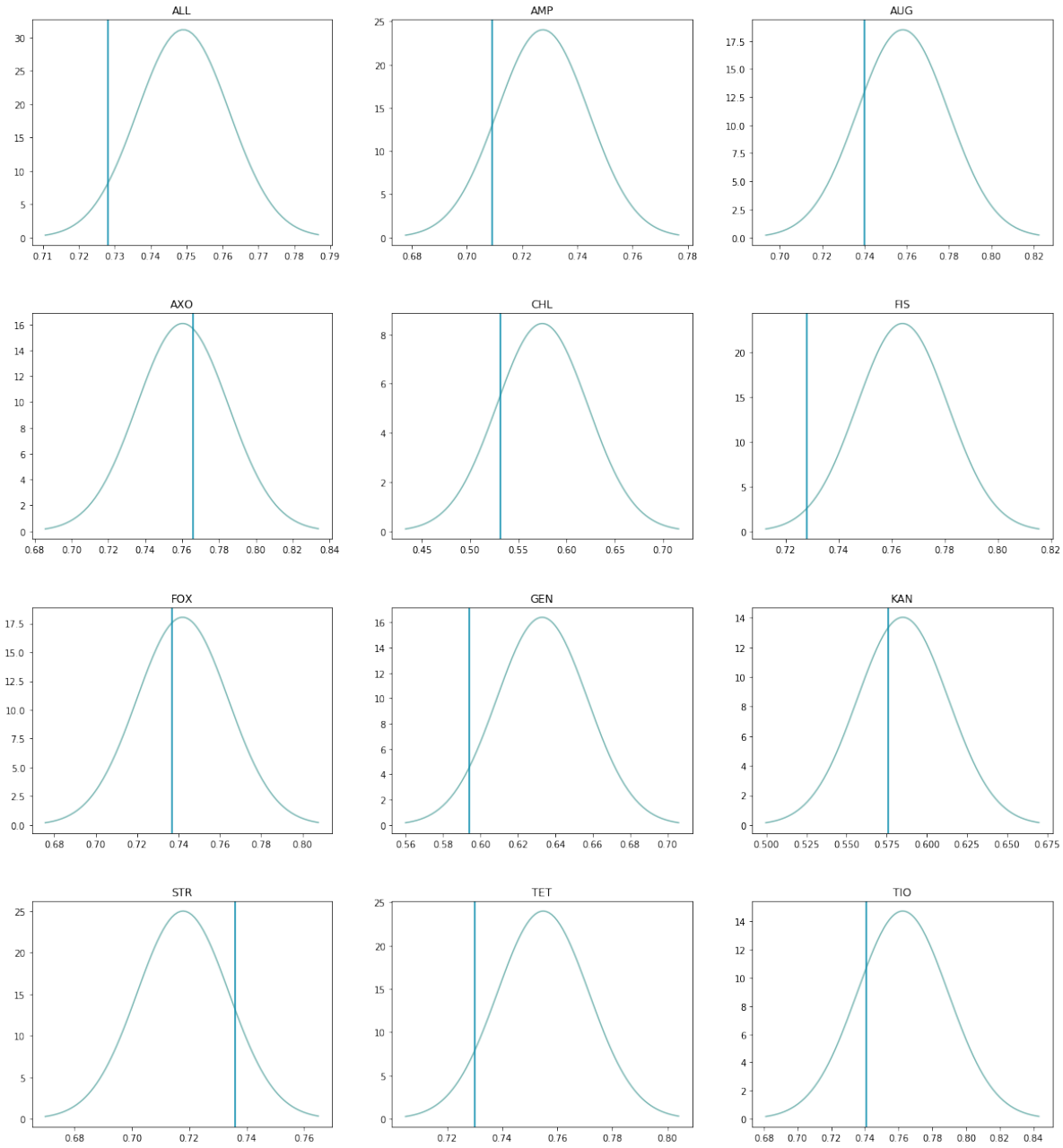

Figure 4.4: F1 scores (vertical blue line) obtained selecting a set of 20 genes with the best scores.

4.2.2 Prediction using 1-step Random Walk in a network with direct physical protein interactions

Figure 4.5 shows histograms obtained from 30 randomized experiments using 20 conserved genes. From these values, normal distributions were approximated and plotted on figure 4.6. The vertical blue line on figure 4.6 is the F1 score obtained using AMR genes as sources to calculate scores from 1-step kernel method.

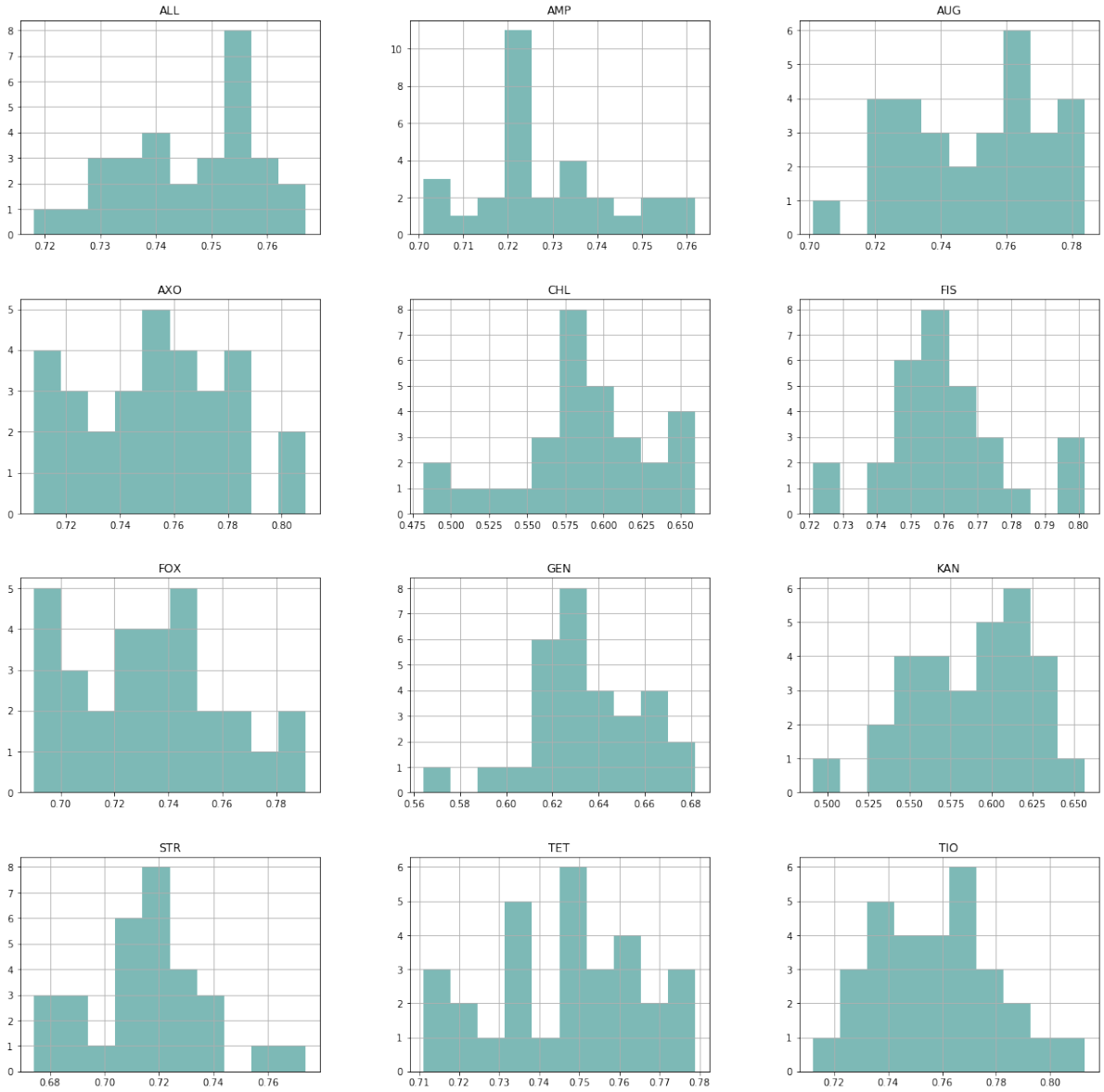

Figure 4.5: Histograms for F1 scores from 30 experiments using sets with 20 genes and a randomized group of of genes as sources.

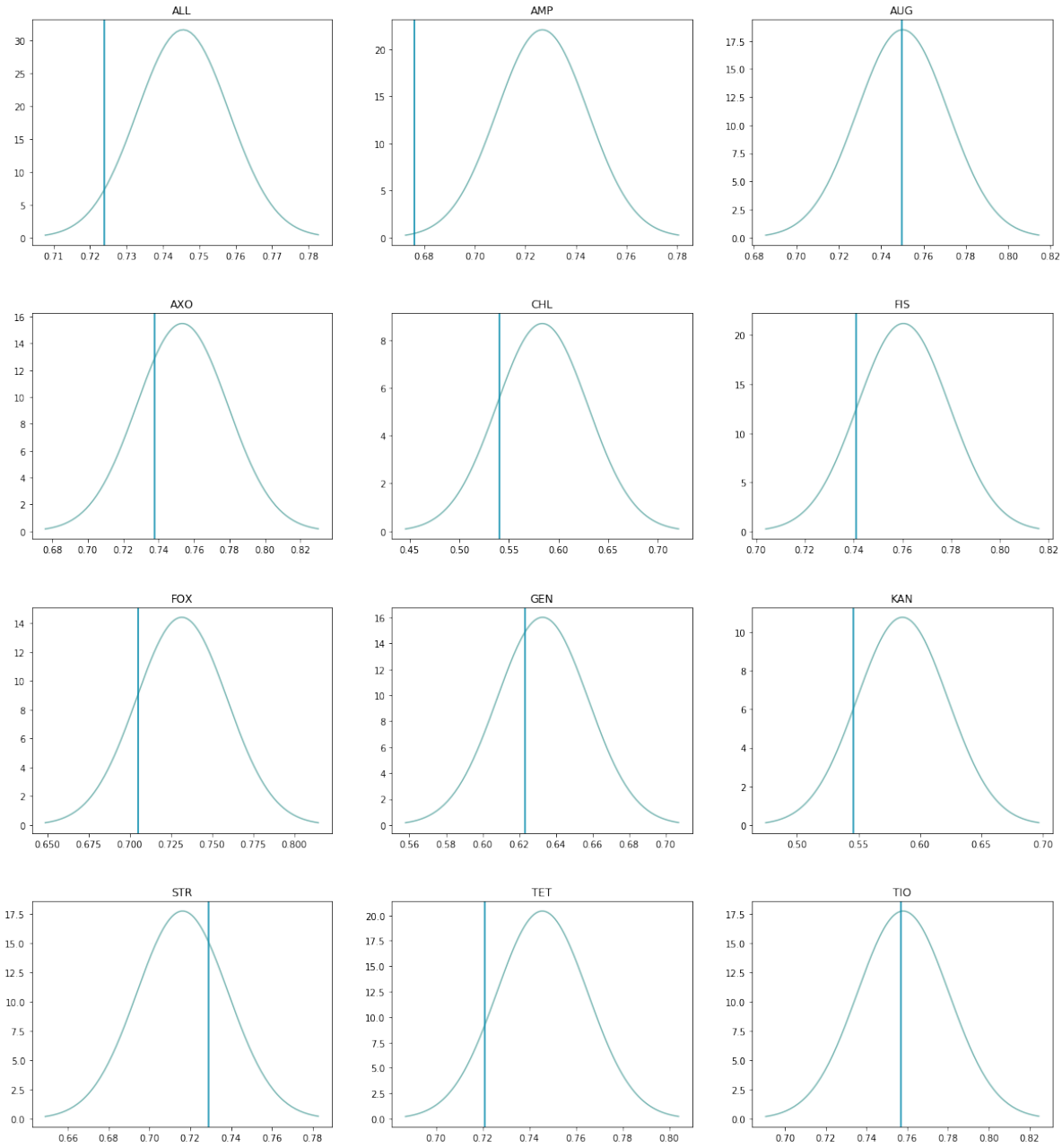

Figure 4.6: F1 scores (vertical blue line) obtained selecting a set of 20 genes with the best scores.

4.2.3 Prediction using 1-step Random Walk in a network with co-occurrence of proteins in physical complexes

Figure 4.7 shows histograms obtained from 30 randomized experiments using 20 conserved genes. From these values, normal distributions were approximated and plotted

on figure 4.8. The vertical blue line on figure 4.8 is the F1 score obtained using AMR genes as sources to calculate scores from 1-step kernel method.

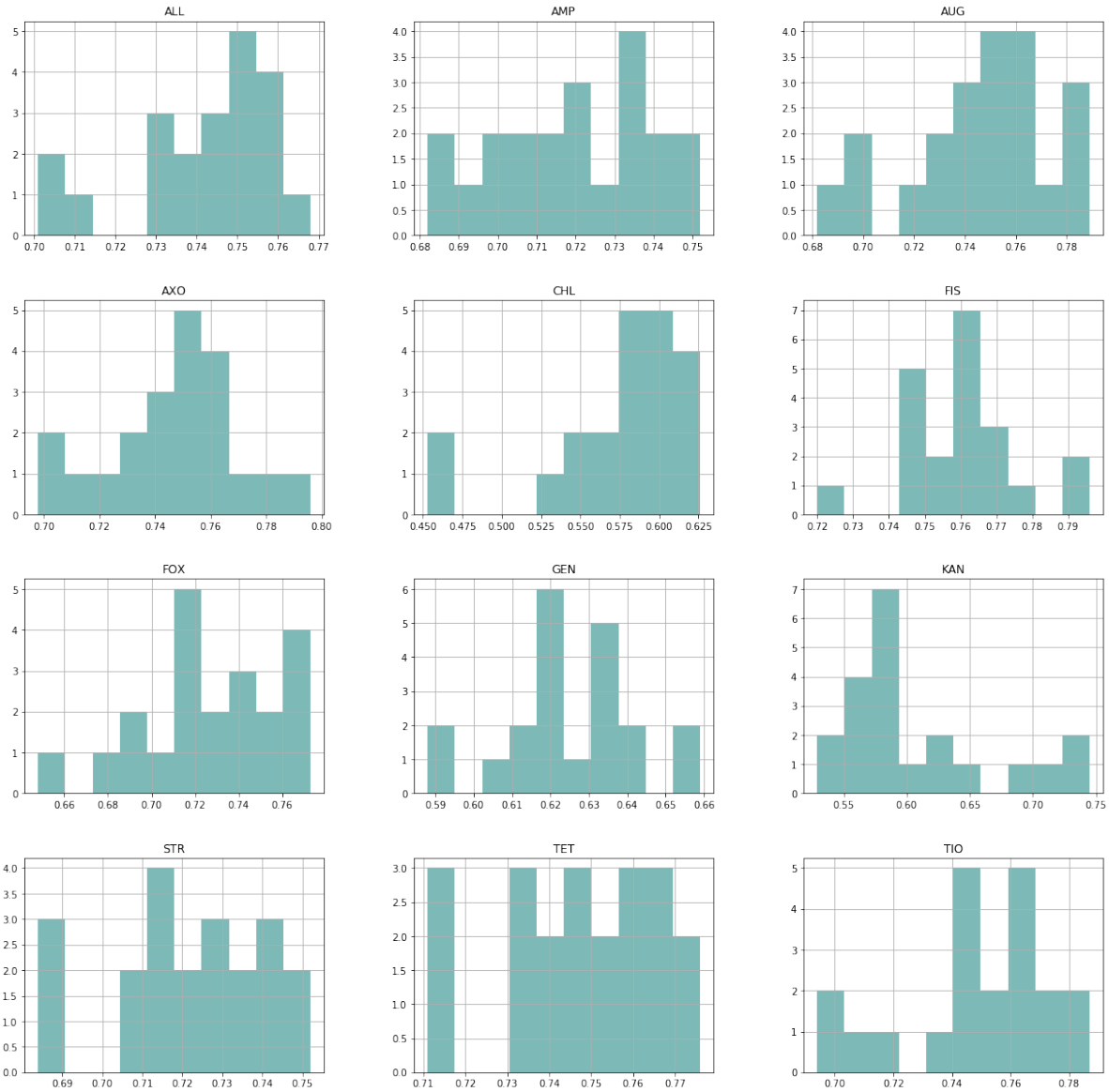

Figure 4.7: Histograms for F1 scores from 30 experiments using sets with 20 genes and a randomized group of of genes as sources.

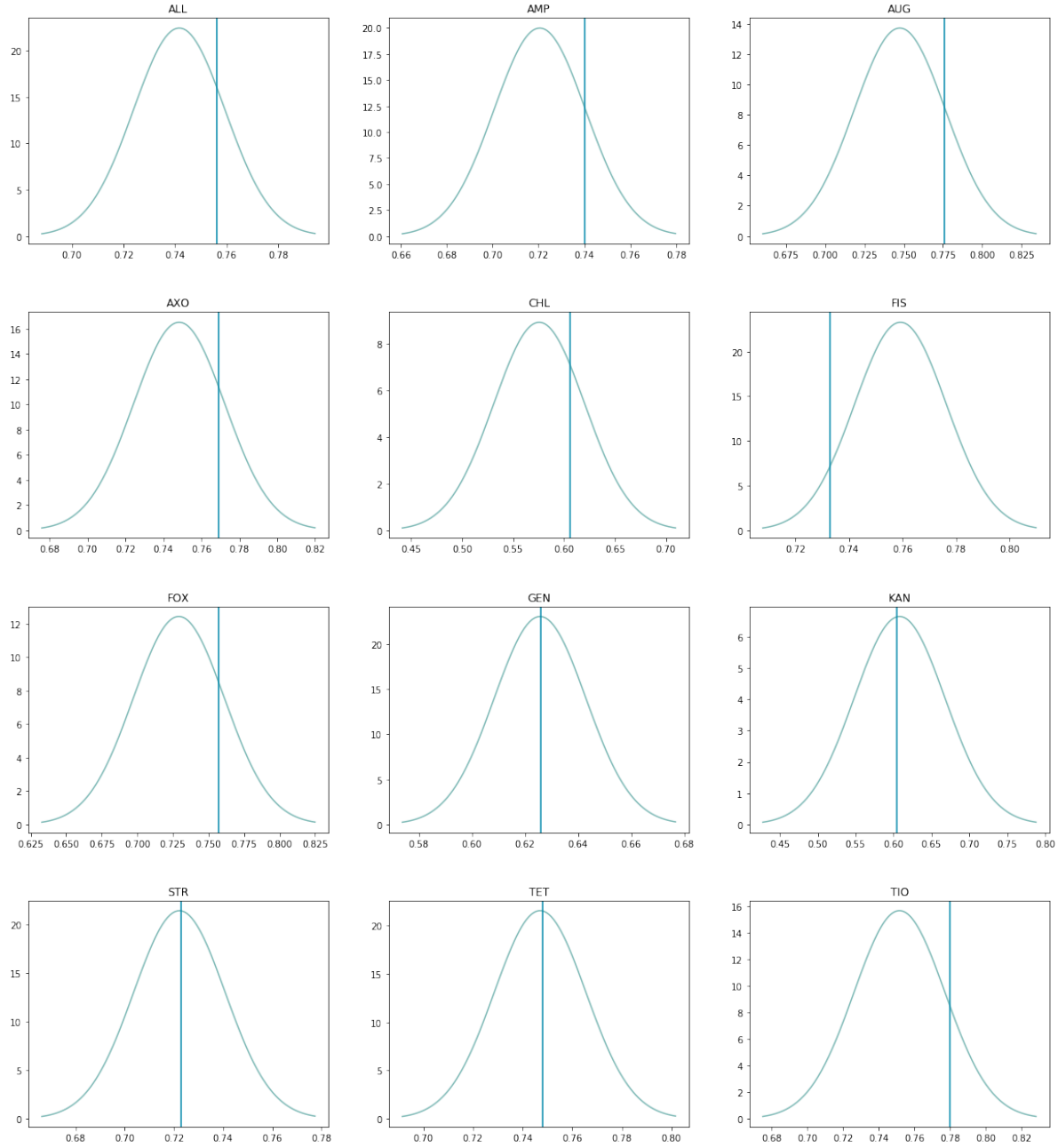

Figure 4.8: F1 scores (vertical blue line) obtained selecting a set of 20 genes with the best scores.

5

Discussion

The results of the experiments conducted in this study did not support the original hypothesis that conserved genes related to AMR genes in terms of physical interactions or protein complexes would achieve higher score in predicting susceptible/resistant phenotype compared to a random choice of conserved genes. In this chapter, we will discuss potential explanations for the unexpected results.

- low scores for an antibiotic probably are associated to the low number of samples for that specific antibiotic: chl has few resistant strains; kan has few strains at all

- p-values: none smaller than 0.1

One possible limitation for the lack of significant improvement in predictive score could be related to the limitations of the dataset used in this study. With relation to AMR genes, it is possible that we are using, within the conserved gene sets, genes that are responsible for antibiotic resistance mechanisms, but have not yet been associated with them. We are assuming that AMR genes have better features to predict the AMR phenotype, therefore, this could increase the mean of F1 scores distribution.

- amr per antibiotic: we tried to perform a prediction using a single amr gene to see if a specific amr gene could perform better to an specific antibiotic, but they predict in a very similar way to all antibiotics

- sample size

- ppi quality

- number of conserved genes: we are using a set of conserved genes given by [17], which is not a complete set of conserved genes. This is made because the methodology to get the sequences from PATRIC database is not clear in this paper.

- Another potential explanation could be related to the proposed predictive model. What could we do? simplify the model? +++

- The hypothesis itself: there is a possibility to have compensatory changes along the entire genome without relation to AMR genes

6

Conclusion

A

Data and Experiment

1) Describe how the previous experiment was reproduced

2) Github:

codes to construct gene sets

protein sequences from ncbi - file name amr genes - specialty genes ppis - file names of both - cite https://paccanarolab.org/s2f to get transferred ppi

BIBLIOGRAPHY

[1] Nobel Prize Outreach. Sir Alexander Fleming - Biographical. https://www.nobelprize.org/prizes/medicine/1945/fleming/biographical/. Accessed: 2022-05-17.

[2] Nobel Prize Outreach. Sir Alexander Fleming - Nobel lecture: penicillin. http://www.nobelprize.org/nobel_prizes/medicine/laureates/1945/fleming-lecture.htm. Accessed: 2023-01-02.

[3] Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*, 399, 2022.

[4] A. MacGowan and E. Macnaughton. Antibiotic resistance. *Medicine*, 45(10):622–628, 2017.

[5] J.A. Ayukekbong, M. Ntemgwa, and A.N Atabe. The threat of antimicrobial resistance in developing countries: causes and control strategies. *Antimicrobial Resistance & Infection Control*, 6, 2017.

[6] C.A. Michael, D. Dominey-Howes, and M. Labbate. The antimicrobial resistance crisis: Causes, consequences, and management. *Frontiers in Public Health*, 2, 2014.

[7] L. L Silver. Challenges of antibacterial discovery. *Clin. Microbiol. Rev.*, 24, 2011.

[8] A. Zorzet. Overcoming scientific and structural bottlenecks in antibacterial discovery and development. *Ups. J. Med. Sci.*, 119, 2014.

[9] Jessica D. Forbes, Natalie C. Knox, Jennifer Ronholm, Franco Pagotto, and Aleisha Reimer. Metagenomics: The next culture-independent game changer. *Frontiers in Microbiology*, 8, 2017.

[10] P. F. McDermott, G. H. Tyson, C. Kabera, Y. Chen, C. Li, J. P. Folster, S. L. Ayers, C. Lam, H. P. Tate, and S Zhao. Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal salmonella. *Antimicrobial Agents and Chemotherapy*, 60, 2016.

[11] M.J. Ellington, O. Ekelund, F.M. Aarestrup, R. Canton, M. Doumith, C. Giske, H. Grundman, H. Hasman, M.T.G. Holden, K.L. Hopkins, J. Iredell, G. Kahlmeter, C.U. Köser, A. MacGowan, D. Mevius, M. Mulvey, T. Naas, T. Peto, J.-M. Rolain, Ø. Samuelsen, and N. Woodford. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the eucast subcommittee. *Clinical Microbiology and Infection*, 23(1):2–22, 2017.

[12] A. C. Schürch and W Schaik. Challenges and opportunities for whole-genome sequencing–based surveillance of antibiotic resistance. *Annals of the New York Academy of Sciences*, 1388, 2017.

[13] N. C. Gordon, J. R. Price, K. Cole, R. Everitt, M. Morgan, J. Finney, A. M. Kearns, B. Pichon, B. Young, D. J. Wilson, M. J. Llewelyn, J. Paul, T. E. A. Peto, D. W. Crook, A. S. Walker, and T. Golubchik. Prediction of staphylococcus aureus antimicrobial resistance by whole-genome sequencing. *Journal of Clinical Microbiology*, 52(4):1182–1191, 2014.

[14] Michelle Su, Sarah W. Satola, and Timothy D. Read. Genome-based prediction of bacterial antibiotic resistance. *Journal of Clinical Microbiology*, 57(3):e01405–18, 2019.

[15] D. W. Eyre, D. De Silva, K. Cole, J. Peters, M. J. Cole, Y. H. Grad, W. Demczuk, I. Martin, M. R. Mulvey, D. W. Crook, A. S. Walker, T. E. A. Peto, and J. Paul. WGS to predict antibiotic MICs for Neisseria gonorrhoeae. *Journal of Antimicrobial Chemotherapy*, 72(7):1937–1947, 2017.

[16] G. H. Tyson, P. F. McDermott, C. Li, Y. Chen, D. A. Tadesse, S. Mukherjee, S. Bodeis-Jones, C. Kabera, S. A. Gaines, G. H. Loneragan, T. S. Edrington, M. Torrence, D. M. Harhay, and S. Zhao. WGS accurately predicts antimicrobial resistance in Escherichia coli. *Journal of Antimicrobial Chemotherapy*, 70(10):2763–2769, 2015.

[17] M. Nguyen, R. Olson, M. Shukla, M. VanOeffelen, and J.J. Davis. Predicting antimicrobial resistance using conserved genes. *PLoS Computational Biology*, 16(10):1–24, 10 2020.

[18] M. Nguyen, T. Brettin, and S.W. et. al. Long. Developing an in silico minimum inhibitory concentration panel test for klebsiella pneumoniae. *Scientific Reports*, 8(421):1–11, 1 2018.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani. An introduction to Statistical Learning with Applications in R. *Springer Texts in Statistics*, 103, 2013.

[20] S.B. Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39, 2013.

[21] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[22] C. X. Chan and M. A. Ragan. Next-generation phylogenomics. *Biology direct*, 8, 2013.

[23] J.J. et. al Davis. The patric bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Research*, 48(118):D606–D612, 01 2020.

[24] J.J. Davis, S. Gerdes, G.J. Olsen, R. Olson, G.D. Pusch, M. Shukla, V. Vonstein, A.R. Wattamm, and H. Yoo. Pattyfams: Protein families for the microbial genomes in the patric database. *Frontiers in Microbiology*, 7(118):1–12, 02 2016.

[25] A.L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12, 2011.

[26] I. W. Taylor and J. L. Wrana. Protein interaction networks in medicine and disease. *Proteomics*, 12, 2012.

[27] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11, 2010.

[28] A.J. Smola and R.I. Kondor. Kernels and regularization on graphs. *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings. Springer Berlin Heidelberg*, 2777, 2003.

[29] R.I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. *19th International Conference on Machine Learning. Proceedings*, 2002.

[30] S.M. Jajere. A review of salmonella enterica with particular focus on the pathogenicity and virulence factors, host specificity and antimicrobial resistance including multidrug resistance. *Vet World*, 12, 2019.

[31] M. Torres, H. Yang, A. E. Romero, and A. Paccanaro. Protein function prediction for newly sequenced organisms. *Nature Machine Intelligence*, 3, 2021.

[32] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, and C. Mering. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user uploaded gene/measurement sets. *Nucleic Acids Research*, 49, 2020.