A Graph-Theoretic Approach to Predict Antibiotic 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

FUNDAÇÃO GETULIO VARGAS

ESCOLA DE MATEMÁTICA APLICADA

BRUNA FERNANDA FISTAROL

A GRAPH-THEORETIC APPROACH TO PREDICT ANTIBIOTIC RESISTANCE

Rio de Janeiro

2023

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Fistarol, Bruna Fernanda
A graph-theoretic approach to predict antibiotic resistance / Bruna Fernanda Fistarol. – 2023.

44 f.

Dissertação (mestrado) – Fundação Getulio Vargas, Escola de Matemática Aplicada.
Orientador: Alberto Paccanaro.
Inclui bibliografia.

1. Aprendizado do computador. 2. Genoma. 3. Antibióticos. 4. Arvores de decisão. 5. Estatística - Métodos gráficos I. Paccanaro, Alberto. II Fundação Getulio Vargas. Escola de Matemática Aplicada. III. Título.

CDD – 519.2

Elaborada por Marcelle Costal de Castro dos Santos - CRB7-016/20

FUNDAÇÃO GETULIO VARGAS
MESTRADO EM MODELAGEM MATEMÁTICA
ESCOLA DE MATEMÁTICA APLICADA

BRUNA FERNANDA FISTAROL

"A GRAPH-THEORETIC APPROACH TO PREDICT ANTIBIOTIC RESISTANCE''.

DISSERTAÇÃO APRESENTADA AO CURSO DE MESTRADO EM MODELAGEM MATEMÁTICA PARA OBTENÇÃO DO GRAU DE MESTRA EM MODELAGEM MATEMÁTICA.

DATA DA DEFESA: 16/06/2023

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

PRESIDENTE DA COMISSÃO EXAMINADORA: PROFº ALBERTO PACCANARO

PROFº ALBERTO PACCANARO
ORIENTADOR

PROFº MOACYR ALVIM HORTA BARBOSA DA SILVA
MEMBRO INTERNO

PROFº JOÃO CARLOS SETUBAL
MEMBRO EXTERNO

RIO DE JANEIRO, 16 DE JUNHO DE 2023.

PROFº CÉSAR LEOPOLDO CAMACHO MANCO
DIRETOR

PROFº ANTONIO DE ARAUJO FREITAS JUNIOR
PRÓ-REITOR DE ENSINO, PESQUISA E PÓS-GRADUAÇÃO

ABSTRACT

For a given bacterial strain, antimicrobial resistance (AMR) genes are those that have been identified to confer antibiotic resistance. Usually, the bacterial whole genome sequence or AMR genes are used to predict antibiotic resistance. Interestingly, it has recently been shown that machine learning methods can predict antibiotic resistance also using sets of core genes chosen at random.

This study investigates a possible reason for this phenomenon. We hypothesized that core genes closely related to AMR genes in terms of physical interactions could contain information about antibiotic resistance. To test our hypothesis, we quantified the relation between core genes and AMR genes in terms of distances on the interactome that were measured using kernel methods on protein-protein interaction networks. This allowed us to select sets of core genes that were evaluated in terms of their ability to predict antibiotic resistance using a machine learning model based on decision trees.

Our experiments could not conclusively establish the existence of a relation between the distance of core genes from AMR genes and antibiotic resistance. While our results can possibly be justified by the poor quality of the currently available data, our work also suggests the existence of compensatory alterations in bacterial genomes that could be used in the future for antibiotic resistance prediction.

Key-words: Prediction of Antimicrobial Resistance. Machine Learning. Bacterial Genomes.

ACKNOWLEDGEMENTS

This is the part where I talk with my heart. My heart speaks in Portuguese :)

Quando decidi cursar o mestrado, ao final da graduação, meu objetivo era lapidar o conhecimento adquirido. Aprendi inúmeras ferramentas, mas sentia que não sabia muito bem o que fazer com elas. Embora já tivesse tido contato com pesquisa antes, nunca havia dado meus próprios passos. Hoje, ao final do mestrado, sinto que cumpri meu objetivo: caminhei um pequeno trecho, ainda que tenha sido em direção a resultados diferentes do esperado - o que faz parte do processo na maioria das vezes, na verdade. Fico feliz em ver o quanto cresci. Por isso, começo agradecendo ao Alberto, meu orientador, por me guiar nessa jornada. O modo como ensina, mentora e dá atenção aos detalhes tornou a experiência muito mais motivante e enriquecedora. Não menos importante, reconhece e compreende as dificuldades. Ser humano ímpar. *Grazie mille*, Alberto!

Agradeço também pela oportunidade de fazer parte do Paccanaro Lab, onde estive com pessoas fantásticas com as quais pude aprender - desde machine learning até palavrões em espanhol, rsrs. Essa experiência também me ajudou a lidar com minhas dificuldades de comunicação. Estando no lab eu tive um incentivo maior para falar com as pessoas, seja tirando dúvidas, discutindo algum problema ou conversando. Meu muito obrigada, Suzana, Mateo, Ruben, Santiago, Phil, Aldo e Marimar. Agradeço também à Suzana e ao Santiago por terem participado das minhas reuniões semanais durante a discussão do meu projeto.

Ao Breno, meu noivo, agradeço por me incentivar e estar ao meu lado todos os dias. Minha trajetória no mestrado foi repleta de dúvidas e percalços, e o fato de fazer com que eu me mantivesse acreditando no meu potencial foi essencial. Eu te amo!

Por falar em incentivo, agradeço à minha mãe e meu pai, que apoiam minhas decisões mesmo que não compreendam muito bem como funciona a carreira acadêmica, algo bastante distante da realidade de onde vim. Obrigada por terem orgulho da pessoa que estou me tornando.

Alguns anos depois de me entusiasmar com a matemática, também me tornei uma admiradora da biologia. Agradeço à Nislâine pela forma como me introduziu a esse mundo fascinante. Tenho muita sorte de ser amiga de uma pessoa que admiro muito - a biologia não foi a única que me cativou.

Agradeço ao Marcelo Alex, por ter me recebido em seu grupo de pesquisa em 2017 e me dado a oportunidade de trabalhar com a “bancada molhada” durante dois anos. Essa

experiência foi essencial na formação do meu perfil interdisciplinar, mas mais que isso, ganhei uma família em um momento que precisei muito. Serei eternamente grata.

Por fim, nada disso seria possível se eu não tivesse tido a oportunidade extraordinária de compor a primeira turma de alunos convidados a estudar na FGV através do CDMC, em 2017. Agradeço imensamente ao Professor Camacho e à Cássia por todas as oportunidades, suporte e apoio desde o início da graduação.

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 [2]. 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 [3].

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 [4].

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 [5].

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 [6, 7]. 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 [8, 9]. Once there is no straightforward solution for the issue of antimicrobial resistance, it is necessary to understand it by studying underlying mechanisms in order to develop tools like diagnostic methods, treatment protocols, or alternative therapies, that can support healthcare professionals in making more effective choices in the management of infections.

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 using a sample from a patient, 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 core. Core genes are more trivial to identify through phylogenetic

analysis [10].

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 [11–17]. 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 core DNA sequences can provide a prediction regarding the susceptible or resistant phenotype of bacteria to a given antibiotic [18]. However, to perform the prediction the core genes are randomly selected and no information related to the genes is used other than its preserved sequence.

The present study proposes to investigate whether certain groups of core genes can have a better performance of predicting the resistance phenotype rather than a random selection. Assuming that AMR genes characterize resistant bacteria, we believe that the predictive power of core genes is related with them.

We will use information about the interaction of core genes with AMR genes to develop methods to select core genes. The central 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 get mutations related to antimicrobial resistance.

In Chapter 2, we will see how a previous study used core bacterial genome sequences to make predictions of the resistance phenotype. In Chapter 3 we present two ideas relying on Network Medicine in order to select core genes to repeat the experiments described in Chapter 2. The first idea is based on the shortest path between core genes and AMR genes in the protein interaction network. The second idea diffuses antimicrobial resistance genes in the protein interaction network in order to map relevant core genes with respect to diffusion. 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 core genes

The current DNA sequencing technology allows computational methods to analyze bacterial genomes. Additionally, strategies for identifying antimicrobial resistance profiles without bacterial culture can enable faster diagnoses. However, sequencing without culturing bacteria may not provide complete microbial sequences. The nature of resistance genes, often associated with plasmids and not core sequences, can also make identification difficult. Despite many computational studies developing algorithms to recognize patterns in bacterial DNA sequences, most rely on complete microbial sequences. The challenges related to obtaining complete bacterial genome sequences without additional experiments motivates research into the predictive potential of partial bacterial DNA sequences.

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 [19]. 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 core genes among members of the same species and unrelated to the antibacterial resistance phenotype [18]. The model was constructed in order to classify the phenotype of a specific bacterium as susceptible or resistant given its core genes 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 that combines a series of simple tests in a logical manner. Each test compares an attribute to a threshold value from a set of possible values, which helps determine the appropriate path through the tree. [1].

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, a training set is utilized to determine how the prediction space is divided. This is achieved by selecting the best split, which corresponds to the threshold value that yield the lowest classification error. For instance, consider a scenario where datapoints have features X_1 and X_2 and the threshold values t_1, t_2, t_3, and t_4 are defined. These threshold values divide the prediction space into five regions, as shown in figure 2.1 [1] and they are chosen in order to have the lowest number of misclassified datapoints from the training set.

To address some limitations of decision trees, they are often combined with techniques that can enhance their performance, such as ensemble methods like gradient boosting.

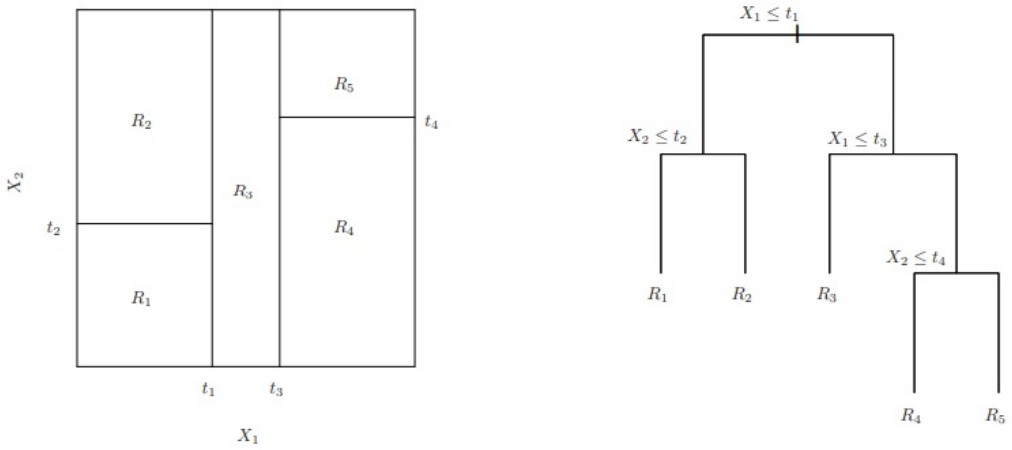

Figure 2.1: The output of recursive binary splitting on a two-dimensional example and a tree corresponding to this partition, respectively [1].

2.1.2 Gradient Boosting

Decision trees can suffer from high variance, which means that minor changes in the data used to train decision trees can lead to very different model performance. Boosting is a general approach that can be applied to many statistical learning methods to reduce this variance. It is an ensemble technique, which means it combines multiple learning algorithms - called weak learners - to obtain a better predictive performance, referred to as a strong learner. In this context, 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, which are the differences between the observed values and the predicted values from the model. By focusing on the residuals, each tree attempts to minimize the errors made by the previous tree. Every time a tree is added to the ensemble, they are weighted in order to give higher importance 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. The loss function, which measures the difference between the actual and predicted values, varies depending on the problem being solved (e.g., mean squared error for regression problems, cross-entropy loss for classification problems). The algorithm then trains a new model

to predict the residuals. These residuals are added to the previous predictions, and this process is repeated until convergence or a predefined stopping criterion is met.

By iteratively combining weak learners and focusing on minimizing the errors from previous iterations, gradient boosting can effectively reduce the high variance of individual decision trees and improve the overall predictive performance of the ensemble model.

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].

2.2 PATRIC Database

Genome and laboratory-derived antimicrobial susceptibility test from Nguyen et al. [18] 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 funded by the National Institute of Allergy and Infectious Diseases (NIAID) and is freely accessible to the public [22].

On December, 2022, after the data collection present in this work, PATRIC database was combined with the Virus Pathogen Resource Database (ViPR) and the Influenza

Research Database (IRD), creating the unified database named Bacterial and Viral Bioinformatics Resource Center (BV-BRC) [23].

2.2.1 Protein families (PLFam)

The analyses of Nguyen et al. [18] 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]. In order to reference analogous proteins from different subspecies with the same label and create the dataset, the authors reannotated all subspecies genes so that they had the same set of protein family calls. Also, all families with a PATRIC annotation associated with AMR were removed from their study.

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.

2.2.2 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 core genes. Given some species of bacteria, which has several strains, the authors choose at random a group of core genes in terms of protein families. Then, for each strain, these core genes are selected to construct the model dataset. Figure 2.2 shows how a species of bacteria has several strains (all vertical lines) associated to a susceptible or resistant phenotype (S or R) with relation to an antibiotic. Each strain has a set of core genes (all horizontal lines), where selected core genes are represented by orange horizontal lines.

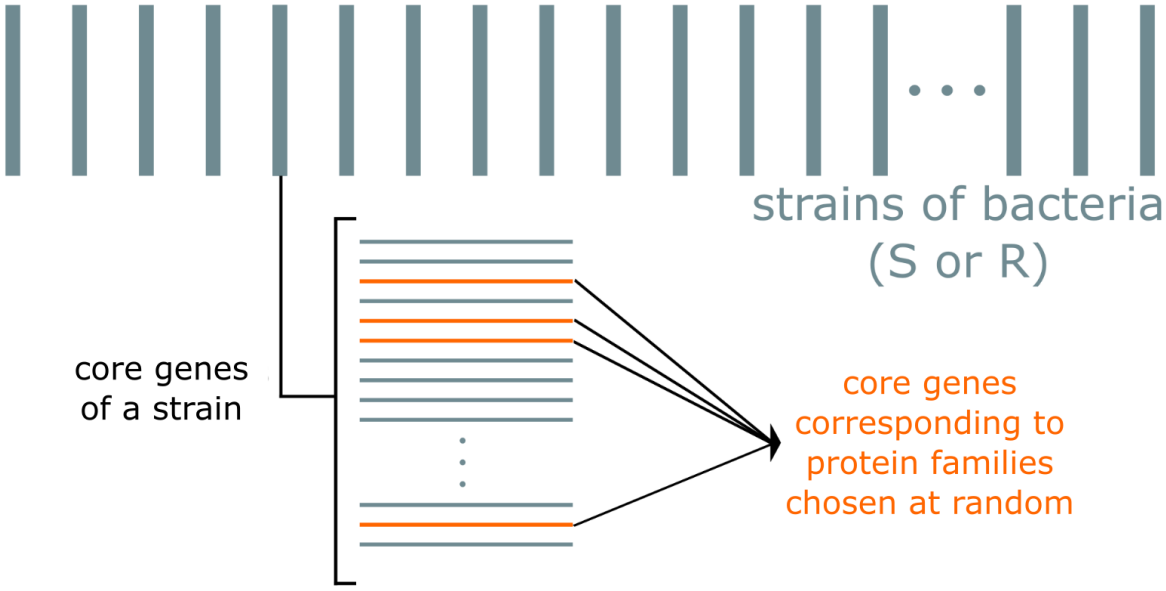

Figure 2.2: Framework of the data used by Nguyen et al. The vertical lines are different strains (or subspecies) of a bacteria associated to a susceptible or resistant phenotype (S or R) with relation to an antibiotic. All core genes are represented by horizontal lines, where the orange lines correspond to the set of protein families chosen at random.

Having the sequences associated to a strain, it is necessary to codificate this information in order to have it as a model input. This is made counting oligonucleotide k-mers.

2.3.1 Oligonucleotide k-mers

Oligonucleotide k-mers of DNA sequences are sub-sequences containing k nucleotides. 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 2.3: 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 and dataset size, as k-mer-based methods do not require pairwise sequence comparisons. This makes them particularly suitable for large-scale genomic analyses and comparisons [25].

2.3.2 Model Dataset

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 core 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% 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.

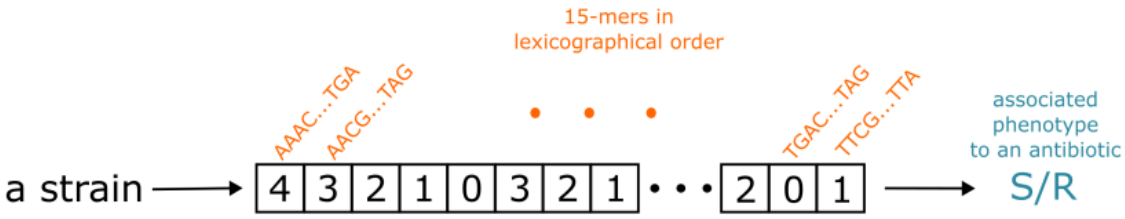

Figure 2.4: Structure of data used as features by the model, where a strain is represented by a vector containing the frequency counts of all occurring 15-mers in the group of selected core genes, organized by lexicographic order. This vector is associated to a resistant or susceptible phenotype (S/R).

However, there are no criteria to create the sets of core genes used to run the experiments - they are constructed at random from the complete set of core genes. This fact motivates us to explore the possibility to select these genes using some criterion based on the interaction between core genes and AMR genes and verify whether we can improve the prediction using specific sets of core genes.

3

SELECTING CORE GENES BASED ON INFORMATION FROM PROTEIN INTERACTION NETWORK

Given the fact that core 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.

Remembering that, from a clinical perspective, we are interested in providing information on the resistance phenotype as quickly as possible and using the available information - bacterial genome sequence - in the best way. Detection of the resistance phenotype can be accelerated when using laboratory strategies that do not require bacterial cell culture. However, this makes it difficult to obtain complete coverage of the bacterial genome in its sequencing. Since sequences of core genes are more likely to be obtained, the objective is to verify if we can improve the performance of the prediction of the bacterial resistance phenotype when we use specific sets of core genes, instead of selecting them randomly.

In this sense, two methods will be proposed in this chapter 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. It involves integrating multiple sources of biological and clinical data to build comprehensive models to identify novel therapeutics by analyzing complex networks of interactions between different biological components, like proteins, genes, metabolites, RNA molecules or even diseases and phenotypes. [26]. 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 [27].

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 [26].

3.2 The Project

Since the previous study was based on protein-encoding genes, a protein interaction network will be used to analyze distance between core genes and AMR genes as a graph, where proteins are nodes and an edge is a physical or functional interaction between two proteins.

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

The approach is based on the hypothesis that core 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 core genes represented in figure 3.1 could have a better predictive power than choose core genes at random, based on their proximity from AMR genes. We can use measures in the graph to find these genes and use them to predict the resistance phenotype.

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

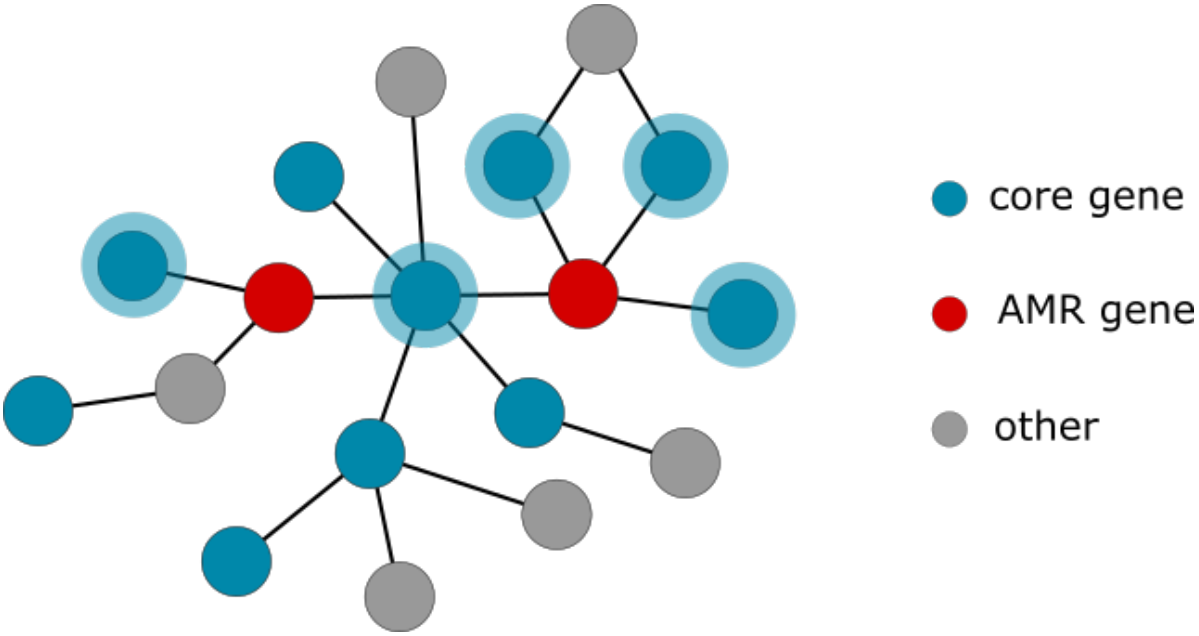

Figure 3.1: Illustrative example of a bacterial protein interaction network where core genes (blue) closer to AMR genes (red) are highlighted.

3.3 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 order to interpret these methods to understand what they are and how they can be used in our problem, some definitions are necessary.

The definitions below are based on [28], [29] and [30].

A kernel function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a measure of similarity between objects. In particular, they can be used to analyze and model the relationships between different components of biological networks - in this case, proteins. It implicitly constructs 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, $x' \in \mathcal{X}$. 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 Laplace Operators

We know that biological structures can be represented as a graph. Now, we need to have a way to write it in order to construct a kernel function satisfying its definitions. In this sense, the Laplacian of a graph (defined below) is used.

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

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

3.3.2 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,

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

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

We can think of the regularization operator as implicitly defining a kernel function on the nodes of the graph. Specifically, the Graph Laplacian regularization encourages the model to assign similar function values to nodes that are close in the graph, which can be interpreted as a kind of similarity measure.

In other words, nodes that are close to each other in the graph (i.e., they are connected by a path of small length) are considered to be similar, and hence they are encouraged to have similar function values. This aligns with our intuitive understanding of biological components: if two genes or proteins interact directly or are part of the same pathway, they are likely to be involved in similar biological processes, and hence they should be treated similarly by our model.

3.3.3 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

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

Different choices to $r(\lambda)$ define different kernel functions, which depends on the specific context. In our case, we would like to look at core genes closer to AMR genes, therefore, a diffusion kernel considering AMR genes as source nodes may be suitable.

3.3.4 *p*-Step Random Walk

The p-step random walk kernel can be interpreted as a measure of the connectivity or the interaction intensity between two nodes in the graph. It counts the number of paths of length p between two nodes, where a path is a sequence of edges that connects the nodes. This means that two nodes are considered similar if there are many paths of length p connecting them.

The p-step random walk kernel is obtained with the following choice of $r(\lambda)$ to the kernel function,

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

which provides the kernel

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

where a acts as a regularization term.

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.

The value of K_{ij} is proportional of the probability of arrive i j from i in a random walk after p steps. The parameter p allows you to control the complexity of the interactions you are considering. For small values of p, the kernel will only consider short paths between nodes, while for larger values of p, the kernel will consider longer paths.

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, a binary matrix of adjacency. All datasets and codes to construct gene sets can be found in https://github.com/brunafistarol/MasterProject.

3.4.1 Naive selection of core genes

A very simple way to separate core genes in different groups is based on the shortest path to any AMR gene. The goal is to verify whether core genes with a smaller shortest path to any AMR gene can provide a better prediction score than core genes far from AMR genes.

As an example, if a core 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 core gene will be part of a group with label 1.

3.4.2 Selecting core 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 core genes

```
Data: core_nodes, AMR_ nodes
Result: sets of core genes according to the shortest path to an AMR gene
for i in core_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,

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

on the kernel given by p-step random walk,

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

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 core 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.

core 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 core genes.

3.5 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 [31].

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 core genes has 11 genes annotated as AMR genes, which were found after the publication of Nguyen et al. [18]. 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*. We assume that we can do this given that we are working with core genes.

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* to train/test/validate the model.

3.5.2 Core genes

The group of core genes used to repeat the experiments is a group of 500 genes given by the previous study [18]. At the moment that this data was collected (December 1, 2018),

core 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 the 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 core 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 [32], since *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 to predict the PPI was obtained from National Center for Biotechnology Information (NCBI) database, searching for protein sequences of the specific subspecies of *Salmonella enterica*.

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 [33]. 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

In this chapter, we will present the results obtained from the resistance phenotype prediction using the sets of core genes constructed as it was described on the previous chapter.

We began running the experiments from Nguyen et al. [18] using sets of core genes according to the shortest path from a core gene to any AMR gene (as it is described on algorithm 1). To construct these groups, we used a set of 500 core genes from *Salmonella enterica* given by Nguyen et al. [18], and the PPI network with direct physical interactions between protein (described on section 3.5.4.1). The number of genes in each group according to the shortest path to any AMR gene is in the table 4.1

There are 9 core genes with distance zero from an AMR gene, which means these genes were annotated as AMR genes after the data collection from Nguyen et al. [18], because all core genes annotated as AMR genes were removed from the study. We will not consider these 9 genes to run experiments for core genes close to AMR genes. Instead, we will use these 9 genes as a separated batch to run the model in order to verify whether the best prediction score comes from this group of genes.

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

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

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, where 1 is the best score. In short, this metric indicates how good a set of genes can predict if a bacterial strain is resistant or not to some antibiotic.

An experiment is made running the prediction model described for Nguyen et al. [18] using a set of genes constructed here. The red columns indicates the F1 score obtained from 9 core genes annotated as AMR. The blue columns are the mean of several experiments using 9 genes: basically, there are 83 genes with the length of the shortest path to any AMR gene equals to 1 (as indicates the table 4.1), 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 in our study, we expected a decreasing score as the core genes move away from AMR genes. However, this pattern is not observed in my experiments.

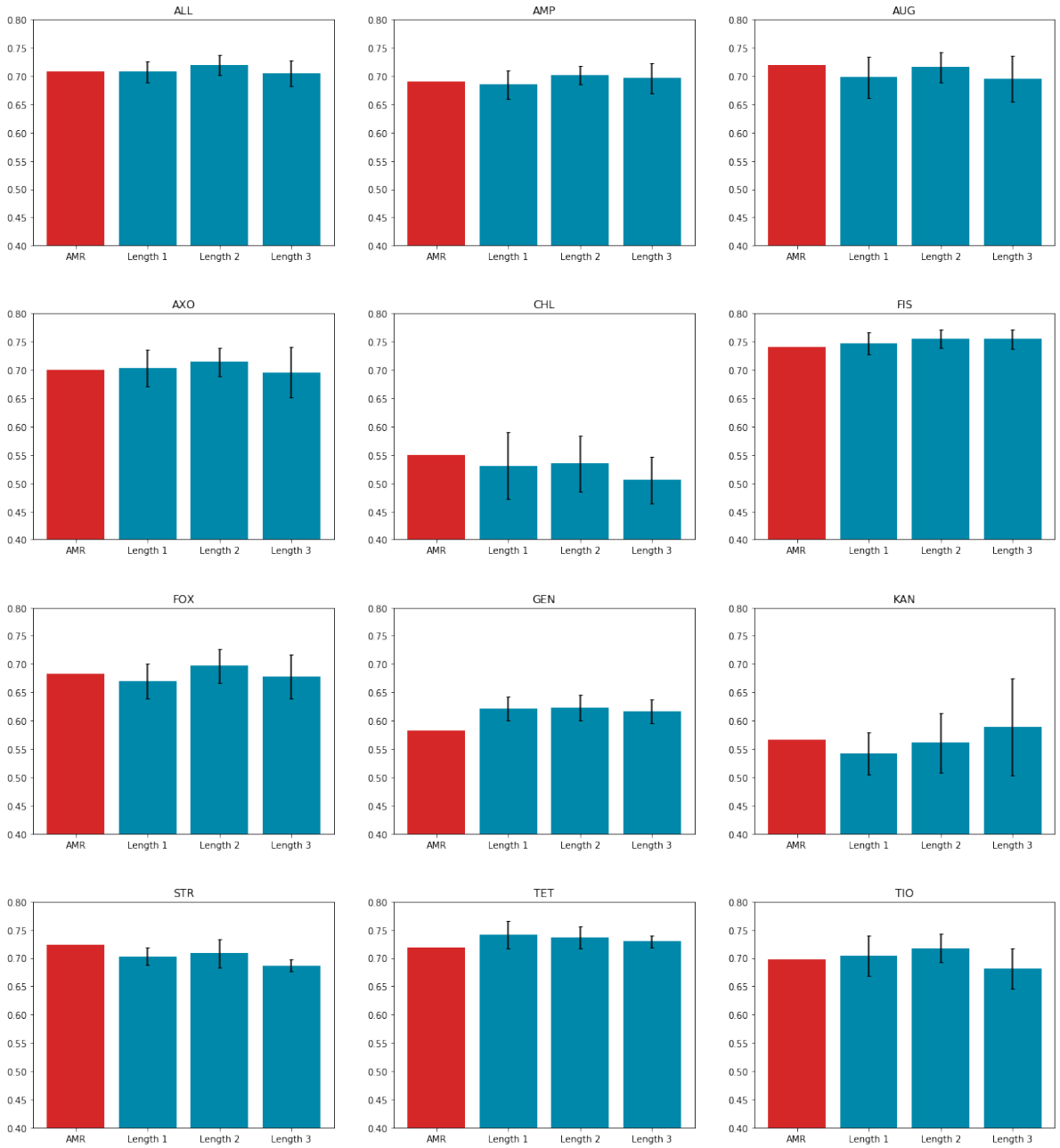

Figure 4.1: F1 scores obtained selecting genes based on distance from an AMR gene to a core gene.

We can look at the figure 4.2 to think about a possible reason for these results. The core 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

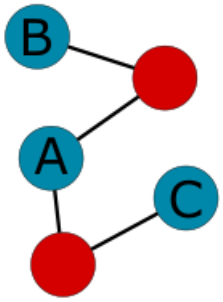

Figure 4.2: Interactions between protein showing a possible reason to not have a good result with a naive selection.

assumption is correct. This motivates the next approach, which makes a diffusion along the PPI using AMR genes as source nodes, trying to summarize the effect of multiple AMR genes.

4.2 Prediction based on scores from kernel method

Kernel methods provide a way to summarize the contribution of all AMR genes. The idea is to measure the additive effect of AMR genes through the diffusion of these genes in the interaction network between proteins. Using the method described in section 3.4.2 it is possible to get scores for each core gene in the PPI network in terms of a diffusion along the graph. The idea is to check if a group core genes with the highest score in terms of kernel measure. (i. e., more connected with AMR genes) can have a significant improve on its prediction performance when compared to a normal distribution of scored obtained running several experiments with core genes randomly selected.

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 density of the normal distributions is plotted with three standard deviation for each side around its mean.

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

The objective is summarize the effect of AMR genes along the physical interaction network making a diffusion from AMR genes to core genes up to two edges from AMR genes, since genes on the same pathway tend to be close on the interactome. The 20 core genes with the highest score were used to run the prediction and get the F1 score from this prediction, which is compared to a distribution of F1 scores from randomized experiments.

Figure 4.3 shows histograms obtained from 30 randomized experiments using 20 core 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.

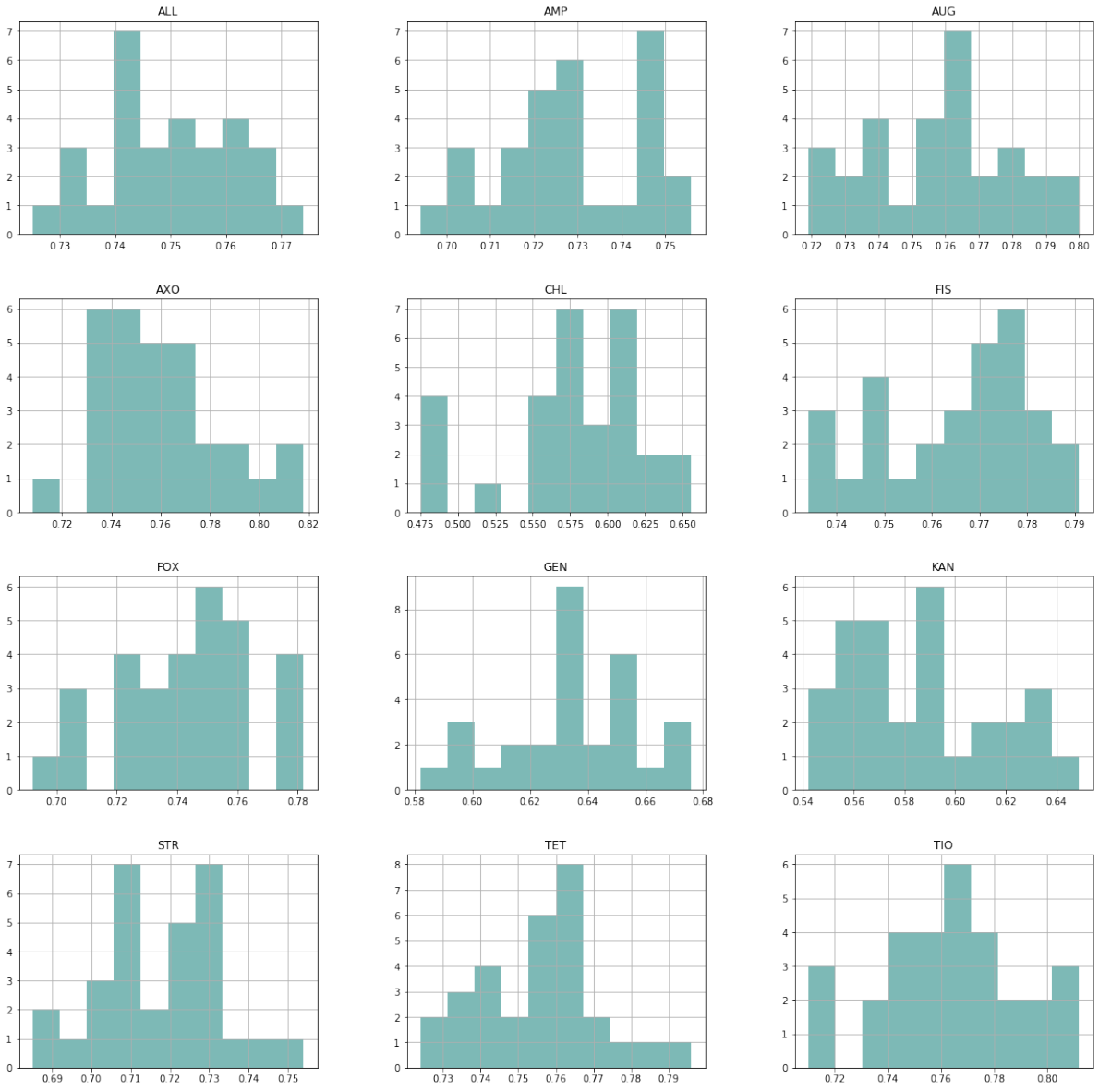

Figure 4.3: Histograms for F1 scores from 30 experiments using sets with 20 genes and a randomized group of genes as sources.

There was no statistically significant difference in the performance of the models for any of the antibiotics. Since the interaction network between bacterial proteins is highly connected, we could be capturing unrelated genes when using $p = 2$. Therefore, we also tried to reproduce the experiment using this same interaction network with $p = 1$.

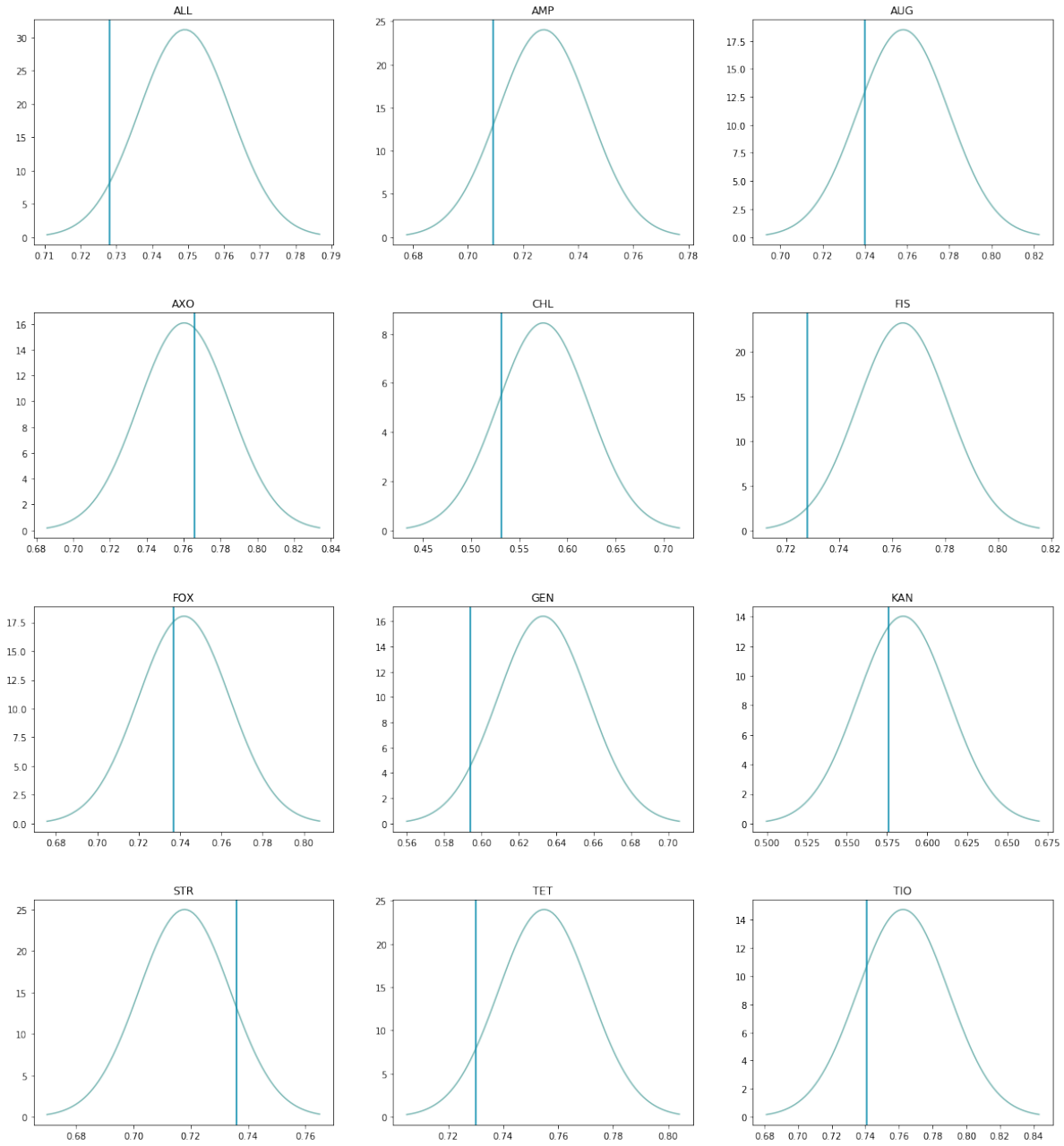

Figure 4.4: F1 scores (vertical blue line) obtained selecting a set of 20 genes with the best scores. The density of the normal distributions from randomized experiments is plotted with three standard deviation for each side around its mean.

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

Analogously to the previous experiment and using $p = 1$, figure 4.5 shows histograms obtained from 30 randomized experiments using 20 core 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.

Likewise, there was no statistically significant increasing in the performance of the models for any of the antibiotics. So, we tried to use an interaction network in which the edges express the probability of finding two proteins working in the same physical complex in the cell, with $p = 1$.

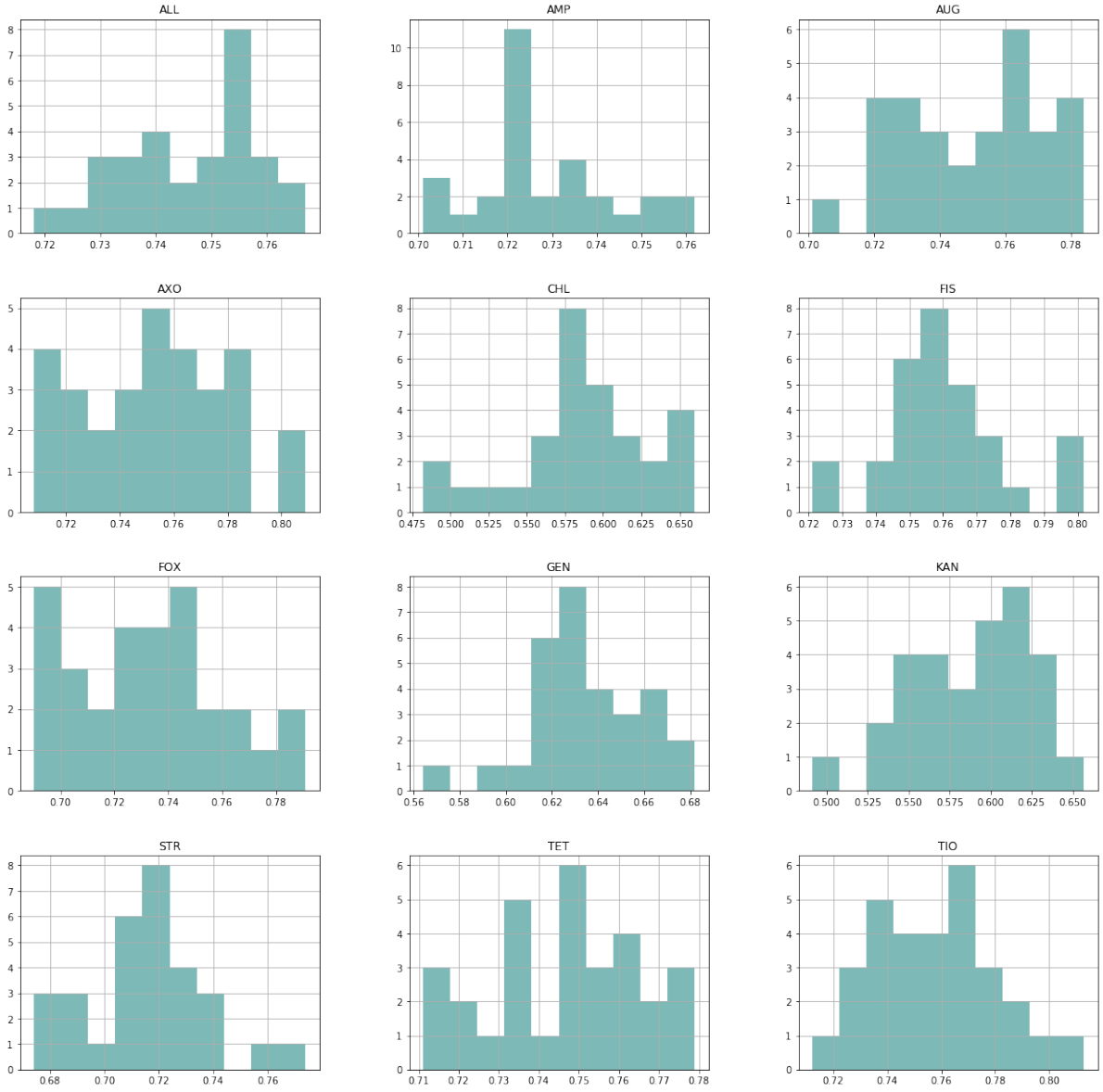

Figure 4.5: Histograms for F1 scores from 30 experiments using sets with 20 genes and a randomized group of of genes as sources.

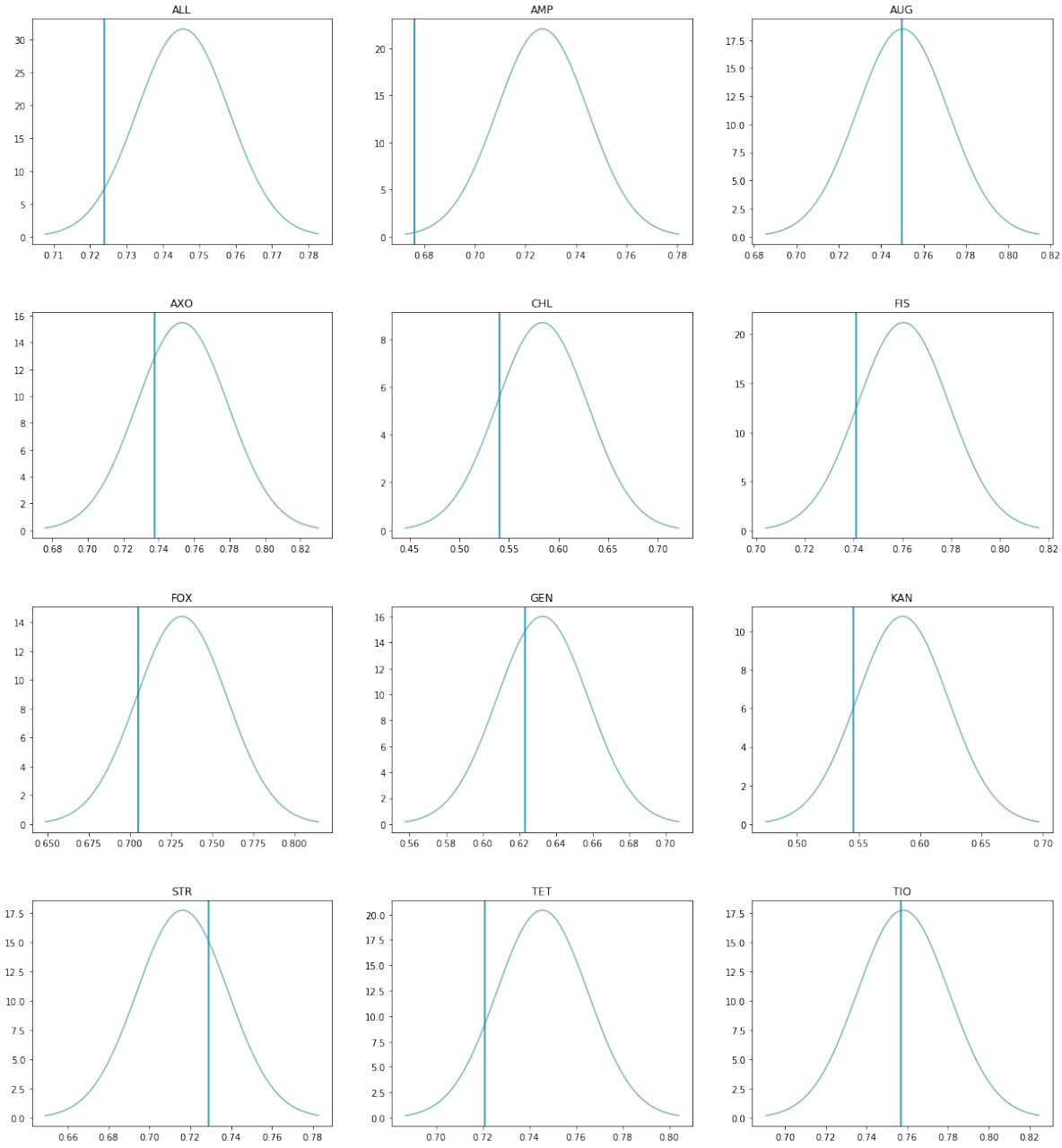

Figure 4.6: F1 scores (vertical blue line) obtained selecting a set of 20 genes with the best scores. The density of the normal distributions from randomized experiments is plotted with three standard deviation for each side around its mean.

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

The objective is summarize the effect of AMR genes along the co-complex network making a diffusion from AMR genes to core genes up to one edge from AMR genes. In other words, we are selecting core genes with the highest probability to be found in a physical complex with AMR genes. The 20 core genes with the highest score were used to run the prediction and get the F1 score from this prediction, which is compared to a distribution of F1 scores from randomized experiments.

Figure 4.7 shows histograms obtained from 30 randomized experiments using 20 core 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.

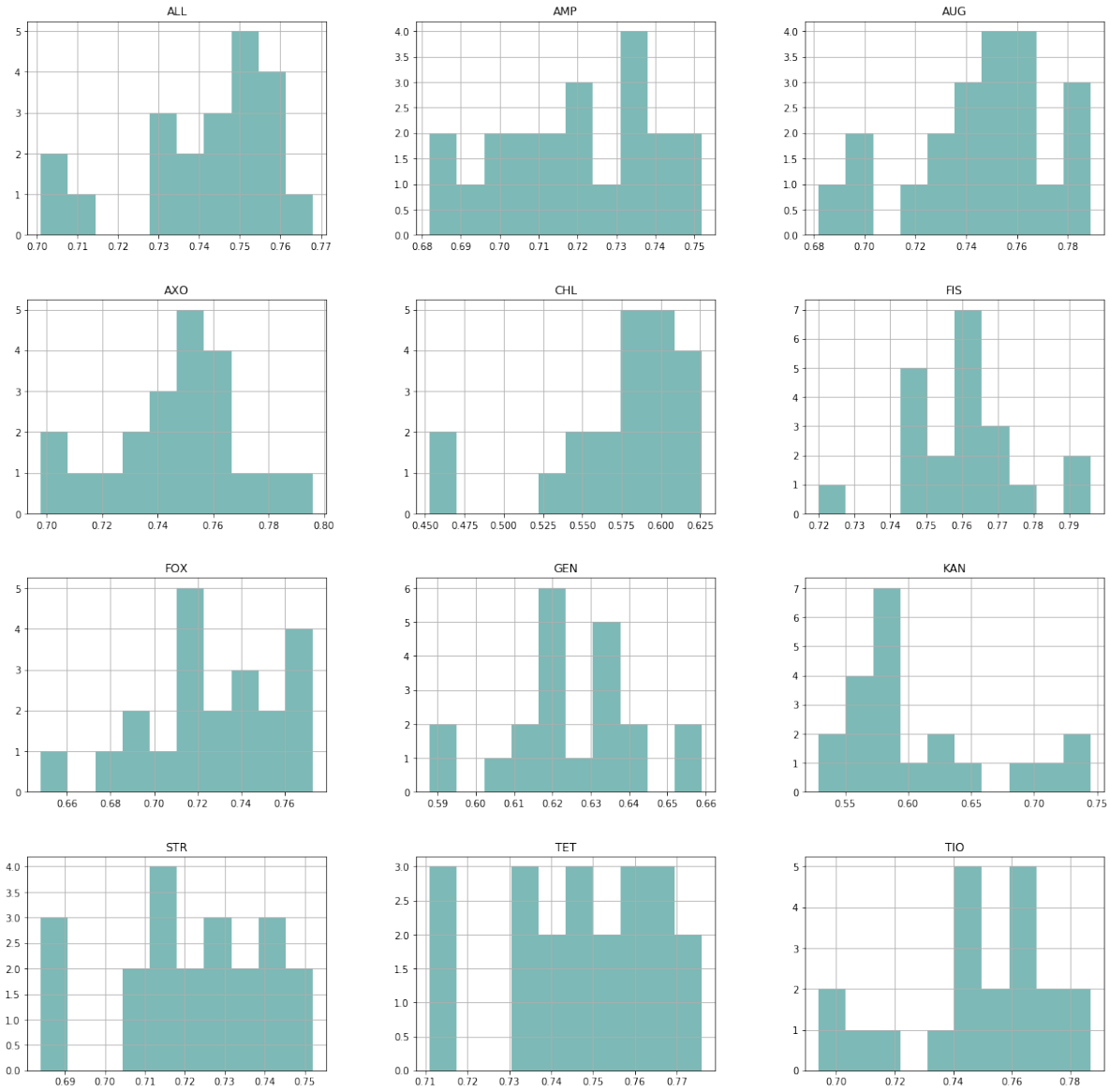

Figure 4.7: Histograms for F1 scores from 30 experiments using sets with 20 genes and a randomized group of of genes as sources.

Again, no statistically significant increasing in the performance of experiment prediction was observed compared to the distribution of randomized experiments. Possible reasons for the observed results will be discussed in the next chapter.

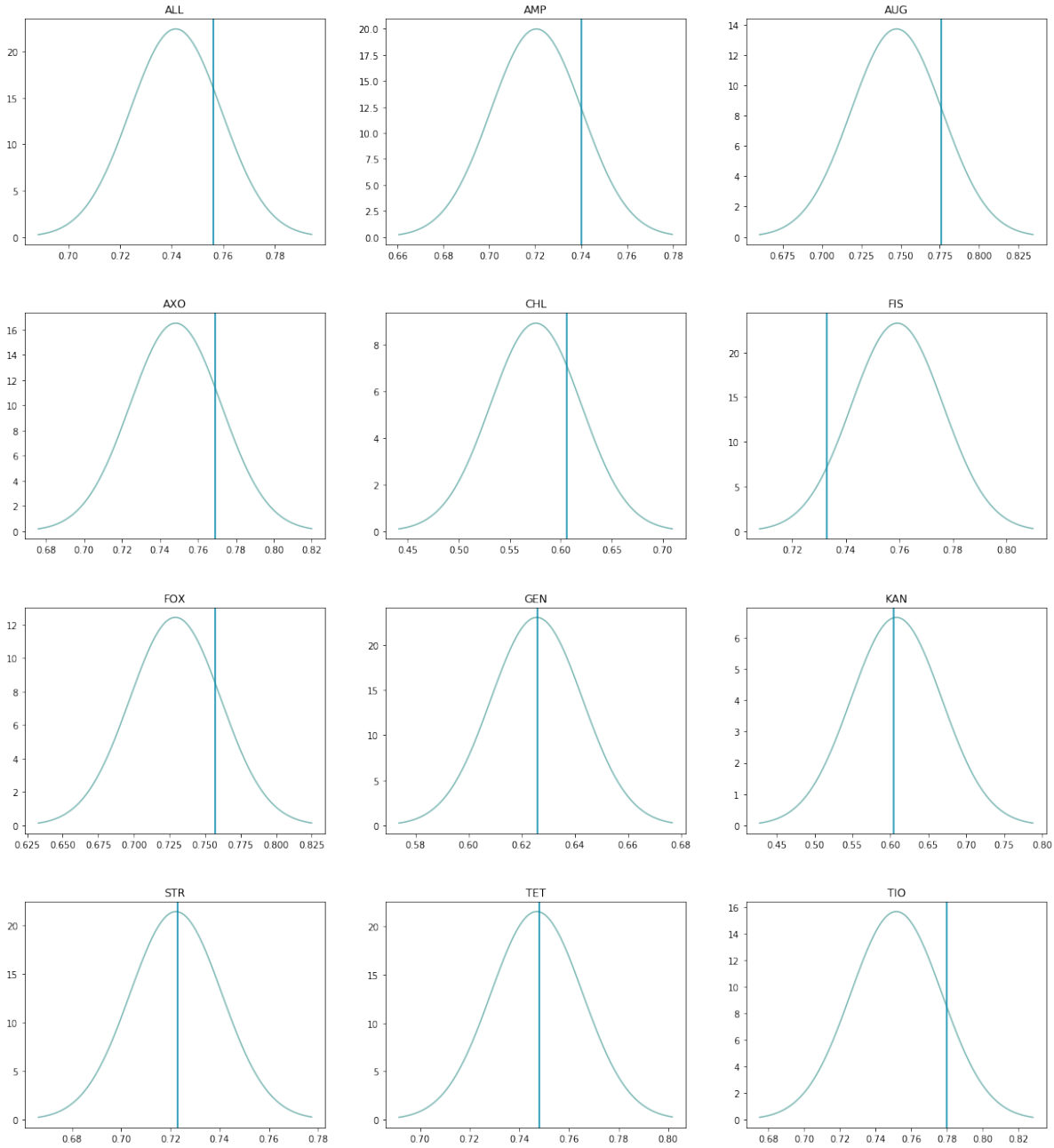

Figure 4.8: F1 scores (vertical blue line) obtained selecting a set of 20 genes with the best scores. The density of the normal distributions from randomized experiments is plotted with three standard deviation for each side around its mean.

5

Discussion

In this chapter we will discuss the implications of our findings and explore potential reasons for the observed results in the context of our hypothesis and the overall project. We will also consider the limitations of our study and provide suggestions for future research.

The primary objective of our study was to investigate whether core genes that are closely related to AMR genes in terms of physical interactions or protein complexes would yield better prediction of antibiotic resistance compared to a random selection of core genes. The results obtained did not have a significant increase in F1 scores, since the found p-values are greater than 0.1 when we compare the F1 score from the experiment using selected core genes to the normal distribution of F1 scores obtained from a randomized group of genes.

There are some possible explanations for the lack of improvement in predictive performance observed in our study:

1. **Unknown AMR genes**: It is possible that we are using, within the core 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. We found 9 AMR genes in the set of 500 used core genes, which were not annotated as AMR when the data was collected (2018, December). New genes are discovered and associated with the resistance phenotype as resistance mechanisms are studied over time.

2. **Quality of protein-protein interaction data**: The quality and completeness of the protein-protein interaction data used in our study could have affected the accuracy of our network-based approach. The PPI containing direct physical interactions is constructed based on laboratory derived experiments, while the PPI of physical complexes contains inferred interactions. Incomplete or inaccurate interaction data would result in an incomplete representation of the relationships between core genes and AMR genes.

3. **Number of core genes**: Our study used a set of core genes provided by Nguyen et al., which contains a predetermined number of core genes. A more complete set of core genes could potentially yield better prediction results.

However, the unexpected results obtained in this study could also suggest the existence of compensatory changes in bacterial genomes. Compensatory mutations or changes are genetic alterations that mitigate the fitness costs associated with resistance to antibiotics. The development of antibiotic resistance in bacteria often comes at a fitness cost, meaning that the resistant bacteria may grow and reproduce more slowly than their non-resistant counterparts in the absence of the antibiotic. Over time, the bacterium or its descendants may acquire additional mutations that help to restore fitness. These compensatory mutations could, for instance, further modify the altered target so it functions better, adjust metabolic pathways to compensate for the wasted resources, or re-balance cellular processes affected by the increased efflux activity [34].

Compensatory changes may occur in various parts of the genome. This would suggest that any part of the bacterial genome could be used to predict antibiotic resistance. Therefore, further research should investigate the effect of these compensatory changes on the selection of core genes for predicting antibiotic resistance.

6

Conclusion

In this thesis, we have investigated the potential of selecting core genes based on their relationship to known AMR genes in protein-protein interaction networks to predict antibiotic resistance. Our work involved the application of kernel methods on bacterial protein interaction networks, the selection of gene sets using these methods, and the assessment of these gene sets' predictive capabilities using a machine learning model based on decision trees. Our results did not support our original hypothesis that core genes related to AMR genes in terms of physical interactions or protein complexes would achieve higher scores in predicting susceptible/resistant phenotype compared to a random choice of core genes.

Upon further analysis, we propose that the unexpected results could be due to the possibility of compensatory changes occurring throughout the bacterial genome rather than being specifically related to AMR genes. This finding suggests that any part of the bacterial genome could potentially be used to predict antibiotic resistance.

However, it is important to note that there are limitations to our study, including the small sample size for some antibiotics, potential limitations in the quality of the PPI network used, and the use of a set of core genes that may not be complete.

We suggest increasing the sample size for each antibiotic, improving the quality of the PPI network, and using a more comprehensive set of core genes. Additionally, we

suggest further investigation into the possibility of compensatory changes occurring throughout the bacterial genome.

In conclusion, our study provides valuable insights into the potential limitations of using core genes to predict antibiotic resistance and highlights the need for further research in this area.

BIBLIOGRAPHY

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

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

[3] 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.

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

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

[6] 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.

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

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[23] R. D. et al. Olson. Introducing the bacterial and viral bioinformatics resource center (bv-brc): a resource combining patric, ird and vipr. *Nucleic Acids Research*, 2023.

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

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

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

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

[29] 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.

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

[31] 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.

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

[33] 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.

[34] R. C. MacLean and T. Vogwill. Limits to compensatory adaptation and the persistence of antibiotic resistance in pathogenic bacteria. *Evolution, Medicine, and Public Health*, 2015.