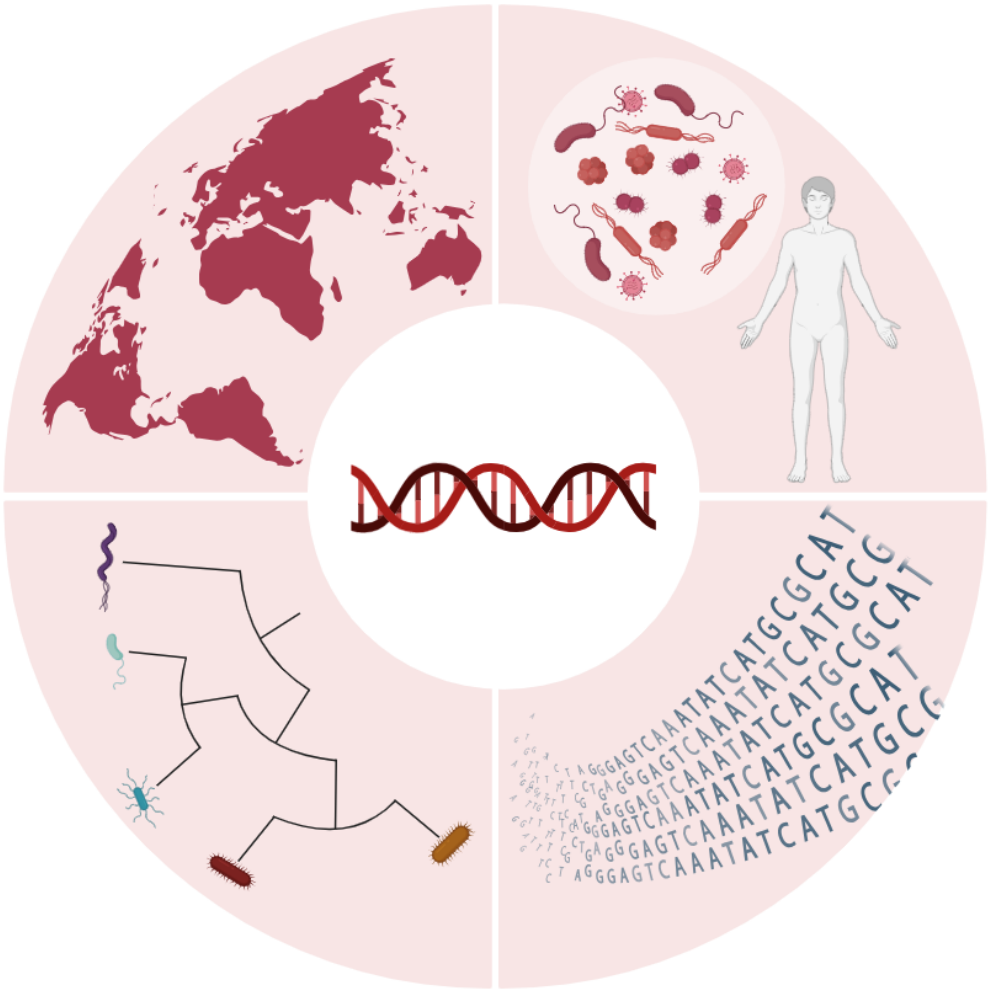
CHALMERS
UNIVERSITY OF TECHNOLOGY



What is a successful antibiotic resistance gene? A conceptual model and machine learning predictions

Master's thesis in Biotechnology

ELINOR EINARSSON & STINA TORELL

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2024
www.chalmers.se

Degree project report 2024

What is a successful antibiotic resistance gene? A conceptual model and machine learning predictions

ELINOR EINARSSON
STINA TORELL

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2024

What is a successful antibiotic resistance gene?
A conceptual model and machine learning predictions
ELINOR EINARSSON & STINA TORELL

Supervisor: David Lund, Department of Mathematical Sciences
Examiner: Erik Kristiansson, Department of Mathematical Sciences

Degree project report 2024
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Cover: Illustration of the conceptual model. Illustrated using BioRender.

Typeset in LaTeX
Gothenburg, Sweden 2024

What is a successful antibiotic resistance gene?
A conceptual model and machine learning predictions
ELINOR EINARSSON & STINA TORELL
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Antibiotic resistance is a global public health threat and it causes bacterial infections to become more difficult to treat. The spread of antibiotic resistance genes (ARGs) is predominantly driven by horizontal gene transfer (HGT) that enables bacteria to share genetic information directly between cells. The ability of an ARG to spread is influenced by a range of factors, and has become a popular field of research, aiming to find characteristics that enable rapid antibiotic resistance dissemination. This facilitates the identification of ARGs that possess the ability to disseminate rapidly, and for proactive measures against the dissemination to be implemented.

Bioinformatics tools were used to study the prevalence of 4775 known ARGs in 867 318 bacterial genomes. A conceptual model describing the success of an ARG was developed containing four different measures of dissemination, over taxonomic barriers, in different GC-environments, geographical dissemination, and dissemination to pathogenic bacteria. By using a top-down approach studying the success of a gene, the thesis complements research studying factors that characterizes successful and rapid HGT. The conceptual model resulted in a success-score for each ARG that reflected the overall performance in the four components. Among the ARGs found to be highly successful the most common class was multidrug resistance, followed by aminoglycoside, β-lactam, and MLS antibiotic resistance. Furthermore, the success-score together with information about the genes, were used to investigate the possibility to predict the success of an ARG with the use of machine learning in a binary classification Random forest algorithm. The model was built to evaluate the predictive performance using decreasing amounts of observations of each gene. As expected, the predictive performance of the model improved as the number of observation increased. Based on only one observation, it was possible to predict the class of each gene with an average sensitivity of ~70% at 90% specificity, and with 250 observations a sensitivity of 98% could be attained. Sequence related features such as gene length and codon usage were important when only a few observations of a gene were used, but as the number of observations grew, non-sequence related features such as number of countries and pathogens a gene was found in, became more relevant. A meta-analysis also aims to explore the managerial and policy implications of antibiotics resistance, and findings include that policies facilitating for machine learning are important to implement. This study can be used as a starting point in the modelling of antibiotic resistance gene success, aiming to help identify emerging ARGs that have the possibility to become future threats.

Keywords: Antibiotic Resistance, Bioinformatics, Horizontal Gene Transfer, Successful ARGs, Machine Learning, Random Forest, Managerial implications

Acknowledgements

First of all, we would like to thank our examiner Erik Kristiansson and supervisor David Lund. Thank you Erik for giving us the opportunity to do this project and for the warm welcome to your research group. Thank you David for your supervision, helping us with all our questions from programming to microbiology.

We would also like to thank the whole Kristiansson research group at the Department of Mathematical Sciences, and Joakim Larsson from the Sahlgrenska Academy for participating in the workshop, and contributing with valuable insights and lively discussions.

Finally we would like to thank our families and friends for supporting us during our time at Chalmers.

Thank you all!

Elinor Einarsson and Stina Torell, Gothenburg, January 2024

Contents

List of Figures

List of Tables

1
Introduction

Antibiotic resistance has been described as one of the main public health threats in modern history, by the World Health Organization (WHO) [1]. The crisis is global, and affects countries at all income levels. Antibiotic resistance occurs when bacteria has resistance genes or mutations in genes that allows it to survive exposure to antibiotic agents [1]. With the spread of resistance genes among human pathogenic bacteria, the risk of not being able to treat common human infections, or to receive difficult to treat infections after surgery or cancer chemotherapy, will increase. It is also estimated that in 2019, 1.27 million people died as a direct result of antibiotic resistant bacteria [2]. The excessive use of antibiotics is the main catalyst of resistance in pathogenic bacteria [1].

New forms of antibiotic resistance can arise via mutations and be inherited in mitosis under selection pressure. However, the most efficient way for antibiotic resistance genes (ARGs) to be spread is through horizontal gene transfer (HGT) that enables genetic material to be directly transferred between two bacterial cells [3]. The ability of a gene to be spread between bacteria, and become "successful" in dissemination, differs per gene, and it depends on several different factors that are not yet fully understood. The microbial environment of the host, [4], the resistance mechanism of the gene [5] and the phylogenetic relationship between the host and receiving bacteria [6] are three factors that are proposed to influence the gene transfer ability. Studies have also explored the role of the nucleotide and codon composition of the gene in influencing gene transfer, by analyzing the fitness costs associated with non-optimal codon usage [6, 7, 8]. The findings remain inconclusive and more research on the topic is needed.

The previously mentioned studies focus on factors that facilitate a successful horizontal transfer, allowing to look at the characteristics of a gene, its microbial environment, its host and DNA sequence, and judge the gene's ability to spread. There is an opportunity to complement this information by modelling the already existing resistance genes and in a top-down approach. By making assumptions of what components constitute a successful ARG on a societal and bacterial level, a model can be built that describes each gene's success. The model could serve as a tool to make predictions on the likelihood of success of emerging ARGs to set targeted measures and restrict dissemination of resistant genes.

With recent improvements in genome sequencing methods there is also a growing availability of sequencing data. Sequenced bacterial genomes and ARG sequences

available in public databases have made it possible to study the transmission of antibiotic resistance genes and facilitated the application of bioinformatics - the use of tools from data science, machine learning, mathematics and statistics to study large and complex biological systems. Furthermore, the integration of machine learning introduces the ability to make predictions based on patterns in data, for example predicting the future spread of genes, allowing for the targeted identification of genes at high risk of spread.

1.1 Aim

The thesis has two key aims, with the common goal of contributing to antibacterial resistance research, specifically regarding the ability to predict genes that will be successful in the future, to enable appropriate proactive measures to be taken.The first aim is to develop a conceptual model that describes the success of an antibiotic resistance gene. The model should give each ARG a "success-score" reflecting its overall performance in the model's constitutive components. The components will be formed based on assumptions of what defines successful ARG dissemination on a societal and bacterial level, to capture multiple perspectives. The second aim is to build a Random forest [9] machine learning model to investigate if it is possible to predict the success of a gene, as defined by the conceptual model. It will also be investigated how much information is required to classify ARGs correctly, with an acceptable degree of sensitivity and specificity. To achieve the goal, bioinformatics tools will be employed to perform sequence alignments between ARGs and bacterial genomes, and analyse taxonomic information and meta data.

The thesis contains a stand-alone section that relates to the managerial implications of antibiotic resistance. The aim of the chapter is to explore appropriate policies and management practices that facilitate for the effective use of resources which will be critical in a society where antibiotic resistance is on the rise.

2
Theory

This chapter will explain the theory behind the most important concepts related to the project. First, antibiotic resistance is introduced, including the mechanisms of antibiotic resistance gene transfer and factors that can affect this transfer. An explanation of important antibiotic classes, and a review of the resistance mechanisms against the antibiotics are also included. Secondly, the chapter explains the use of machine learning in the area of bioinformatics with focus on the Random forest algorithm [9].

2.1 Antibiotic resistance and gene dissemination

Antibiotic resistance is the phenomenon when bacteria expresses mechanisms to survive exposure to an antibiotic agent. The resistance to antibiotics can be either natural or acquired. Natural resistance includes intrinsic resistance where the genetic material that confers resistance is always expressed in a species. Induced resistance refers to when the gene naturally occurs in the species, but is only expressed after antibiotic exposure [10]. Like any other new trait, antibiotic resistance can also be acquired as a result of mutations and be passed down through mitosis under selective pressure [3]. However, the most effective mean of spreading acquired antibiotic resistance genes is through a process known as horizontal gene transfer (HGT).

2.1.1 Horizontal gene transfer

HGT is the process where genetic material is transferred horizontally between individual bacteria [3, 11]. As a bacterial cell receives new foreign DNA, it can adapt to new environments and get a competitive edge over other organisms [11]. The existence of HGT also means that genetic material, such as genes encoding antibiotic resistance, can spread rapidly between different species of bacteria [3].

Transduction, transformation, and conjugation are the main types of HGT where the latter is the mode most common. It involves the formation of conjugative pili that connects two living bacteria and allows for transfer of genetic material through direct contact between the cells. Transduction occurs when genetic transfer is mediated by a phage (viral or bacterial), and finally transformation happens when the DNA from lysed cells is taken up from the extracellular environment [11]. The mechanisms of HGT are represented in figure 2.1.

There is evidence that HGT is strongly driven by common phylogeny, meaning that genes often spread more easily among closely related species [12, 13]. A possible explanation to this is that species that share common ancestry also share common mechanisms of exchanging materials [14]. Another plausible explanation for HGT being more prominent between closely related species is the impact of gene fitness cost, where HGT is favoured for a gene that is more compatible with the host genome [15]. Tuller *et al.* [8] mentions that GC-content can impose constraints on codon usage, resulting in incompatibility of the host and gene pair. It is also known that GC-content correlates with the phylogenetic distance [6].

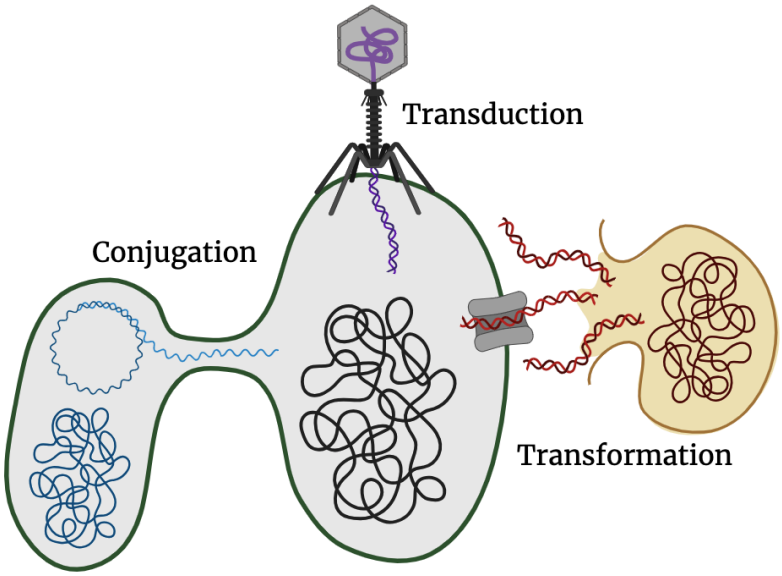


Figure 2.1: *Graphic representation of the main types of horizontal gene transfer: transduction, transformation, and conjugation. Illustrated using BioRender.*

HGT can be detected through a sliding window method, where the GC-content of a genomic section is compared with the typical content of the entire region [16]. A horizontally transferred gene will adapt its GC-content to the new host context at a rate of 0.045-0.91% per million years, therefore if any gene section is remarkably different compared to the bacterial genome, it is likely horizontally transferred [5]. Sanchez-Osuna *et al.* [5] uses this fact, and examines the GC-dissemination of various gene groups such as *NDM* β-lactamases and *qnr* quinolones, by comparing the GC-content of genes with the GC-content of the host. The authors showed that 18.7% of the gene groups had "dissemination bands", meaning that a gene group is found in a range of host GC-contexts. Strikingly, 42% of those gene groups had been previously reported as having widespread dissemination. The authors conclude that the method of detecting GC dissemination bands, therefore could be a tool to detect widespread HGT.

2.1.2 Antibiotic classes and their mechanisms of action

There are multiple types of antibiotics, often classified by the mechanism of action to kill bacteria [10]. Consequently, the antibiotic resistance genes are often

classified based on the antibiotic class they confer resistance against. Acquiring an understanding of these mechanisms of action is therefore important when studying such genes. The most common targets for antibiotics are cell wall synthesis, protein synthesis, nucleic acid synthesis, the cell membrane and various metabolic pathways [10]. β-lactam antibiotics are prominent antibiotics that target the cell wall synthesis. By interfering with the formation of cross links between the peptides in the peptidoglycan chain, the cell wall is weakened which leads to cell lysis [17].

For protein synthesis there are two key subgroups of antibiotics, those that target the 30s ribsomal sub-unit and those that target the 50s sub-unit [17]. Aminoglycosides belongs to the 30s group, and their action results in premature termination and misreading of mRNA. The tetracyclines also target the 30s unit, but act by preventing ribosomal binding to tRNA. On the other hand, the chloramphenicols also counteract the binding of tRNA but to the ribosomal 50s sub-unit. Macrolides, lincosamides, and streptogramins, often grouped as MLS, target the 50s unit interfering with the peptidyl transferase center resulting in incomplete peptide chains being detached [17].

The quinolones, and specifically the flouroquinolones, is the most common group of antibiotics that interfere with the nucleic acid synthesis. They target the DNA gyrase and topoisomerase IV enzymes that normally act to reduce strain on the DNA as the strands separate [17], and instead transforms them into enzymes that excessively nicks and fragments the bacterial genome [18]. The aminocoumarins is another antibiotic group that targets DNA gyrase. It prevents the binding of ATP to the enzyme, which in turn hinders DNA replication and transcription [19]. Apart from nucleic acid synthesis, some antibiotics can cause cell death by damaging the DNA helix structure. These antibiotics are called metronidazoles and are a common subtype of nitroimidazoles [20].

Sulfonamides and diaminopyrimidines are two classes of antibiotics that target the folic acid metabolic pathway [21]. The disruption of folic acid synthesis hinders other important pathways, for example the synthesis of the nitrogenous bases and various amino acids. More specifically, sulfonamides are competitive inhibitors of the dihydropteroate synthase enzyme, that participates in the first regulation step in the biosynthesis of folate. The inhibition leads to cell death due to lack of thymine. Additionally, the target of folate pathway by the diaminopyrimidines happens through competitive inhibition of dihydrofolate reductase [21].

Another interesting class of antibiotics is peptide antibiotics. These can have a variety of mechanisms and can be both natural, as a part of a bacterium's natural defense system, and synthetic [22]. An example is lipopeptides, a group of antibiotics that work to depolarize the cell membrane [10]. Finally, separate from the traditional antibiotics, there are disinfectants and anti-septics that are used on the skin and on hard surfaces, for example in hospitals [23].

2.1.3 Resistance mechanisms of antibiotic resistance genes

As previously mentioned, antibiotic resistance genes can be grouped by the antibiotic drug they confer resistance against. For example the *tet*-genes encodes proteins that confer resistance to tetracycline antibiotics. Genes can also be classified based on their mechanism of resistance. This section will outline the most common antibiotic resistance mechanisms: limiting uptake, antibiotic inactivation, target modification, and efflux pumps [10].

Limiting antibiotic uptake is a mechanism where the bacteria hinders the antibiotic from entering the cell. As an example, *Staphylococcus aureus* has acquired genetic material that produces a thicker cell wall, so that certain vancomycin antibiotics cannot enter. Another example is bacteria that regulate the number, or selectivity, of porins in the outer membrane as a way to stop the uptake of drugs [10]. The cell can also alter the drug, and cause its inactivation. For example, the cell can hydrolyze the drug, or add an acetyl or phosphoryl group. A common type of antibiotic resistance enzymes are the β-lactamases that hydrolyze the β-lactam antibiotic leading to drug inactivation [10].

By target modifications, the cell alters its own components interacting with the drug. For example, vancomycin ARGs modify the peptidoglycan layer which limits the glycopeptide antibiotics ability to disrupt it [24]. Similarly, the antibiotic resistance gene *tet(W)* acts by altering the conformation of the ribosome and protects it from binding with tetracycline [25]. Lastly, bacteria can also use efflux pumps to transport the drug out of the cell. Efflux pumps are encoded by chromosomal genes, as they are used to transport out toxic chemicals from the cell, and can be mutated to confer a high resistance to antibiotics [10]. One example of an efflux pump encoding gene is *tet(A)* that pumps out tetracycline from the cell [26]. Apart from other mechanisms of action, efflux pumps alone can make the bacteria multi drug resistance. This makes effective efflux pumps exceptionally good traits for bacterial survival [27].

2.2 Machine Learning tools in bioinformatics

Bioinformatics is a fairly new research field, emerging from the prosperity of next generation sequencing technologies [28]. Methods aim to find patterns in large amounts of biological data, and have made it significantly less challenging to study genes, such as ARGs, on a large scale. The use of machine learning algorithms in bioinformatics, adds a dimension of prediction and interpretation, making it a popular field of research [29]. One type of machine learning algorithm used for classification is the Random forest first described by Breiman [9], which has shown great potential when it comes to classify biological data. For example Wang *et al.* [30] managed to predict protein-protein interaction sites with a higher accuracy than other tools, and Le *et al.* [31] presented the Random forest as the best in predicting disease-gene associations. The details of the algorithm will be described in the following sections.

2.2.1 Building a Random forest algorithm

Decision trees are easy and robust tools when it comes to data mining, the identification of patterns in large volumes of data [32]. A major application area for decision trees is object classification, where the nodes in the tree represent features that describe each object in the data set. The Random forest algorithm is a classifier that assembles a large number of decision trees, each voting for a class of the object. Subsequently, the conclusive class is chosen based on the majority of the decision tree votes. Hence, the larger the majority the more confident is the classifier [9]. A basic representation of the Random forest algorithm is visualised in figure 2.2.

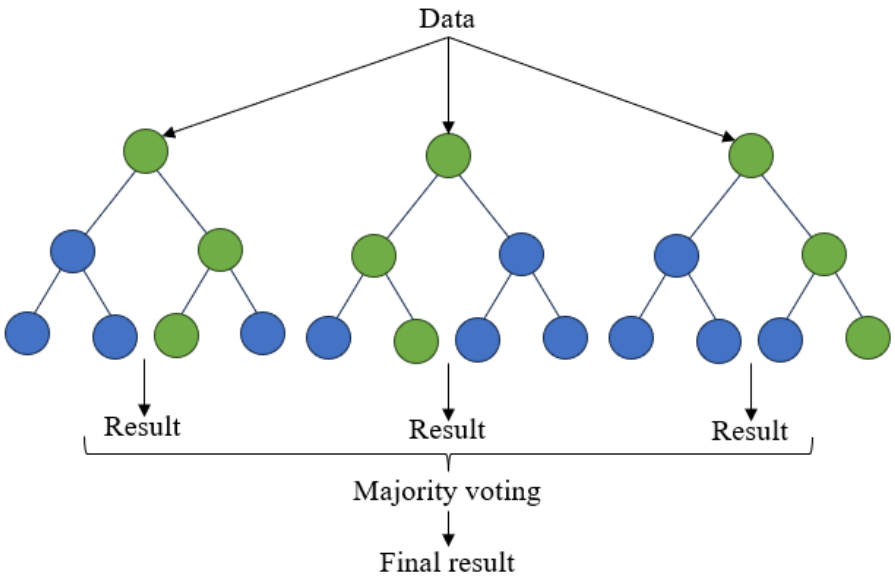


Figure 2.2: *Graphic representation of a Random forest algorithm. A large number of decision trees are assembled, each voting for a class of the object. The majority of the classes voted for is selected as the conclusive class [9].*

Each tree in the Random forest is built using bootstrap data set, which are data points randomly selected from the original data with replacement. The trees are grown node by node starting from the top and for each split in the tree, randomly selected features will be chosen as candidates for the split [9]. The number of selected features can be varied, but it has been found that the square root of the total number of available features often produces near optimal results [33].

To select the best fit feature for each split, τ, the Gini impurity, $i(\tau)$ for each leaf in the split is calculated with

$$i(\tau) = 1 - p_1^2 - p_0^2 \tag{2.1}$$

where $p_k = (N_k/N)$, N_k is the number of samples in a binary classification classified as $k = \{0, 1\}$ from the total number of samples, N. The total Gini impurity for the two leafs, a and b, of the split is further calculated as

$$i_{tot}(\tau) = \frac{N_a}{N} \times i_a(\tau_a) + \frac{N_b}{N} \times i_b(\tau_b). \tag{2.2}$$

Here N_a and N_b are the sample size in leaf a and b respectively. For each of the feature candidate the total Gini impurity, $i_{tot}(\tau)$, is calculated and finally the feature with the lowest $i_{tot}(\tau)$ is chosen [34]. There is no pruning in the construction of the trees meaning that all trees will be built to their largest extent [9].

Another way to describe the importance of a feature is with the Mean Decrease Accuracy (MDA). MDA measures, for each feature, how much accuracy is lost if the feature is removed. Thus, the higher MDA the higher impact of the feature on classifying the objects correctly [9].

2.2.2 Predictions and Confusion Matrices

In a binary classification model data points will be assigned a probability to belong in each of the two classes. In a Random forest this is based on the voting of the trees. Together with a probability cutoff, the data points are predicted to belong to a class. Figure 2.3A illustrates tests X, Y and Z, their probability to belong in each class, and the resulting prediction at two different positive probability cut-offs. Given that the actual class of each data point is known, there are four possible outcomes of the model, true positive, false negative, true negative and false positive [35], illustrated as a confusion matrix in figure 2.3B. The prediction and confusion matrix varies with the probability cutoff.

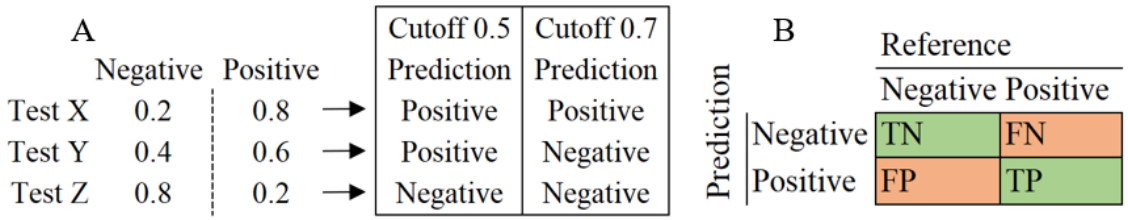


Figure 2.3: *A) An illustration of a binary classification model, and the resulting class prediction of tests X, Y and Z at two probability cutoff points 0.5 and 0.7. Note that a cutoff of 0.7 means that that the probability of a positive test must be >0.7 for the test to be classified as positive. B) A confusion matrix representing the outcomes of a prediction.*

To evaluate the performance of the machine learning model, the concepts sensitivity and specificity are used. The sensitivity is the probability that a reference positive test is classified as a true positive, TP, and not as a false negative, FN. [35]. It is calculated based on the confusion matrix as

$$Sensitivity = \frac{TP}{TP + FN}. \tag{2.3}$$

On the other hand, specificity is the probability that a reference negative test is classified as a true negative, TN, and not as a false positive, FP [35], and calculated as

$$Specificity = \frac{TN}{TN + FP}. \tag{2.4}$$

2.2.3 Receiver Operating Characteristics curve

A Receiver Operating Characteristics curve (ROC curve) is a visualization of a model's performance in terms of sensitivity and specificity at all probability cut-off points [35]. On the Y axis is the sensitivity of the model, also called the true positive rate, and on the X axis is 1 - Specificity, also known as the false positive rate. [35]. Figure 2.4 illustrates example ROC curves, where the darker the curve, the better is the overall performance. A ROC curve along the black dashed line $X = Y$ imply that the model's performance is poor, and predicts with low sensitivity and specificity. The optimal model would pass through the black point, indicating that it can predict with full sensitivity and a false positive rate of 0 [35]. The area under the curve (AUROC) gives the overall performance of the model, and is a number between 0 and 1. In figure 2.4 the AUROC of the light blue curve is filled in.

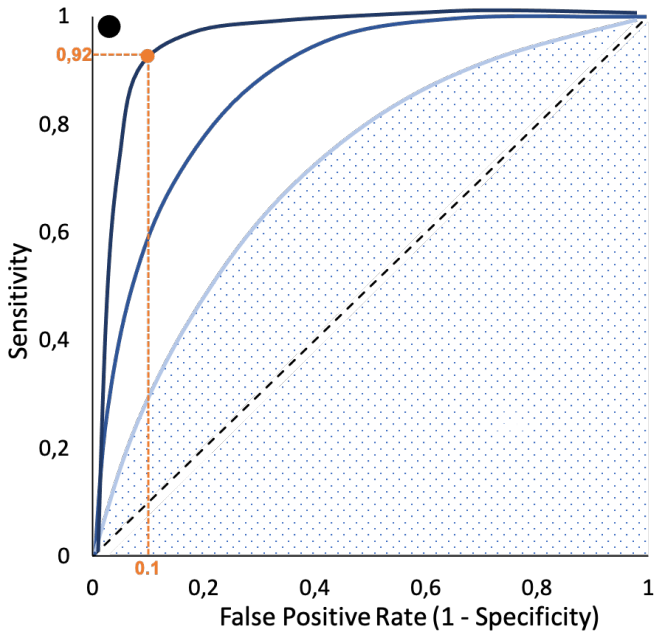


Figure 2.4: *A representation of ROC curves. A darker curve implies an overall better performance of the model. The orange line illustrates how to read the sensitivity and specificity at a point along the curve.*

The ROC curve also illustrates the performance of a model at a given false positive rate. For example, the orange lines in figure 2.4 show that at a false positive rate of 0.1 (90% specificity), the dark blue model has a sensitivity of ~0.92.

3
Methods

In this chapter, the methods for the construction of the conceptual model will be explained in detail, including data collection and processing, and calculation of the different components of the success-score. The method used to implement the Random forest to predict ARG success will also be described. The chapter is ended with a description on how the managerial implications of antibiotics resistance are explored in the thesis.

3.1 Data collection

4775 antibiotic resistance genes were collected from The Comprehensive Antibiotic Resistance Database (CARD) (date accessed: 2023-08-29) [36], and 867 318 bacterial genomes from National Center for Biotechnology Information (NCBI) Assembly (date accessed: 2022-04-04), [37]. The genomes that did not pass NCBI taxonomy check, or where contamination was suspected, were removed to secure robustness of the result.

A database was constructed from the ARG sequences using a built-in function in blast for bioconda v.4.8.2. A parallel nucleotide blast was conducted using blastn in BLAST v.2.5.0 with the genomes as query and ARG database as subject [38, 39]. The output contained 121 346 839 alignments and was filtered on percent identical $> 90\%$ and subject coverage $> 70\%$. In the case of overlapping genes, the gene with highest percent identical, or if equal percent identical highest subject coverage, was kept. For genes with indistinguishable percent identical and subject coverage, random selection was implemented. The same approach was executed for genes that were encountered more than once in the same genome. After the filtrating 88 999 443 alignments were kept containing 4029 ARGs.

3.2 Conceptual model and score generation

A conceptual model to describe the success of an ARG was developed and the general workflow from raw data to the success-score is presented in figure 3.1. The model was developed containing four separate components that each generated a component score. The four components were taxonomic spread, spread in different GC-contexts, geographical spread, and presence in pathogens. A workshop was held together with scientists from *Chalmers University of Technology* and *Gothenburg University* to discuss and improve the model.

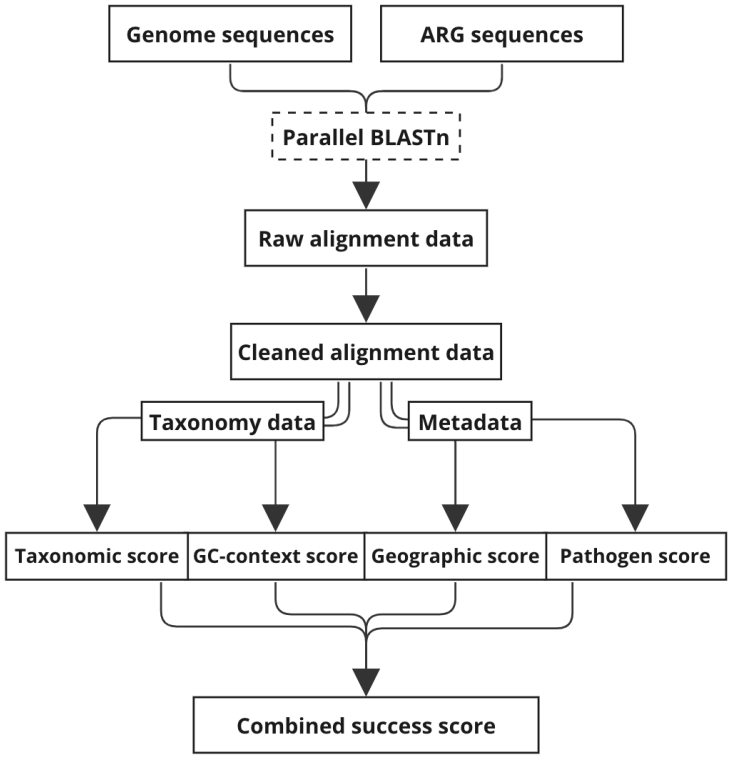


Figure 3.1: *Graphic representation of the workflow. Boxes with continuous borders represent files and dashed borders implies a process.*

3.2.1 Taxonomic dissemination score

The taxonomic dissemination score was generated by combining the cleaned BLAST alignment file with the taxonomic classification of each genome, received from NCBI (date accessed: 2022-04-04) [37]. The NCBI taxonomy is based on phylogenetic information and represents the evolutionary relation between species as it is understood today [40]. The maximum taxonomic distance between the species that each ARG was found in was calculated.

To calculate the taxonomic distance a phylogenetic tree was built based on the taxonomic information given for each bacterial species. A simple example of a phylogenetic tree including four bacterial species is displayed in figure 3.2. In the tree built for the conceptual model, all species that were used in the BLAST were included. To calculate the taxonomic distance between two species that an ARG was found in, their closest common ancestor according to the taxonomic information was identified, and the number of nodes required to reach the common ancestor was counted. The closest common ancestor of the two species *Alicyclobacillus acidiphilus* and *Aerococcus viridans* is *Bacilli* in the taxonomic level Class, which gives a taxonomic distance of 4. While the taxonomic distance between *Alicyclobacillus acidiphilus* and *Alicyclobacillus mengziensis* is 1, since their closest common ancestor is the Genus *Alicyclobacillus*. A gene found in all four species displayed in the example would therefore get a maximum taxonomic distance of 4. Note that in the model it

is assumed that the distance is equal between all taxa. In reality, the distance can vary greatly, and two taxonomic groups can be very similar, or very different.

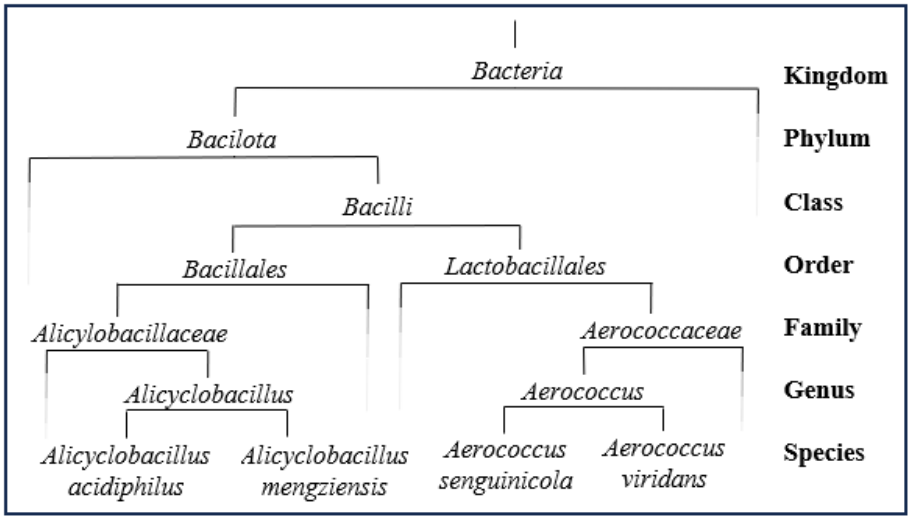


Figure 3.2: *Representation of a simple phylogenetic tree for four selected bacterial species. The levels of the taxonomy is presented to the right of the tree.*

The maximum taxonomic distance for each ARG was then multiplied with the total proportion of species that the gene was found in to premier ARGs that are abundant in a diverse host range. The total taxonomic dissemination score, T_S, was calculated as

$$T_S = max(T_d) \times (N_{species}/N_{species,tot}) \tag{3.1}$$

where $max(T_d)$ is the maximum taxonomic distance of the bacterial species. $N_{species}$ represent the number of different species in which the ARG is found, and the $N_{species,tot}$ is the total number of species used for the blast and equals 21992. Genomes that had incomplete taxonomic lineage were removed from the analysis. The analysis resulted in 4028 ARGs with a taxonomic score.

3.2.2 GC-context score

A GC-context score per ARG was calculated based on the BLAST alignment. The score is a combination of descriptive statistics that describe the spread of GC-content of the host species that an ARG is found in. A high score means that the ARG is compatble with existing in a range of environments. The GC-content of all genomes that an ARG had aligned with was calculated, and a representation of the output is visible in table 3.1. An average of the GC-content for each species per ARG, was then calculated. This is represented in table 3.2. For instance, the average of the *E. coli* genomes was calculated for ARG X and ARG Y separately.

Table 3.1: *Example extract from the blast result. ARG X is found in genomes A, B, C and D, where A and C belong to E. coli, B to K. pneumoniae and D to S. enterica. The GC-content of the genomes is presented in the table. Note that both ARG X and ARG Y are found in genome C. This means that genome C contained two different resistance genes.*

ARG	**Genome**	**Species**	**GC-content**
ARG X	Genome A	*Escherichia coli*	50.6%
ARG X	Genome C	*Escherichia coli*	50.4%
ARG X	Genome B	*Klebsiella Pneumoniae*	57.4%
ARG X	Genome D	*Salmonella enterica*	52.4%
ARG Y	Genome C	*Escherichia coli*	50.4%

Table 3.2: *Example table after taking the average GC-content per species per ARG. Note that in the actual results file, each ARG could be found in up to hundreds of species.*

ARG	**Species**	**GC-content**
ARG X	*Escherichia coli*	50.5%
ARG X	*Klebsiella pneumoniae*	57.4%
ARG X	*Salmonella enterica*	52.4%
ARG Y	*Escherichia coli*	50.4%

After calculation of the average GC-content of each species an ARG was found in, the descriptive statistics were calculated to describe the GC-context range for each ARG separately. The inter quartile range (IQR) as well as the difference between the 90^{th} and 10^{th} percentile and the maximum difference in GC-content were first calculated as

$$\begin{cases} IQR = 75^{th} \text{ percentile value} - 25^{th} \text{ percentile value} \\ R_{90,10} = 90^{th} \text{ percentile value} - 10^{th} \text{ percentile value} \\ R_{max,min} = \text{ maximum value} - \text{ minimum value.} \end{cases} \tag{3.2}$$

The sample standard deviation, STD, and median absolute deviation, MAD, of the GC-content was also calculated using

$$\begin{cases} STD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \\ MAD = \text{Med}(|x_i - \text{Med})|) \end{cases} \tag{3.3}$$

where N is the number of individual GC-content values x_i, and $\overline{x}$ the mean of those. The median absolute deviation is the median of the absolute value of the difference between each GC-content value x_i and the median of the sample, $\text{Med}(x)$.

Lastly, a Wilcoxon rank sum test was performed on the distribution of GC-context per ARG. This was done to statistically measure if the GC-context of each ARG

could represent the total "background" distribution of GC-context of all species. The background distribution was found by calculating the average GC-content of genomes per species. A gene spread over multiple GC-contexts representing the background distribution, that does not discriminate based on GC-content, should be considered successful. Because of the formulation of the problem, the aim was to look for non-significance, since this meant that the distribution was representative of the background, thus

$$\begin{cases} \mathbf{H}_0 = \text{The background and the ARG sample have the same continuous distribution of GC-context} \\ \mathbf{H}_1 = \text{The GC-context distribution of an ARG is different from the background distribution.} \end{cases} \tag{3.4}$$

The generated *p-value* was the final contribution to the GC-context score describing the distribution of GC-context.

Each statistics *STD, MAD*, $R_{max,min}$, *IQR*, $R_{90,10}$, and *p-value* was transformed with a min-max transformation, to ensure all statistics having a similar weight by subtracting the lowest value and dividing by the total range. This scales each statistic to a [0,1] range where the previously highest value would attain a value of 1, and the lowest 0.

$$x_{transformed} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{3.5}$$

x_i is the data point to be scaled, x_{min} is the lowest value of the statistic, and x_{max} is the highest value of the statistic.

The aggregation was performed with the L2-norm, which is the square-root of the sum of the squares of each individual statistic. The final GC-score, GC_S, for an ARG is

$$GC_S = \sqrt{STD^2 + MAD^2 + (R_{max,min})^2 + IQR^2 + (R_{90,10})^2 + (p-value)^2}. \tag{3.6}$$

The GC-score requires the gene to be found in at least two species, otherwise the GC-score is set to 0. 2511 ARGs received a score greater than 0.

3.2.3 Geographic dissemination score

A geographic dissemination score was generated by combining metadata from NCBI (date accessed: 2023-08-29) [37], containing information about the geographic location of the collected bacteria, with the blast alignment file. The combined file was filtered and data with mislabeled countries were removed. From the original 867 318 genomes, 758 928 were kept after the filtering. The continent for each country was added using the *pycountry* package v.22.3.5 [41] in python. For each ARG the

unique number of countries, $N_{countries}$, and continents, $N_{countries}$, were calculated, hence

$$G_S = N_{countries} \times N_{continents} \tag{3.7}$$

where G_S is the geographic score. Of 4029 ARGs, 3899 were annotated correctly and received a geographical score.

3.2.4 Presence in pathogen

The conceptual model also included a presence in pathogens component. The score was calculated based on the number of pathogens that an individual gene was found in. This represents the pathogen-score for the ARG and the more pathogens an ARG is aligned with, the higher is the pathogen score, P_S.

$$P_S = N_{pathogen} \tag{3.8}$$

The pathogens included came from the PathoSystems Resource Integration Center (PATRIC) [42] and is a comprehensive list of 227 pathogenic bacteria.

3.2.5 Aggregation of dissemination score

The four separate dissemination scores were individually transformed with a Z-transformation, this ensured all components contributed equally to the final score. The Z-transformation is conducted using the following formula,

$$Z_i = \frac{S_i - \overline{S}}{s} \tag{3.9}$$

where Z_i is the new z transformed score, S_i is the score of sample i, $\overline{S}$ is the mean of all scores and s is the sample standard deviation. The aggregated score, S_{tot}, was then generated by taking the median value of the four separate scores,

$$S_{tot} = M(T_S, GC_S, G_S, P_S). \tag{3.10}$$

3.3 Building the Random forest model

A classification Random forest was developed to investigate the ability to predict the success of an ARG. The general workflow consisted of assigning a binary class to each ARGs based on the success-score, generating features from alignment data and ARG sequences, and finally constructing a Random forest model.

3.3.1 Generation of data for the predictive model

To resemble a real life scenario, random sampling of the data was done to be used in the Random forest algorithm. Different scenarios were accounted for, were up to 1, 5, 10, 20, 50, 100, and 250 observations of each ARG were randomly selected from the filtered blast alignment file. The features were then calculated based on only the

randomly selected observations. The features can be divided into five groups (gene, geographic, GC, taxonomic and pathogen features) that are summarized in table 3.3. The one observation scenario contained only gene and pathogen features. The majority of the features were calculated in the same way as in the score generation and they will not be described again.

Table 3.3: *A compilation of all features used to describe the ARGs, divided per feature group.*

Gene features	**Geographic features**	**GC features**
Gene length	Number of countries	IQR
Gene GC-content	Number of continents	$R_{90,10}$
Codon usage		$R_{max,min}$
		Mean GC-content
Taxonomic features	**Pathogen features**	STD
Taxonomic distance	Number of pathogens	MAD
Number of species		

The gene features included the length, the GC-content, and the codon usage of the ARGs calculated from the gene sequence data. The codon usage was calculated by first finding the start codon (ATG, GTG, TTG, CTG, or ATC), and for that reading frame calculate the ratio of each codon. For the codon 'AAA' the ratio was calculated as

$$Ratio_{AAA} = \frac{N_{AAA}}{N_{tot}} \tag{3.11}$$

where N_{AAA} is the number of 'AAA' codons, and N_{tot} is the total number of codons in the gene.

The taxonomic, GC, geographic and pathogen features were, compared to when generating the scores, kept separate and not aggregated. The taxonomic features included the max taxonomic distance, and the number of species a gene was found in, unlike the score were it was divided with the total number of species. The GC features included the $STD, MAD, R_{max,min}, IQR$, and $R_{90,10}$ calculated in the same way as for the score generation, and the mean value of the GC-context values. Note that the features were not min-max transformed, as was the case for score generation. Additionally, the Wilcoxon rank sum test *p-value* is not included as a feature. This is because the background distribution from the blast result is not relevant in a real world scenario as there is no "reference" background.

One implication of this method of generating features was that when selecting only for few observations, it was likely that some features could not be generated due to the fact the most of the features needed the ARG to be found in at least two different species. When this happened this feature was ignored for that ARG.

3.3.2 Implementation of the model

Following the development of the conceptual model, and the assignment of a success-score to each ARG, a set of binary classifier Random forest models were implemented. This was done to investigate if it was possible to predict if an ARG would have a high or low success-score and how much information that would be required to do so with an acceptable sensitivity and specificity.

The Random forest models were implemented in three key stages that fit the exploratory nature of our thesis. Firstly, a general screening and exploration of the model was done to test its behavior under different conditions. The model's performance under various conditions was explored. For example, the sample sizes, the number of input features for a decision tree, and the cutoff for the binary classification were altered. Secondly, the models were built and will be described in detail below. Thirdly a cross validation of the models were performed and the predictive performance assessed. The possibility of a three-class classification model was also explored, but was decided against evaluating it in detail. This exploration enabled an understanding of how the model works, and what performance that could be expected.

The models were built using the *RandomForest* package v.4.6-14 [43], in R. Two models were built, in the first model all available features were included, and in the second model only the features with MDA above 1.5% were included as features. The table with MDA values was automatically generated from the *RandomForest* package. To explore the amount of information needed to predict the success, both models were built on a range of observations, from 1 observation up to 5, 10, 20, 50, 100, 250 and all available observations.

The model had a set of parameters that we explored during the first stage of the process, and kept constant during the model building. The rest were set to default as per the *RandomForest* package. The binary classification cut-off was set to 500, meaning that the 500 ARGs with the highest success-score were assigned into class "1", and the remaining ARGs were assigned to class "0". During the screening stage, it was found that the model was quite sensitive to the large class imbalances in the data set. To reduce this imbalance between class '1' and '0' in the training data set, the smaller class ('1') was up-copied to the double , and the larger class ('0') was down-sampled to half of the original. The sampling was done by random selection from the original data with replacement.

The parameter that decides the number of variables randomly selected at each node was set to the square root of the total number of features. In the *RandomForest* package, this parameter is called mtry. The manual user guide of the package mentions that Random forests are generally not very sensitive to the exact value of the mtry, as long as it is approximately right, and that square root of the number of features often produces near optimal results [33]. This was also explored during the screening stage.

In the third stage of the model implementation, the performance of the model was evaluated. ROC-curves were plotted using the *pROC* package v.1.18.0 [44], and confusion matrices were generated using the *caret* package v.6.0-90 [45]. The models were also cross validated using a five-fold cross validation to understand how the performance depended on the selection of training and testing data. Furthermore, a deeper analysis of the five observation data set was done by generating additional replicates of five observations.

3.4 Managerial implications - meta analysis

This section is based on a meta-analysis of current literature and reports. The first step is an analysis of the current resistance situation in the two case countries Italy and Sweden in terms of resistance data, prescription rates and public awareness surveys. The second step involves analyzing the cost implications of antibiotic resistance, focusing on a hospital-level analysis to identify the specific cost items affected by antibiotic resistance. Finally, the two analyses serve as the foundation for a discussion and reflection on management practices and policy-setting.

The section focuses on the bacteria-human interface, prescriptions to humans. As mentioned in the introduction, there are many possible interfaces for antibiotics and bacteria, all of which must be managed through policy and regulation, but these will not be addressed in this chapter. Furthermore, the focus will be on the direct costs associated with antibiotics resistance on a hospital level. There are also other types of costs such as indirect costs from the loss of the sick patient from the workforce, and opportunity costs when research efforts must be focused on development of better antibiotics regimens, instead of on other diseases. These costs are outside the scope of this chapter.

4
Results

The following chapter contains the results of the two sections of the project. Firstly, the results of the conceptual model will be presented, and compared with external data sets. This is followed by the results of the machine learning implementation. The results of the stand-alone chapter relating to managerial implications of resistance are also presented in this chapter.

4.1 The four components of the conceptual model

The conceptual model contains four components that are built upon assumptions of what defines a successful ARG, and that intend to capture different perspectives of what constitutes success. The assumptions are that a successful ARG has the ability to spread

- over large taxonomic distances and to a variety of different species,
- to a broad spectrum of GC-contexts,
- globally over a large number of countries and continents, and
- to many different pathogenic bacteria.

Hence, the components are: taxonomic spread (T_S), spread in GC-context (GC_S), geographical spread (G_S), and spread in pathogens (P_S), and will be described briefly in this section. An illustration of the conceptual model and the components is presented in figure 4.1. A final success-score is generated by aggregating the scores of the individual components. A comprehensive description of the components, how they were generated and aggregated is found in the methods section.

For the taxonomic component of the conceptual model, we assume that a successful ARG has the ability to become widespread across a diverse host range, and over large taxonomic distances. The assumption is based on the evidence that dissemination is strongly affected by common phylogeny [12, 13], as described in the theory section. A gene that is spread over a large distances should therefore be considered successful since it managed to overcome phylogeny barriers. The score, (T_S), is calculated as

$$T_S = max(T_d) \times (N_{species}/N_{species,tot}) \tag{4.1}$$

where T_d is the maximum taxonomic distance of two bacterial species (see Methods). $N_{species}$ represent the number of different species in which the ARG is found, and

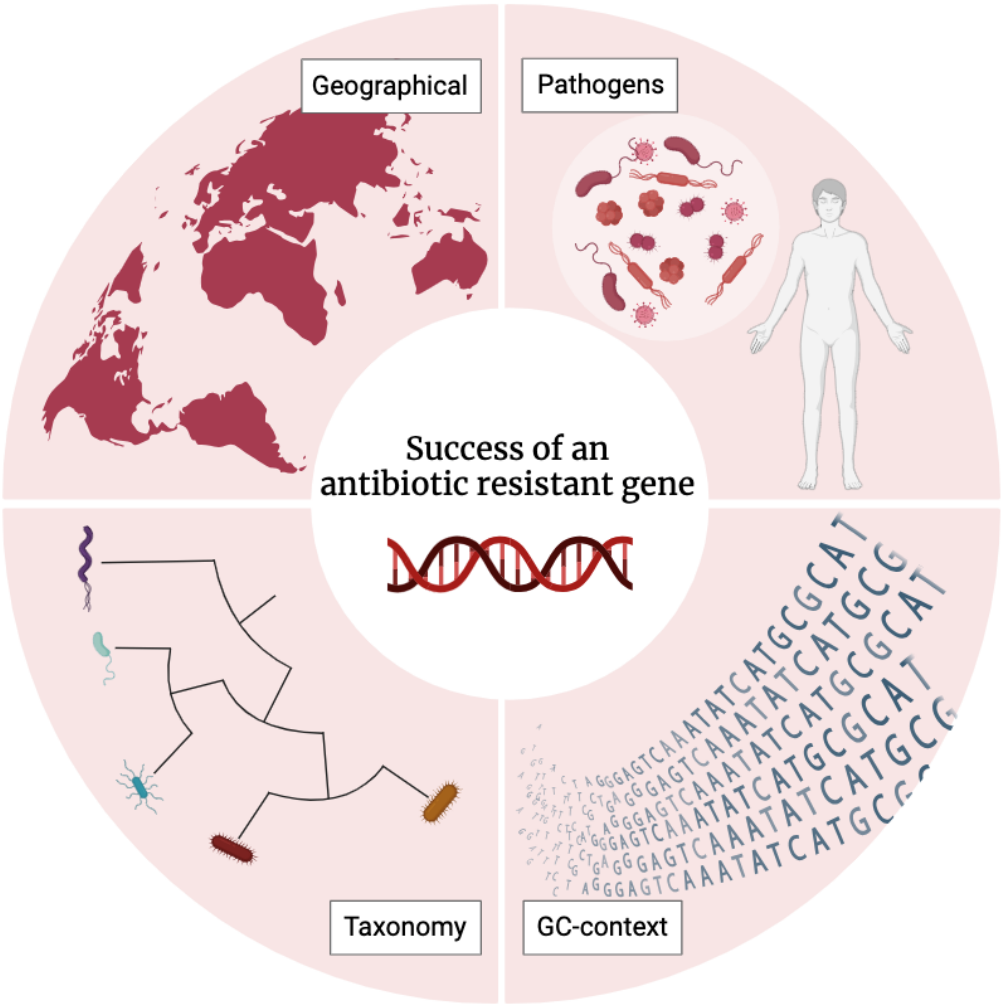


Figure 4.1: *The conceptual model of the success of an antibiotic resistant gene (ARG). The model consist of four different components: 1) dissemination in taxonomically different hosts, 2) dissemination in different host GC-contents, 3) geographical dissemination, and 4) dissemination in human pathogens. Illustrated using BioRender.*

the $N_{species,tot}$ is the total number of species used in the blast.

The GC-context component is built under the assumption that a successful ARG can be spread to hosts with a large variety of GC-contents. And as mentioned by Sánchez-Osuna *et al.* [5] in their study on GC-dissemination, for a rapidly disseminated genes, the ARG GC-content should be independent of the host genome GC-content. A gene found in multiple GC-contexts has also overcome constraints that can be imposed by the GC-content differences. The GC_S score, combines several statistics used to characterize the spectrum of GC-contents in which an ARG is found. The score is calculated by first min-max transforming the statistics, and then aggregating them with the following formula

$$GC_S = \sqrt{STD^2 + MAD^2 + (R_{max,min})^2 + IQR^2 + (R_{90,10})^2 + (p - value)^2} \quad (4.2)$$

where STD is the sample standard deviation and MAD is the median absolute deviation. $R_{max,min}$ describes the total range of GC-values, the IQR describes the inter quartile range of host GC-content, and $R_{90,10}$ describes the range between the 90^{th} and 10^{th} percentile. The $p - value$ is generated from a Wilcoxon ranksum test, examining if the distribution of the observations for the ARG comes from the same distribution as the distribution of GC-values of the full blast data. Note that we are looking for a non-significant result, as that would mean we cannot reject the hypothesis that the distributions are from the same source.

The conceptual model contains a component that describes the ability of a gene to spread between countries and continents. The score, G_S, is calculated as

$$G_S = N_{countries} \times N_{continents} \quad (4.3)$$

where $N_{countries}$ and $N_{continents}$
represents the number of countries and continents the ARG was found in, respectively.

The final component of the conceptual model is the presence in pathogenic bacteria. This component serves to include the direct risk for human health. The score, P_S, is the absolute number of pathogens, $N_{pathogen}$, an ARG has spread to.

$$P_S = N_{pathogen} \quad (4.4)$$

To ensure all components contributed equally to the aggregated score, the component scores were z-transformed using the formula

$$Z_i = \frac{S_i - \overline{S}}{s} \quad (4.5)$$

where Z_i is the new z transformed score, S_i is the score of sample i, $\overline{S}$ is the mean of all scores and s is the sample standard deviation. The aggregated score, S_{tot}, was then generated by taking the median value of the four separate scores.

$$S_{tot} = Med(T_S, GC_S, G_S, P_S) \quad (4.6)$$

This resulted in a final success-score, where each ARG is assigned a score between -0.547 (lowest score) to 7.35 (highest score). The top 100 genes, their success-scores and individual component scores are presented in table A.1 in Appendix.

The Pearson correlation between the final success-score, S_{tot}, and the individual scores, T_S, GC_S, G_S, and P_S, was examined and presented in figure 4.2. The result reveals that the pathogen score P_S is highly correlated (0.922) with S_{tot}, while the GC-context score GC_S has the lowest correlation to S_{tot} (0.531). Furthermore, the highest correlation in the individual scores appears between G_S and P_S at 0.702. While the lowest correlation, 0.195, occurs between GC_S and P_S.

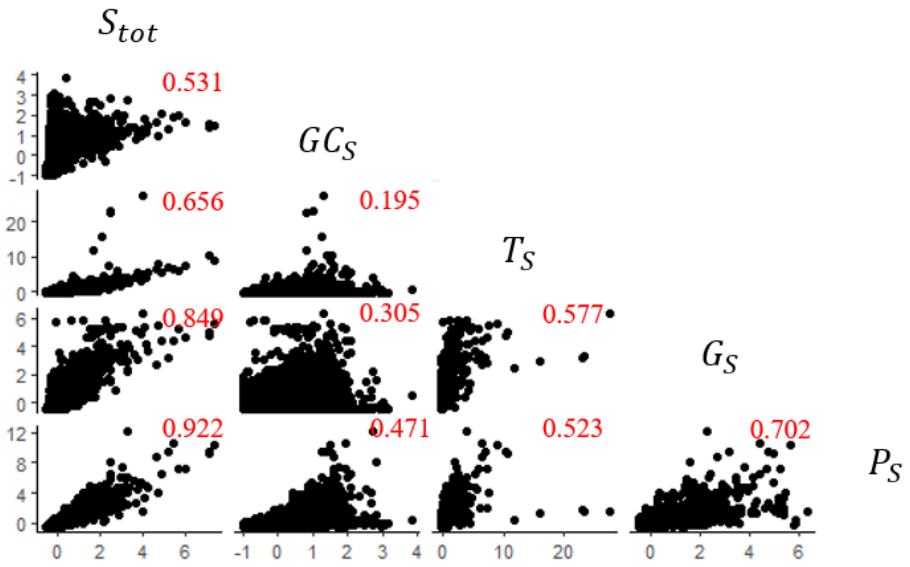


Figure 4.2: *The Pearson correlation (red numbers) between the total success-score, S_{tot}, and the individual scores T_S, GC_S, G_S, and P_S.*

4.2 The scores and most successful genes

The 100 highest scoring ARGs (S_{tot} ranging from 7.35 to 1.93) were categorised based on which drug class they confer resistance against, as specified in CARD, and presented in figure 4.3. ARGs classified as macrolides, lincosamides and streptogramins were assigned to a MLS class. Cephalosporin, cephamycin, carbapenem, and penam were assigned to β-lactams. ARGs that were associated with two or more drug groups were classified as multidrug. The largest ARG-classes were multidrug, and aminoglycosides with 25 and 23 genes respectively, followed by β-lactams with 9 genes and MLS containing 8 genes.

Aminoglycosides and MLS antibiotics target the ribosome in some way, while β-lactams target the cell wall synthesis. What can be noticed that in each category there are often multiple variants of the same type of gene. For example, there are multiple variants of AAC genes (aminoglycoside acetyltransferases) that inhibit aminoglycosides through acetylation, and they primarily differ by the position of the acetylation. There are also multiple APH genes (aminoglycoside phosphotransferases) that differ by the position of the phosphorylation which it uses to inactivate the aminoglycoside [46].

To set the results in context, the success-score was compared with previous sets of genes [47, 48, 49]. Zhang *et al.* [48] aims to identify ARGs that pose high risk for human health and uses three different criteria: enrichment in human-associated environments, gene mobility, and host pathogenicity. The high risk ARGs are those that meet all three criteria (highest risk), and criteria 1 and 2 (future threats). Similarly, Qian *et al.* [47] also focuses on the risk for human health. Four different scores measuring human accessibility, clinical availability, mobility and human pathogenic-

Multidrug		Aminoglycoside		Beta-lactams	MLS
AAC(6')-Ib-cr5	*mdtE*	*AAC(3)-Ib*	*acrD*	*CTX-M-15*	*ErmB*
acrB	*mdtF*	*AAC(3)-IId*	*ANT(2'')-Ia*	*mecA*	*ErmC*
AcrE	*mdtN*	*AAC(3)-IIe*	*ANT(3'')-IIa*	*OXA-1*	*mphE*
AcrF	*MexD*	*AAC(6')-Ib7*	*ANT(3'')-II-AAC(6')-IId*	*PC1-blaZ*	*msrE*
ceoB	*MexF*	*AAC(6')-Ie-APH(2'')-Ia*	*APH(3')-Ia*	*TEM-1*	*oleB*
cpxA	*MexI*	*aad(6)*	*APH(3'')-Ib*	*TEM-116*	*oleC*
CRP	*MuxB*	*aadA*	*APH(3')-IIa*	*TEM-208*	*srmB*
Ecol-acrA	*oqxB*	*aadA16*	*APH(3')-IIIa*	*TEM-245*	*tlrC*
Ecol-mdfA	*Paer-CpxR*	*aadA2*	*APH(3')-VIa*	*TEM-34*	
evgS	*smeB*	*aadA5*	*APH(6)-Id*		
H-NS	*smeE*	*aadA6*	*kdpE*		
marA	*TolC*	*aadA9*			
mdsB					

Tetracycline	Fluoroquinolone	Phenicol	Peptide	Diaminopyrimidine	Disinfecting agents and antiseptics
tet(A)	*emrA*	*catB3*	*bacA*	*dfrA1*	*qacE*
tet(B)	*emrB*	*catI*	*BRP(MBL)*	*dfrA12*	*qacL*
tet(C)	*emrR*	*cmlA5*	*eptA*	*dfrA14*	*qacEdelta1*
tet(L)	*mdtH*	*floR*	*YojI*		
tet(M)	*MdtK*	**Sulfonamide**	**Aminocoumarin**	**Nitroimidazol**	**Rifamycin**
tet(O)	*QnrS1*	*sul1*	*mdtB*	*Mrx*	*arr-3*
tet(W)		*sul2*	*mdtC*	*msbA*	

Figure 4.3: *The 100 highest scored ARGs assigned into drug-classes based on the classification in CARD. Macrolides, lincosamides and streptogramins were assigned to a MLS class. Cephalosporin, cephamycin, carbapenem, and penam were assigned to β-lactams. ARGs that were associated with two or more drug groups were classified as multidrug.*

ity, are combined into a risk index. The ARGs in the top 25% are considered high risk genes. A comparison was also made using a set of ARGs described by Zhuang *et al.* [49]. The authors counted the frequency of each ARG in 9374 PubMed publications between years 1990 and 2020 as a measure of relevance, and presented the top 50 most frequently mentioned ARGs.

The comparison was made by grouping the genes based on the risk or relevance level assigned in the three articles [47, 48, 49]. The success-scores of the genes in the groups was then plotted using box plots, as presented in figure 4.4. A Wilcoxon rank sum test was performed to test if the success-score differed significantly between the high risk/top reported ARGs and the remaining ARGs, which would indicate that the success-score can be a tool to identify relevant genes. The tests revealed a high level of significance with a $p - value < 2.22 \times 10^{-16}$ in all cases.

Worth noting is that the articles compared with our model look for ARGs associated with human risk, or the extent to which ARGs have been reported in scientific journals - whereas our scope is larger. This means we are not aiming to capture the same information, and it is not expected that the results should be identical. However, figure 4.4 suggest that the success-score, based on the conceptual model, can provide similar insight into which genes should be "relevant" or "high-risk" genes.

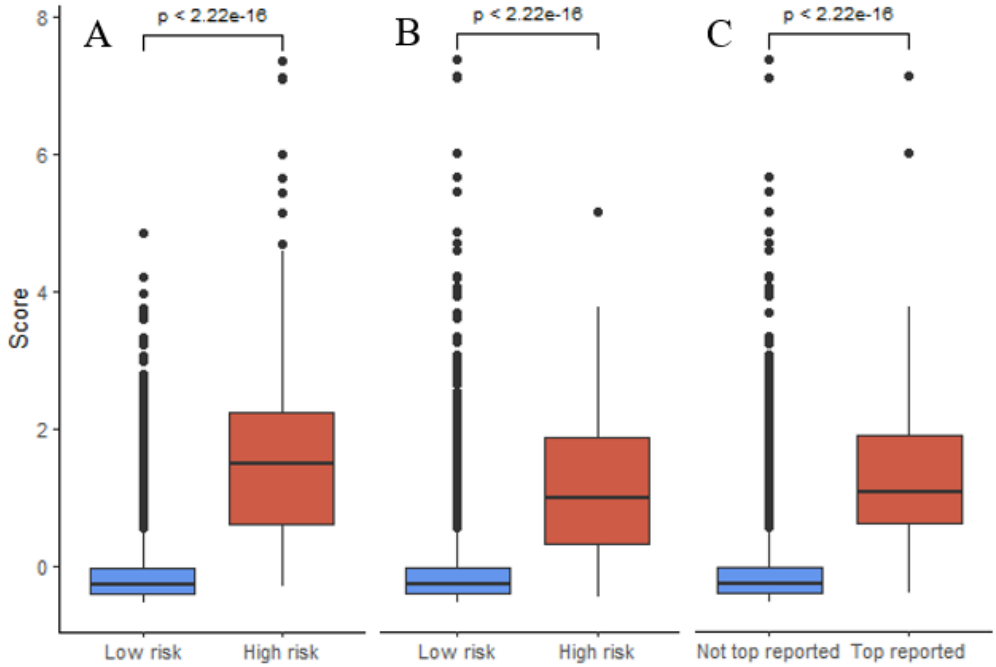


Figure 4.4: *The success-score was compared with ARGs reported in previous studies. Here A) displays the score compared with "high-risk" genes in Qian et al. [47], while B) compares "high risk" ARGs by Zhang et al. [48] with the success-score. C) A third comparison was made with ARGs presented as top reported in 9374 examined PubMed publications between years 1990 and 2020 by Zhuang et al. [49]. A Wilcoxon rank sum test was performed and reveals a high level of significance* $(p - value < 2.22 \times 10^{-16})$ *in all cases.*

4.3 Predictive model

In this section of the report, the results of the Random forest predictive model are presented. The model has been developed to investigate if it is possible to predict if an ARG is successful, per the definition of the conceptual model. The amount of information that is needed to predict with an acceptable degree of sensitivity and specificity is also examined.

The success-score (S_{tot}) from the conceptual model forms the basis of the binary classes used for the predictive model. The 500 genes with the highest success-scores are classified as "1" (highly successful) and the rest as "0". Figure 4.5 shows the distribution of the (S_{tot}) score, as well as the cut-off at 500.

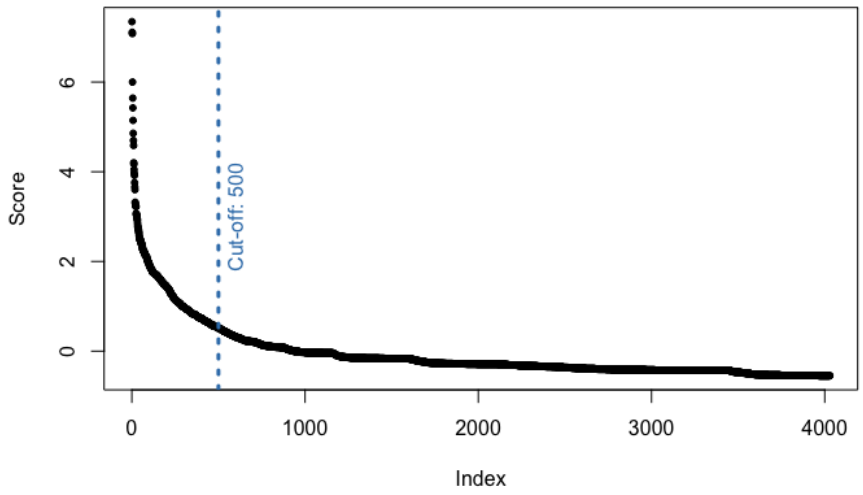


Figure 4.5: *Success-scores, S_{tot}, of the data set, and the cut-off for the binary classification.*

4.3.1 Model including all available features

The first model that was evaluated was based upon all available features, see table 4.1. The model was separately trained and tested on eight data sets, that were based on different amounts of observations of each gene. The eight data sets are therefore referred to by the number of observations used for feature generation. Note that for the one observation data set, the only available features were the gene information features and the number of pathogens.

Table 4.1: *A collection of all features used to describe the ARGs in the Random forest model, divided per feature category. Note that the codon usage feature consists of one feature per codon, in total 64 features.*

Gene features	**Geographic features**	**GC features**
Gene length	Number of countries	IQR
Gene GC-content	Number of continents	$R_{90,10}$
Codon usage		$R_{max,min}$
		Mean GC-content
Taxonomic features	**Pathogen features**	STD
Taxonomic distance	Number of pathogens	MAD
Number of species		

The ROC curves for all data sets, visualizing the performance of the model in terms of sensitivity and specificity at all probability cut-off points, are shown in figure 4.6A. The model was tested in a 5-fold cross validation, and the AUROCs of the folds are shown in figure 4.6B.

The model that is trained with the data set containing all available observations has the highest overall area under curve of 0.995. It also has the curve that most closely resembles the ideal ROC curve visualised in figure 2.4. This result is expected since the full data set is used to build the conceptual model. With the same motivation, it was expected that the one observation data set, should have the overall lowest AUROC. However, that is not the case, figure 4.6B shows that the data set with five observations has the lowest area under the curve of 0.88. Potentially, this could be due to the information in the five observation data set contained potentially conflicting information or outliers, and with few observations each data point becomes important.

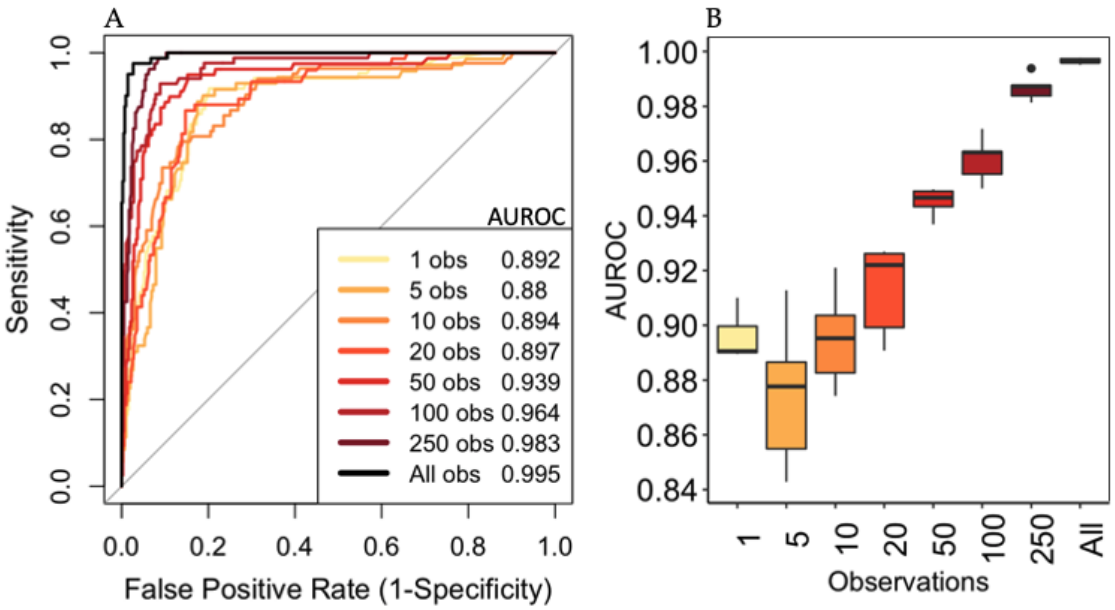


Figure 4.6: *a) A ROC plot showing the performance of the model trained on all available features. There are 8 curves each representing a different data set. The legend specifies the AUROC value of each curve. b) A box plot representing the AUROCs generated in a 5-fold cross-validation of each data set.*

The performance of the model was also measured as the sensitivity at 95% and 90% specificity, corresponding to a false positive rate of 5% and 10% respectively. Due to the large amount of antibiotic genes that are less successful, it will be critical to have a high specificity to limit false positives. Setting of false alarms for even 5% or 10% of the cases could, due to the class imbalance, result in there being more false positives than true positives.

The average sensitivity over five-folds is shown in figure 4.7. Except for a small decline at five observations, the trend at 90% specificity is that the sensitivity increases with the number of observations. At 95% specificity, there is little improvement between one and 20 observations, while there is a gap at 50 observations. At 95% specificity, ≥ 50 observations are required to get a sensitivity of 70%, while at 90% specificity, even 1 observation can be enough to get a 70% sensitivity. This shows

that there is a clear trade-off between sensitivity and specificity, and that this trade-off becomes more apparent the fewer observations there are.

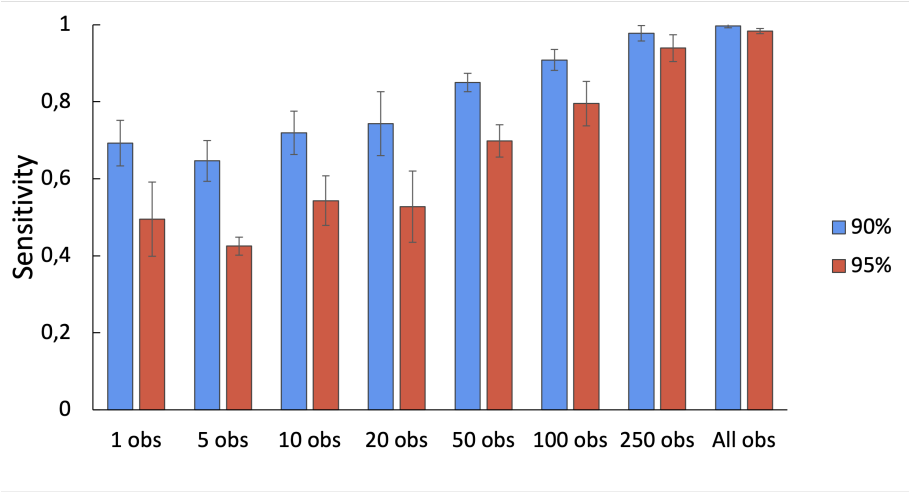


Figure 4.7: *A bar plot showing the average sensitivity over 5-folds of each data set at 90% and 95% specificity. The error bars represent the standard deviation of the sensitivity. The figure is based on the Random forest model trained on all available features.*

For two selected data sets, 5 observations and 100 observations, the confusion matrices at 95% specificity are shown in figure 4.8. The large class imbalance is visible as the number of reference "0" are much higher than the number of reference "1". The impact of the class imbalance is also clear as for the 100 observation case, there are 19 false positives, and 66 true positives, this means of all positive tests, 22% will be classified wrong. For the five observation case, 35% of the positive tests are wrong.

5 observations

		Reference	
		0	1
Prediction	0	288	43
	1	15	28

100 observations

		Reference	
		0	1
Prediction	0	375	18
	1	19	66

Figure 4.8: *Confusion matrices for the 5 and 100 observations data sets, at 95% specificity. The probability cutoff thresholds required for 95% specificity were 0.642 and 0.607 respectively. The matrices are based on the Random forest model trained on all available features.*

4.3.2 Feature importance

Based on the models plotted in figure 4.6, the importance of each feature, measured as MDA, was extracted. Figure 4.9 visualizes the feature importance for 1, 5, 50 and

250 observations. Note that the geographic features and number of species are not included in the model for the 1 observation data set, as those numbers are always 1. Similarly, the GC features and the taxonomic distance are not included since they are based on there being at least 2 species, which is not possible as the set is based on 1 observation.

In the figure there are two key things to notice, the first being that codon usage, gene length and gene GC-content is more important in the data sets with fewer observations. When there are more observations, features such as the number of countries and number of species become more important. The second thing to notice is that the more observations, the higher is the MDA of the most important feature. The linear relationship between success-score and gene length and GC-content was examined and revealed a correlation of 0.144 and -0.009 respectively.

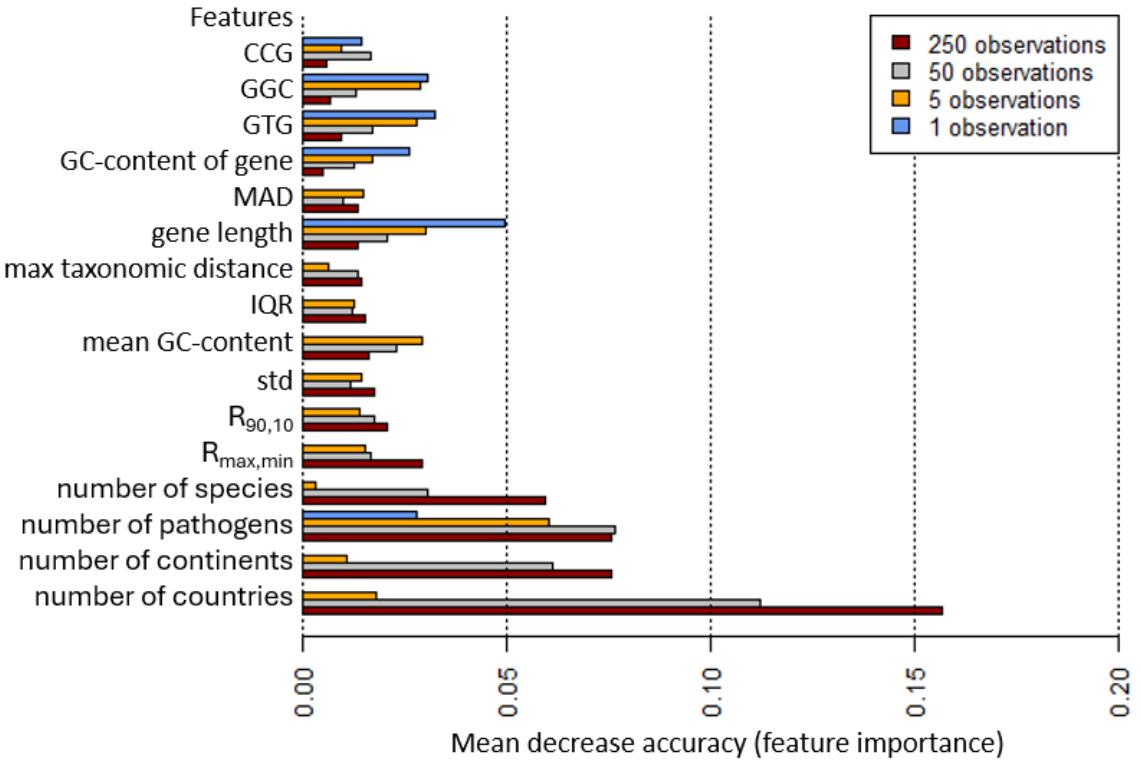


Figure 4.9: *Mean Decrease Accuracy for the 1, 5, 50 and 250 observation data sets. Only three of the 64 codons are included in the figure, these are the codons that are the most important in terms of MDA.*

To see if the model could be improved beyond the initial performance from figure 4.6, only features with a MDA above 1.5% for each set of observations was selected for model training (see figure A.1 available in Appendix) Having a model that can perform on fewer features is beneficial in a real world scenario since less data collection and handling would be necessary. Figure 4.10 shows the ROC curves for the model. Once again the trend is that the more observations, the better is the overall model performance. In terms of average AUROC values, the differences are minor.

This model is slightly better for the 1-50 observations data sets, slightly worse for the 100 observations data set, and equal for the 250 and all observations data sets.

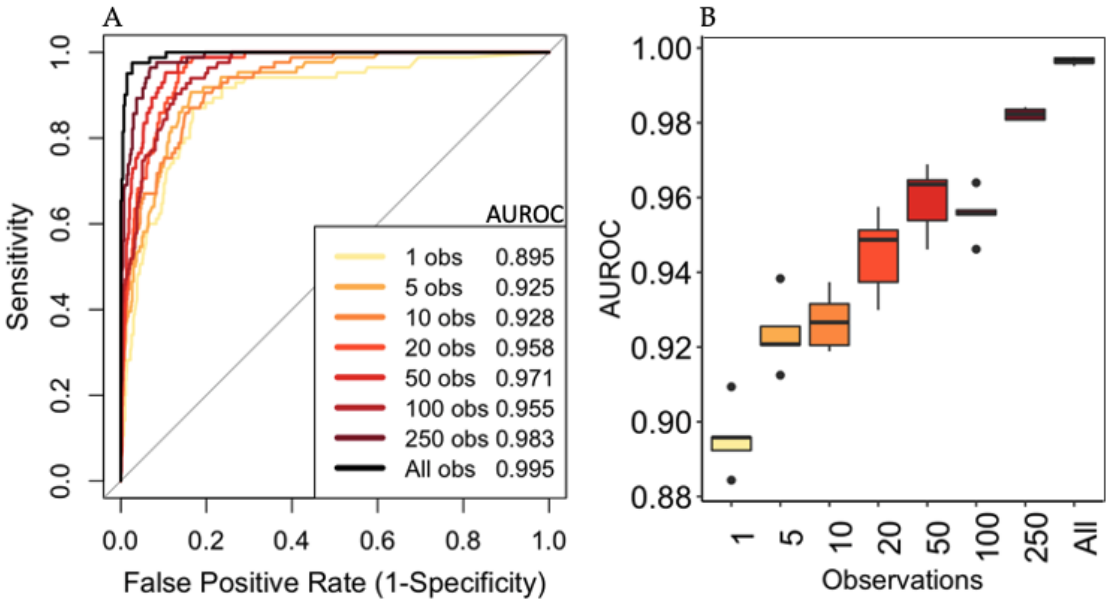


Figure 4.10: *a) A ROC plot showing the performance of the model trained on the features with MDA* $\geq 1.5\%$*. There are 8 curves each representing a different data set. The legend specifies the AUROC value of each curve. b) A box plot representing the AUROCs generated in a 5-fold cross-validation of each data set.*

The sensitivity at 90% and 95% specificity has also been calculated for this model, shown in figure 4.11. Similar to in the model with all features (see figure 4.7) the performance is better at 90% specificity across all sets of observations. In this model, the sensitivity improves gradually, in contrast to the model with all features, where there was little improvement between one and 20 observations. What is interesting to notice, and adds to the previously mentioned AUROC comparison, is that for 100 and 250 observations, the performance is worse at both levels of specificity in this model. Whereas for 1-50 observations, the performance improves at both levels of specificity.

The confusion matrices at 95% specificity, for the model including only the important features, are shown in figure 4.12. At this specificity, 41% of the positive predictions are false positives in the 5 observations data set, and 25% in the 100 observation data set. Compared to the confusion matrices in figure 4.8 these numbers are 6% points, and 3% points higher, respectively. The improvement in sensitivity suggests that reducing the number of features is beneficial for the low observation data sets (1-20 obs). This could be because it reduces sources of conflict and potential outliers, which are very apparent when only few data points are used. On the other hand, it appears to be beneficial to use all features when more observations are available.

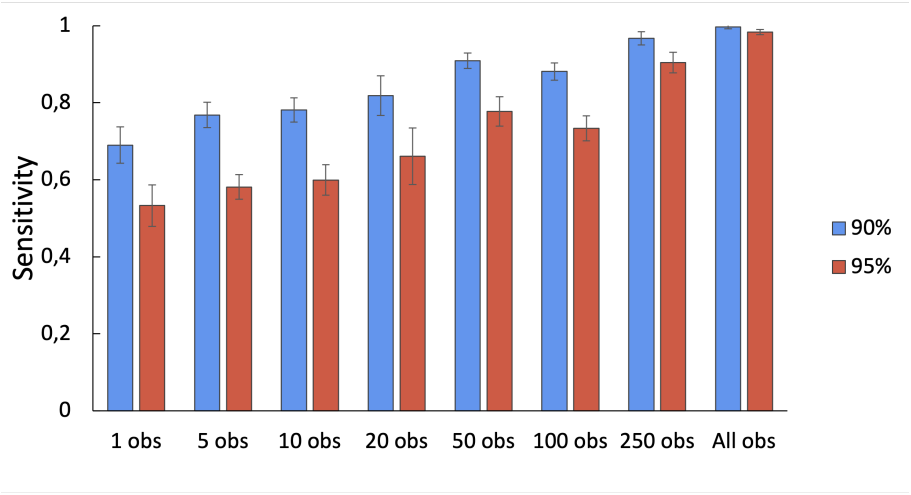


Figure 4.11: *A bar plot showing the average sensitivity over 5-folds of each data set at 90% and 95% specificity. The error bars represent the standard deviation of the sensitivity. The figure is based on the Random forest model trained on features with MDA* $\geq 1.5\%$.

5 observations

		Reference	
		0	1
Prediction	0	649	37
	1	34	49

100 observations

		Reference	
		0	1
Prediction	0	379	25
	1	19	58

Figure 4.12: *Confusion matrices for the 5 and 100 observations data sets, at 95% specificity. The probability cutoff thresholds required for 95% specificity were 0.579 and 0.689 respectively. The matrices are based on the Random forest model trained on features with MDA* $\geq 1.5\%$.

4.3.3 Replicates of 5 observations

From previous discussions, the data is likely prone to outliers and conflicting information, which especially affects the data sets based on fewer observations. Therefore, the model was built and tested on six separate selections of five observations. The sensitivity at 95% and 90% specificity are presented in figure 4.13. Throughout all six replicates there is only minor variation which gives the model some credibility in that it can perform consistently.

In regards to the sensitivity, all replicates perform better at 90% specificity than at 95%. At 95% the model is just slightly better than a random guess and would not be useful in a real life context. At 90% the model performs better, but a 10% false discovery rate will still be an issue, especially due to the large class imbalance that will exist since much more genes are unsuccessful.

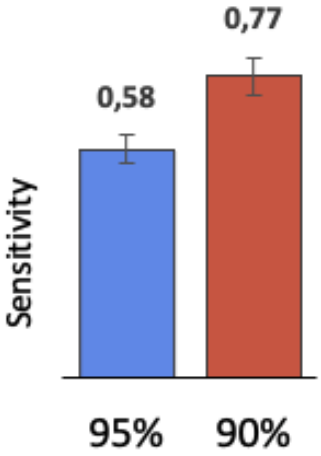


Figure 4.13: *A bar plot showing the average sensitivity at 90% and 95% specificity for six different sets of 5 observations. The error bars represent the standard deviation. The figure is based on the Random forest model trained on features with MDA* $\geq 1.5\%$.

4.4 Managerial implications - resistance and cost analysis

This section contains the results of the two key meta-analyses. The first analysis relates to the resistance in the case countries Italy and Sweden, and the second analysis relates to costs.

4.4.1 Resistance analysis

To analyze the spread of antibiotics resistance in Sweden and Italy, data from the European Centre for Disease Prevention and Control (visualized in the surveillance atlas of infectious disease, ATLAS) is used [50]. Figure 4.14 shows the resistance levels of various pathogens against two common β-lactam antibiotics. From the figure it can be concluded that resistance against these two antibiotics is more widespread in Italy than in Sweden. This pattern is also true when further antibiotics and species are analysed [50]. From the figure it can be noted that the carbapenem resistance differs between the countries. In Italy 87% of *Acinetobacter baumannii* isolates are resistant to carbapenem antibiotics, as well as 27% of *Klebsiella pneumoniae* and 16% of *Pseudomonas aeruginosa* isolates. In Sweden, only *Pseudomonas aeruginosa* is affected significantly, with 12% resistant isolates.

Another notable difference is the difference in Staphylococcus aureus resistance to meticillin. The resistance, commonly known as meticillin resistant *Staphylococcus aureus*, MRSA, is a global health issue and one of the most commonly acquired infections in hospital environments [51]. As visible in figure 4.14, the prevalence of MRSA in Italy is 30%, compared to only 2% in Sweden. The 2% corresponds to approximately 3293 cases of MRSA in Sweden [52], if the resistance levels were at

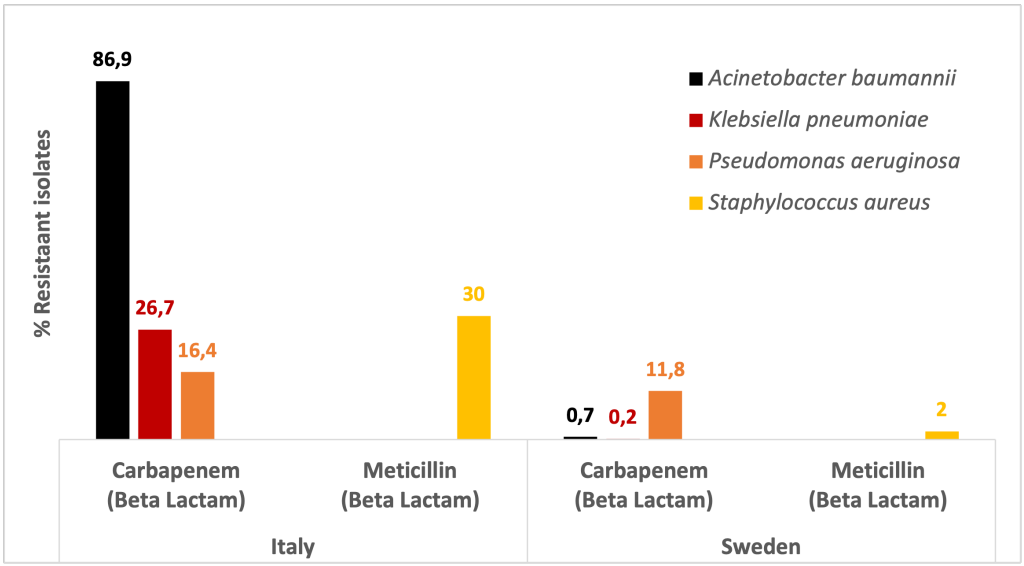


Figure 4.14: *Percentage of resistant isolates for two different beta-lactam antibiotics, for four pathogens in Italy and Sweden in 2021. Based on data from European Centre for Disease Prevention and Control (ATLAS)[50].*

the Italian levels, 30%, it would roughly correspond to an incidence rate of 50 000 cases of MRSA.

Figure 4.15 shows the use of antibiotics in Italy and Sweden between 2010 and 2021 based on data from European Center for Disease Control (ECDC) 2022 [53]. The figure reveals that while consumption is decreasing in both countries, the countries have a difference in prescription rates, with Sweden having 8-10 DDDs lower than Italy. Italy still has prescription levels above Swedish levels in 2010. A detailed view of the proportion of each class of antibiotic in the countries is presented in figure 4.16. In the figure, the Netherlands and Greece are also included to extend the comparison between countries with low and high antibiotics usage. The Netherlands has a low overall antibiotics usage, and Greece has a high usage.

The first thing to note is that the proportional use of the different antibiotic classes differs between the countries. This could be explained by the fact that different antibiotics are used for different types of infections, and that the distribution of infections is different in the countries. Another explanation is that doctors prescribe different antibiotics for the same type of infection, driven by either preference and routines or that there are different resistance patterns in the countries.

Sweden and the Netherlands have similar patterns on all classes except Beta-lactam antibiotics (penicillins) and Macrolides, Lincosamides and Streptogramins (MLS). When comparing Sweden and Italy it is apparent that they share the most common antibiotic, but, for example, the second most common antibiotic class in Sweden, Tetracyclines, is less than 5% of the total consumption in Italy.

A survey from the European Union about knowledge and attitudes towards An-

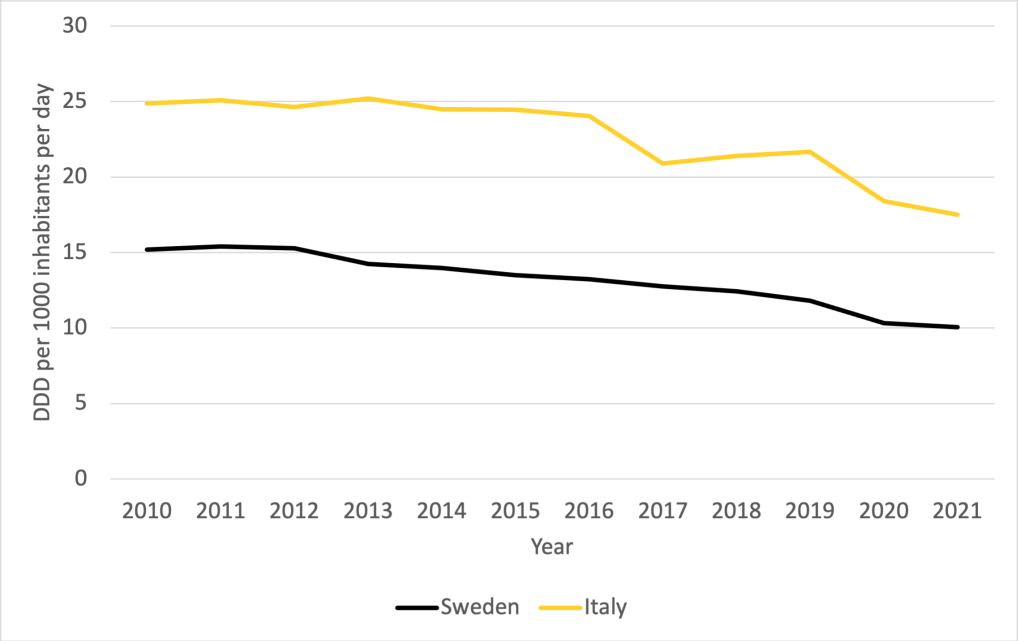


Figure 4.15: *Antibiotic usage in primary care and hospitals over time, in Sweden and Italy. The consumption is given in defined daily doses (DDD) of antibacterials per 1000 inhabitants. Data from ECDC 2022 [53].*

timicrobial Resistance among the public was conducted in 2022 [54]. The result shows a low general awareness among the Italian population, in contrast to a high general knowledge among the population of Sweden. The survey also suggests that the information about correct use of antibiotics is better in Sweden compared to Italy. Moreover, 32% of the Swedish contestants answered that they had been given information about unnecessary use of antibiotics compared to 21% of the Italian contestants [54]

4.4.2 Cost analysis

In a report from 2017, the World Bank group estimated that antimicrobial resistance would lead to a reduction in world GDP by 1.1% - 3.8% [55]. The WBG also predicts that it will cause more economic harm to low-income countries. Based on this high level cost context, we will zoom into the direct costs of antibiotics resistance for a hospital. As a reference point, in 2013, the Swedish public health agency estimated that additional healthcare costs from antibiotics resistance was 160M SEK, based on 4590 clinical infections and 3096 carriers. The cost was based on the cost of inpatient care, outpatient care, primary care and contact tracing [56]

A review by Poudel (2023) gathered information on parameters relating to the burden of antibiotic resistance to hospitals [57]. The review focuses on articles that compare patients with a resistant and non-resistant form of the same infection, to account for the regular treatment costs, and only study the marginal costs. Poudel (2023) reports that the mean excess length of stay is 7.4 days (95% CI: 3.4 - 11.4), with a range of excess cost per patient of $ -2371 to 29289 [57]. The review is per-

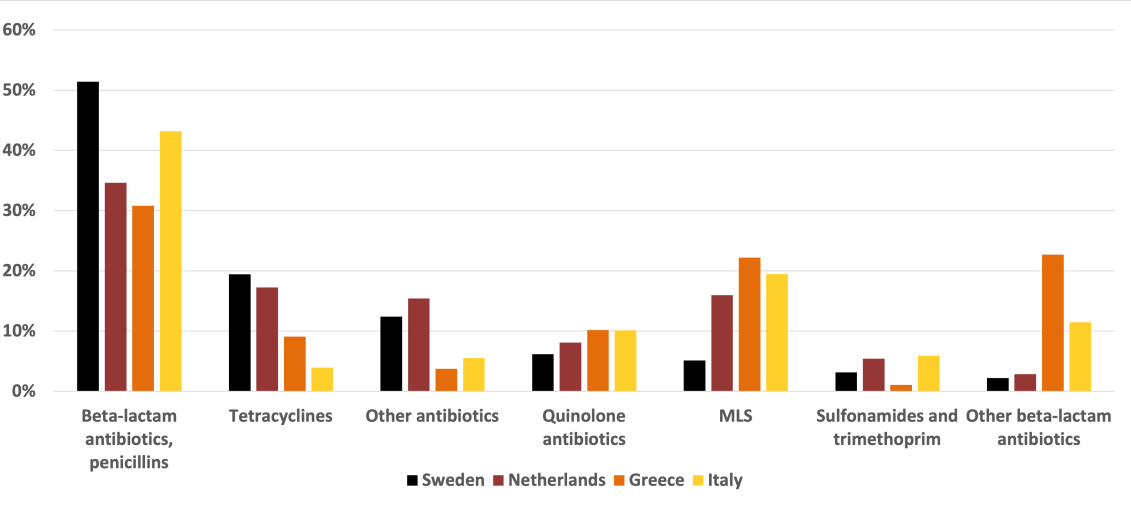


Figure 4.16: *Proportion of each antibiotic used per country, based on data from ECDC [53].*

formed across a range of high and upper-medium income countries, and a variety of infection types explaining the large per patient cost range.

One study mentioned in the review by Poudel (2023) [57], is an article by Thorpe (2018) [58]. In the study that analyzed survey data in the US, the per-infection incremental cost for resistant infections was calculated, concluding that the cost on average increased by $ 1383. Furthermore, Thorpe (2018) [58] mentioned that 61% of all resistant infections are relatively uncomplicated urinary tract infections (UTI), and were to be removed from the data set, the incremental cost increases to US $ 2656. This clearly shows that the cost is driven by the type of infection. An important note is that the costs are for the payer, which in the US is the patient, or the insurance company, and the cost does not necessarily reflect the actual cost distribution of the hospital, but rather how the hospital chooses to price their services.

It is difficult to directly compare costs due to the difference in payment models between countries. As an example, Thorpe (2018) [58] studies the payers' costs, which in the US is the patient, or the insurance company, and the cost does not necessarily reflect the actual cost distribution of the hospital, but rather how the hospital chooses to price their services. On the other hand, it is likely the cost proportions and buckets are similar. The review by Poudel [57] contains seven studies that have broken down the cost by cost item [59, 60, 61, 62, 62, 63, 58]. The breakdowns are presented in figure 4.17.

A meta analysis was performed with the aim of identifying what items are driving the cost of antibiotic resistance. To do so, the cost buckets were coded by category to be able to identify commonly reported cost items. As seen in figure 4.18, the coding resulted in five key buckets. The largest bucket is the cost of medication and treatment which involves both the antibiotic as well as other medications

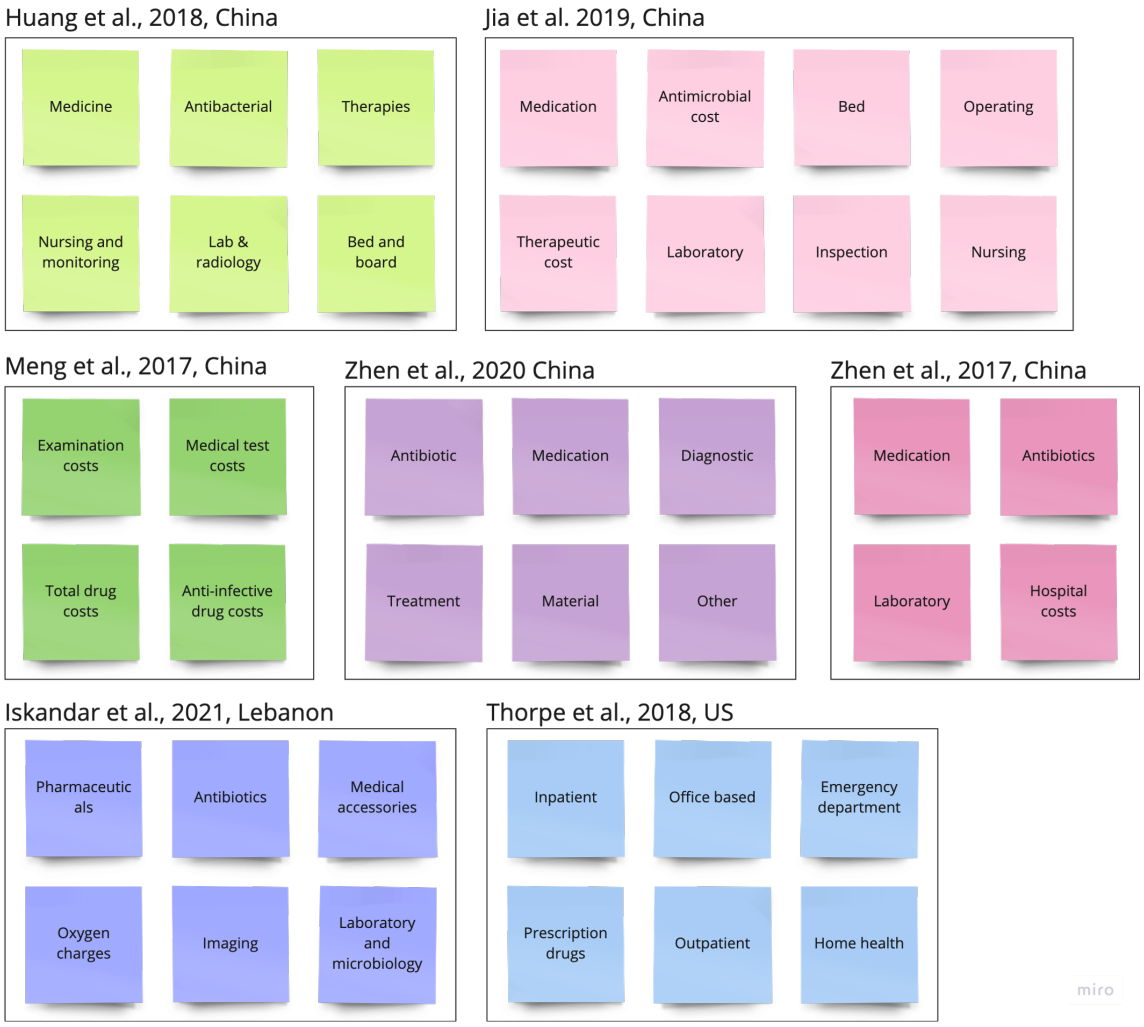


Figure 4.17: *Cost buckets that are impacted by antibiotics resistance, as outlined in seven studies, color coded by article.*

and non-pharmaceutical therapies. Huang (2018) mentions that the non-antibiotic medication is a significant cost bucket because the resistant infection impacts co-morbidities negatively, so that a patient's "regular" medicinal costs also increase [59]. Another important bucket is the cost of diagnostics and laboratory work. This is likely because a patient with an antibiotic resistant infection is likely to requires additional tests which will drive additional costs. The third bucket is the cost of nursing and care. This is related to the length of stay and how much care the patient must have. Finally, there are other hospital operating costs, as well as additional costs for the outpatient and home health care.

Figure 4.18: *Categorising the key cost buckets identified in the articles. The items are color-coded based on the article the item was found in (see figure 4.17).*

5
Discussion

This chapter intends to interpret the results presented in the previous chapter. A discussion is held about whether the genes that have been found as most successful are reasonable and how they differ from the literature. A review of the conceptual model is done in a thorough manner where the improvement potential of the components are discussed as well as further suggestions additional components. The results of the predictive model are also discussed, specifically how the performance is affected by the number of observations, the trade off between sensitivity and specificity and the impact of the number of features.

The discussion section is concluded by a reflection on the managerial and policy implications that follows with the rise of antibiotics resistance. To tie in the managerial implications with the remainder of the thesis, one focus area will be on how machine learning models could be implemented to assist in reducing the impact of antibiotic resistance on healthcare.

5.1 The success-score and top rated genes

The top 100 scored ARGs were grouped by the antibiotic class they confer resistance against (see figure 4.3). The majority of the genes were classed as multidrug, followed by aminoglycoside, β-lactams, and MLS. In the multi-drug class all except *AAC(6')-Ib-cr5* are coding for efflux pumps or sub-units of efflux pumps. As previously mentioned, the efflux pumps are efficient for the bacteria since they alone can confer multi-drug resistance [27]. It is likely, from this point of view, that a successful ARG can confer resistance to many different antibiotics and give the bacteria an evolutionary advantage over other, single drug resistance bacteria.

Apart from the multidrug class, figure 4.3 reveals a variety of other drug classes that genes with a high success-score confers resistance to, all with different mechanisms of action. It is difficult to say whether the results are plausible only based on the drug target. It is expected that the number of resistance genes against a specific drug class would be influenced by the historic usage of the class, due to the selection pressure on bacteria that results from antibiotic exposure. In figure 4.16 it is shown that β-lactams are a prominent class of antibiotics, and in the top 100 table, there are also many β-lactam resistant genes. On the other hand, the aminoglycosides do not appear to be frequently prescribed today (see figure 4.16), however before 1980 they were used as first-line drugs [64]. This historical use of aminoglycosides can

possibly explain the high number of resistance genes on our top 100.

When comparing the list in figure 4.3 with other findings agreement is seen to some extent. For instance Zhuang *et al.* [49] examined 16 ARG families from the 46 chosen articles, and presented tetracycline as far most reported group of ARGS followed by multidrug, beta-lactams, aminoglycosides and sulfonamide . In Zhang *et al.* [48] on the other hand, the β-lactams were the most common followed by the aminoglycosides, MLS and diaminopyimidines . It is reasonable that the most common antibiotic families will vary between the papers, especially since the scope of the articles is different and the authors intend to capture different information. For future research it would be interesting to study the mechanism of action and identify those that are represented among the highly successful genes.

Several of the top 100 genes (see figure 4.3) are commonly reported in other research. For instance, Zhuang *et al.* [49] manually examined 46 representative papers from PubMed and gathered the most frequently mentioned ARGs, per environmental isolation source, in China, Europe, USA, Africa, and Australia. It is clear that from our top 100 list, *ermB, ermC, tet(A), tet(B), tet(W), tet(O), tet(M), sul2*, and *tolC* were some of the most frequently reported ARGs in human fecal samples according to the authors, where *ermB* was reported in all 5 geographical areas. Additionally, the article presented four of the most reported ARGs in hospitals as *mecA, CTX-M-15* and *TEM-1*. These genes were also found in our top 100.

ARGs are commonly found in livestock farms as a consequence of using antibiotics to treat and prevent disease among animals. On the same theme, articles studying ARGs in samples from farm waste water and compost, pigs and chicken feces, and fish tissue have been reviewed by Zhuang *et al.* [49]. Some ARGs presented, also present in our top 100, were *Aph(6')-ld, TEM-1, acrB, ermB, mdtH, mexD, mexF, sul2, tet(A), tet(B), tet(C), tet(L), tet(O),* and *tet(W). Sul2* was found in farms in all geographical areas. Moreover, Wang *et al.* [65] studied the presence of ARGs in human, chicken and pig feces and found high levels of tetracyclines, aminoglycosides and β-lactams. The authors report *tet(Q), tet(W), tet(M),* and *tet(O)* as the most widely detected tetracyclines in all samples studied. Of these, only *tet(W)* is not present in the top 100 success-score. Furthermore, high levels of the MLS ARGs *ErmB, ErmC, ErmF, ErmT* was reported, as well as the tetracycline *tet(L)* and the aminoglycosides *APH(3')-IIIa*, and *AAC(6')-Ie-APH(2')-Ia*. Of these, both of the aminoglycocides, the tetracycline and the MLS ARGs *ErmB*, and *ErmC* received a success-score in the top 100. Similarly, Ma *et al.* [66] studied the resistome from human, pig, and chicken fecal samples. Various tetracyclines were found, where *tet(A), tet(M)*, and *tet(W)* were all present in our top 100 list.

A part from livestock and hospital surroundings, ARGs are often found in the environment and can from there be spread to human pathogens [67]. Berglund [68] present *sul1, dfrA1, dfrA12, qnrS, sul2, ErmB, ErmC*, and a majority of the *tet*-genes as commonly encountered environmental genes that are easily spread among bacteria through HGT. It is therefore reasonable that these genes are also found to

a large extent in our top 100.

To further provide contextual relevance to the score it was combined with ARGs identified as high risk to human health in two articles by Zhang *et al.* [48] and Qian *et al.* [47]. The result revealed a high level of significance in identifying those 'high-risk' ARGs, and is aticipated given that the success-score contains the pathogen dissemination aspect, which was the primary focus in both articles. Nevertheless, as already mentioned, our model offers a wider perspective that is not explored in the other two studies, and captures additional genes not identified by the authors Qian and Zhang. Along with this, the score also manages to capture the most frequently reported genes from PubMed publications highlighted by Zhuang *et al.* [49]. The alignment between our top 100 list, and some of the most frequently mentioned genes [49], as well as from Wang *et al* [65], Ma *et al.* [66], and Berglund [68] gives credibility to our model and shows its ability to capture ARGs that are relevant in research and in the real world. An important note is that the cutoff for a successful ARG in the predictive model is top 500 and not top 100 as in figure 4.3. Hence, some of the genes that were found in the literature but not part of the list in figure 4.3 was still considered successful in the predictive model. This includes for example *vanA*, and *tet(Q)*.

5.2 The conceptual model and its limitations

The conceptual model was designed to capture a wide range of information about a gene, and the components are drastically differ in what they measure. This is beneficial as it allows us to redefine what a successful antibiotic resistant gene is, and find genes that are not typically researched extensively. There is no definition of a "successful ARG", but articles such as Zhang *et al.* [48] and Qian *et al.* [47] often focus on the human health risk as the most important component. Our conceptual model on the other hand serves to extend the definition, and look beyond the genes most commonly discussed, to help find genes that are typically overlooked or that could be dangerous in the future. For instance, a gene disseminated across large taxonomic distance and to numerous countries, but not yet found in pathogens, is likely, in just a matter of time before it is found in a pathogen.

Despite being verified with previous research, there is potential for improvements of the conceptual model. The taxonomic dissemination score was for instance based on the conclusion that a gene is spread more easily among closely related bacteria. NCBI taxonomy, used to determine the phylogenetic relationships, is an easy and fast way of determine the relationship between bacterial species. However, the classification based on taxonomy is sometimes not accurate since the names of species and their taxonomic lineage is dynamic. With new discoveries scientist sometimes move species between nodes in the tree, and when new species are discovered, more nodes are added. This have in some cases lead to publications of genomes not being accurate in their taxonomic lineage or species name [40]. Furthermore, the classification in NCBI taxonomy is based on the ranks of traditional hierarchical

classification proposed by Linnaeus (kingdom, phylum, class, order, family, genus, species), as a consequence this approach the phylogenetic relationships is simplified and some of its complexity is lost [40].

Another way of measuring distance of bacteria is to do so call amplicon sequencing of the 16S rRNA gene that encodes a part of the small ribosomal subunit. The gene is hypervariable between genomes of different organisms and is flanked by highly conserved regions making it a perfect target for amplification. The sequence divergence could then determine the distance among bacteria [69]. This method could possibly remove the bias of having genomes that are labeled wrong, or when traditional hierarchical ranks fails to capture the true phylogenetic relationship. This approach should be considered in future research.

The GC-context score aimed to capture the genes ability to spread over a variety of different genome contexts. Based on the findings by Sanchez-Osuna *et al.* [5], as mentioned in the theory chapter, we assume that the GC-content range of a gene (not a gene group as in the article) could be used as a way to measure dissemination, working based on the idea generated by the authors that for rapidly disseminated genes, the ARG GC-content should be independent of the host genome GC-content. This method has some limitations in our application. It can be argued that the GC-content is a very crude measure of phylogeny. In a way, the GC measure is therefore similar to the taxonomic component of the conceptual model because both components intend to capture similar information. On the other hand, the taxonomic measure is based on dynamic categories that may change over time, so it can be beneficial to include a static measure, like GC-content, as well.

For the geographical component it was assumed that ARGs that were spread to numerous countries and continents were successful in dissemination. An important component of this, that was not taken into account, is that the the size of countries and the ease of spreading between different countries will differ. One alternative approach would be to include the degree of dissemination within a country, for example the proportion of resistant isolates. However, the problem with skewed metadata arises here, as some countries have a higher frequency of genome sequencing than others. Another angle that was not included in the geographic score was the prevalence of ARGs in marine habitats, as all samples from oceans were excluded in the model. It is not only important to look in human related environments when identifying ARGs of risk for humans, but also in environmental bacteria. Research indicates that ARGs in environmental bacteria is the major source of ARGs observed in pathogens in clinical environments [67]. Also, since many aquatic ARGs are located on genetic integrons and plasmids they are at a high risk of being transferred to other species [68].

The pathogen score is a measure of how many pathogens a gene is found in. The purpose of including this component is to capture an element of clinical human risk. One drawback to the way the component is formulated today is the fact that not all pathogens are equally abundant, or equally pathogenic, and to immuno-

compromised people, even non-pathogenic bacteria can be dangerous. A possible improvement to the score is to weigh each pathogen by the previously mentioned factors, as a way to incorporate more information into the score.

In addition to the discussed improvements, there are alternative factors that should be considered in future research. One such important factor is the time dimension. A gene spreading rapidly among various bacterial species is without doubt more successful than a gene that achieves the same spread over a longer period. A reason that this factor was not included in the score was the challenge of assemble comprehensive metadata capable of adequately addressing it. In addition to the time factor, future work should aim to add a perspective of overlapping environments. Common environments has been shown to significantly increase the possibility for an HGT to take place [12], and a successful gene should therefor be addressed as one that has the possibility to be spread over these environmental barriers. However, once again the low availability of meta data relating to isolation source in the current databases hinders this type of analysis.

Along with the numerous components of the conceptual model and possible approaches to modify them, there was always risk to include such different components, as scores could cancel each other out. The method chosen in this thesis was aggregation via the median, meaning that a gene that scored high in one component but low in others, received a low overall score. Alternatively, weights can be assigned to components with high scores, to ensure that successful genes in one component get a high overall score independent of the other components. It is important to match the aggregation method with the intended use case of the model, and consider its implications on the overall classification of genes.

One major source of error in this report is the uneven data. The available genomes are heavily dominated by a few species, and the country of origin of the genomes is not accurately representing global antibiotic resistance dissemination. There are several available methods to perform normalization and reduce the bias in the data. However, we deemed that most methods would drastically reduce the available data. Instead, we decided to keep thee analysis on a species level, instead of analysing individual alignments. With this in mind however, it is likely data availability also skewed the result. For future analysis, an important consideration would be to design the components with the data bias in mind to minimize the impact. Furthermore, it is also important to continue the collection of samples from diverse locations and isolation sources to be able to capture the full extent of the global resistome.

5.3 Predictive model

The purpose of the Random forest model was to understand if it was possible to predict the success of a gene, as defined by the conceptual model, and how much information that it needed to do so at an acceptable degree of sensitivity and specificity. As seen in the results section, when the Random forest was trained on all available information, it predicted successfully with a high degree of specificity and

sensitivity with an AUROC of 99.5%, see figure 4.6. This was expected as the model was fed the same information that was used to build the success-score, which suggests that Random forests can be used for this application.

5.3.1 Simulation of a real life scenario

To reproduce a real life scenario where new a ARG arises, one random observation of each gene was selected for feature generation. To simulate the continued dissemination of the ARG and the growing amount of available information, the number of randomly selected observations therefore also increases, from 1 to 5, 10, 20, 50, 100, 250 and all observations. Building the predictive models under the different scenarios revealed that more information indeed was better, but that the AUROC > 0.84 in all cases. If the model is to be applied in a real life application, it is also required to know the performance in terms of sensitivity ad specificity. The optimal sensitivity and specificity will depend on several factors such as how the model is to be used and the costs of the actions that will be taken based upon the prediction. For example, the model could be used to build a "watch list" for future antibiotic resistant genes. In this case false positives are likely not problematic since they could be ruled out as more information is retrieved. In another case the model could be used to inform policy makers on antibiotic pollution into the environment, or antibiotic treatments, which would imply that high sensitivity and specificity is necessary.

To address the question of how much information is required to make acceptable predictions, we note that the model can to some extent predict a gene's success given only a few observations. However, a high false positive rate must be accepted for all true positives to be captured. For example, the maximum sensitivity that can be reached with 5 observations is ~77% at 90% specificity, and ~58% at 95% specificity (see figure 4.13) using the model with only features with a MDA $\geq 1.5\%$. 250 observations on the other hand yield a sensitivity approaching 1 in the model with all available features (see figure 4.7).

One limitation that must be taken into account is that in a real life application, the first few observations of a new gene are likely to be very close to each other, potentially from the same hospital, or from the same country, and therefore contain limited information. In the thesis, the first few observations could come from anywhere in the world, and contain a lot more information. This means, the predictive model in the thesis rather shows how much data is needed to capture the required information to predict the success-score. It is not realistic that the first observations of a gene are very different, and the information diversity attained from selecting five random observations in the blast data set is much larger than the five first real life observations. In future research, alternative methods to simulate the initial observations of a gene should be developed. For example, there should have been a more systematic selection of the observations to be able to could control how different the initial observations are. For example, only take observations from the same country, or only observation from the same host species.

5.3.2 Feature importance

Comparing the performance of the predictive with model with the model including all available features it should first be noted that the AUROC values are generally similar comparing the two models, and both exhibit the same trend that the higher the number of observations, the higher the AUROC. The second thing to be noted concerns the variation seen when performing cross validation of the models. For the five observation data set, the variation in AUROC is larger when all features are available (see figure 4.6B, and figure 4.10). A third and final point to note is the improvement in sensitivity that occurs as a result of reducing the number of features when there are few observations. Something that could explain both the variation and sensitivity improvements is that features that create bias or unnecessary complexity are removed. It is expected that this effect is the greatest for the five observation data set as fewer observations means that each individual data point is very important in building the model. Fewer observations can also be better from a model feasibility point of view. Since fewer data points must be collected and stored, it will be more cost effective.

The MDA from the Random forest output revealed the importance of each feature for classifying the ARGs. For high number of observations the most important features were related to geography, the number of species and the number of pathogens. This is reasonable since the geographical, taxonomic and pathogenic measures are highly correlated with the total success-score, see figure 4.2. When there are fewer observations, the importance decreases for the geographic features and for the number of species. This could be because those features do not have the ability to distinguish the ARGs as much because the range of possible values is limited to the number of observations.

Another interesting thing to note is that when there is 1 observation, the gene length feature is the most important feature. The other gene features such as GC-content, and particular codon usage are also important. Note that these factors increase in importance since no other information than gene information and number of pathogens is included in the 1 observation features. These features alone could predict the success with a sensitivity of ~70% at 90% specificity when including the most important features. This gives us reason to believe that there is a connection between the ARG dissemination patterns and the gene sequence. No linearity was however found between the gene length and success-score, nor between GC-content and the success-score, implying a more complex correlation between these factors and the dissemination. This correlation pattern remains unknown.

Regarding the codon usage, some codons had a significant contribution in the MDA. What should be noted is that the codon GTG was present as an MDA above 1.5% in all data sets with observations below 100 (figure 4.9 and A.1). GTG is an alternative start codon for bacteria, but translates to Met, just like the more common start codon ATG. Early studies have shown that GTG compared to ATG decrease the translation efficiency in both *E. coli* [70] and *B. subtilis* [71], while others have suggested GTG as a more efficient start codon in *M. gallisepticum* [72]. Nevertheless,

there is evidence that codon usage adaptation effect the fitness cost in bacteria [73] and as a consequence also the efficiency in horizontal gene transfer of antibiotic resistance genes [7, 8]. A plausible explanation to our finding is that genes with start codon GTG is spread among bacteria in certain patterns, favouring HGT in bacterial species with similar codon composition as the gene itself.

5.4 Reflection upon managerial implications

In a qualitative study by Swedish FHM, antibiotic prescribing doctors in in-patient care are interviewed about how they decide what type of antibiotic to prescribe [74]. The study outlines that the choice of antibiotic for each case often is a qualitative decision, driven by multiple factors such as available guidelines, and their personal experience. Figure 4.16 shows the prescription rates for antibiotics across countries. With the knowledge that the choice of antibiotic is not a simple straight-forward choice, it means that the difference in prescription patterns in figure 4.16 can be due to different decision making criteria in the countries. Since the resistance rates also differ between the countries, as seen in figure 4.14, it highlights the need to implement policies for decision making criteria, and continuously evaluate and educate on the topic.

Additionally, the core topic of the thesis concerns the development of a machine learning method for predicting success of antibiotic resistant genes. Being able to integrate such a machine learning method when choosing an antibiotic could be beneficial, as it would be a way to take into account the resistance potential of the infection. Another interesting application of the machine learning method is in regards to the increased demand and cost of lab and diagnostics associated with antibiotic resistance, as seen in figure 4.18. A use case here is to use new diagnostics tools as a way to reduce the time until treatment and to receive the best type of antibiotic from the beginning. This will likely also have positive impact on the cost of medication, as faster treatment will ensure the co-morbidities are not escalated by the infection and will not require as much additional treatment. Similarly, this will likely reduce the length of stay and the requirement on nursing and beds. For Sweden this will be especially important since its the country in EU with the lowest number of hospital beds per capita [75]. Ensuring that policies are formed that integrate and facilitate new technologies as a way to both improve care and reduce impact of resistance on costs, will be important going forward.

Furthermore, as seen in the results section, a factor that differs between Italy and Sweden is the public's awareness of antibiotics resistance. It is important to use antibiotics correctly when prescribed, and to always finish all prescribed antibiotics. Likely, a more educated public will be better at following the instructions since they are more knowledgeable about the issue, thus limiting the unnecessary resistance that could arise when antibiotics are consumed in the "wrong way". Since the gap in knowledge between the countries is large, it is possible that public awareness could be a contributing factor to the difference in resistance, and that there should be policies implemented to raise public awareness of the issue.

6
Conclusion

Antibiotic resistance is a threat to the global public health. With the spread of antibiotic resistant genes, common infections will become more difficult to treat, and surgery and chemotherapy will come with greater risks of receiving infections. It is vital to be proactive in limiting the spread of emerging resistance genes, and with that comes the question of what characterises a successful ARG, and are we able to predict it? Contributing to the answer of this question is also the overall aim of this thesis. We have constructed a conceptual model that describes the success of an antibiotic using four components that each are assumed to describe one dimension of "success". The four components are taxonomic, GC-context, geographic, and pathogenic.

In the thesis it is shown that the conceptual model captures genes that are known to be widely disseminated, which provides evidence that the model can describe success. The dissemination of antibiotic resistance genes is a complex process affected by several different factors, and it is possible to continue to develop the model, by refining the components or by adding additional components to better capture the full picture of dissemination. These components could include the dissemination in different environments and the time taken for a gene to spread, however, it puts additional demand on data availability which today is limited in these regards. A Random forest model was also built to predict the success of a gene, and as expected it was found that the more observations available, the better was the model at predicting the success. Though, even with few observations of a gene the model was functional. This means that this type of predictive model could become an important tool in identification of ARGs that will become successful in the future, allowing for proactive measures.

Bibliography

[1] *Antimicrobial resistance*, Accessed 2023-12-20, World Health Organization. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance`.

[2] C. J. Murray *et al.*, "Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis," *The Lancet*, vol. 399, no. 10325, pp. 629–655, 2022.

[3] J. M. Sousa, M. Lourenço, and I. Gordo, "Horizontal gene transfer among host-associated microbes," *Cell Host & Microbe*, vol. 31, no. 4, pp. 513–527, 2023.

[4] C. S. Smillie, M. B. Smith, J. Friedman, O. X. Cordero, L. A. David, and E. J. Alm, "Ecology drives a global network of gene exchange connecting the human microbiome," *Nature*, vol. 480, no. 7376, pp. 241–244, 2011.

[5] M. Sánchez-Osuna, J. Barbé, and I. Erill, "Systematic in silico assessment of antimicrobial resistance dissemination across the global plasmidome," *Antibiotics*, vol. 12, no. 2, p. 281, 2023.

[6] A. Porse, T. S. Schou, C. Munck, M. M. Ellabaan, and M. O. Sommer, "Biochemical mechanisms determine the functional compatibility of heterologous genes," *Nature communications*, vol. 9, no. 1, p. 522, 2018.

[7] D. Amorós-Moya, S. Bedhomme, M. Hermann, and I. G. Bravo, "Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage," *Molecular biology and evolution*, vol. 27, no. 9, pp. 2141–2151, 2010.

[8] T. Tuller *et al.*, "Association between translation efficiency and horizontal gene transfer within microbial communities," *Nucleic acids research*, vol. 39, no. 11, pp. 4743–4755, 2011.

[9] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[10] W. C. Reygaert, "An overview of the antimicrobial resistance mechanisms of bacteria," *AIMS microbiology*, vol. 4, no. 3, p. 482, 2018.

[11] I. L. Brito, "Examining horizontal gene transfer in microbial communities," *Nature Reviews Microbiology*, vol. 19, no. 7, pp. 442–453, 2021.

[12] Y. Hu *et al.*, "The bacterial mobile resistome transfer network connecting the animal and human microbiomes," *Applied and environmental microbiology*, vol. 82, no. 22, pp. 6672–6681, 2016.

[13] J. G. Lawrence and H. Hendrickson, "Lateral gene transfer: When will adolescence end?" *Molecular microbiology*, vol. 50, no. 3, pp. 739–749, 2003.

[14] E. Butaitė, J. Kramer, and R. Kümmerli, "Local adaptation, geographical distance and phylogenetic relatedness: Assessing the drivers of siderophore-

mediated social interactions in natural bacterial communities," *Journal of Evolutionary Biology*, vol. 34, no. 8, pp. 1266–1278, 2021.

[15] A. San Millan and R. C. MacLean, "Fitness costs of plasmids: A limit to plasmid transmission," *Microbiology spectrum*, vol. 5, no. 5, pp. 10–1128, 2017.

[16] M. Ravenhall, N. Škunca, F. Lassalle, and C. Dessimoz, "Inferring horizontal gene transfer," *PLoS computational biology*, vol. 11, no. 5, e1004095, 2015.

[17] G. Kapoor, S. Saigal, and A. Elongavan, "Action and resistance mechanisms of antibiotics: A guide for clinicians," *Journal of anaesthesiology, clinical pharmacology*, vol. 33, no. 3, p. 300, 2017.

[18] K. J. Aldred, R. J. Kerns, and N. Osheroff, "Mechanism of quinolone action and resistance," *Biochemistry*, vol. 53, no. 10, pp. 1565–1574, 2014.

[19] S. Willcocks, F. Cia, A. Francisco, and B. Wren, "Revisiting aminocoumarins for the treatment of melioidosis," *International journal of antimicrobial agents*, vol. 56, no. 1, p. 106 002, 2020.

[20] C. Weir and J. Le, *Metronidazole.* (Statpearls). StatPearls Publishing, 2023.

[21] D. Fernández-Villa, M. R. Aguilar, and L. Rojo, "Folic acid antagonists: Antimicrobial and immunomodulating mechanisms and applications," *International journal of molecular sciences*, vol. 20, no. 20, p. 4996, 2019.

[22] R. E. Hancock and D. S. Chapple, "Peptide antibiotics," *Antimicrobial agents and chemotherapy*, vol. 43, no. 6, pp. 1317–1323, 1999.

[23] G. McDonnell and A. D. Russell, "Antiseptics and disinfectants: Activity, action, and resistance," *Clinical microbiology reviews*, vol. 12, no. 1, pp. 147–179, 1999.

[24] P. J. Stogios and A. Savchenko, "Molecular mechanisms of vancomycin resistance," *Protein Science*, vol. 29, no. 3, pp. 654–669, 2020.

[25] K. P. Scott, C. M. Melville, T. M. Barbosa, and H. J. Flint, "Occurrence of the new tetracycline resistance gene tet (w) in bacteria from the human gut," *Antimicrobial Agents and Chemotherapy*, vol. 44, no. 3, pp. 775–777, 2000.

[26] M. C. Roberts, "Update on acquired tetracycline resistance genes," *FEMS microbiology letters*, vol. 245, no. 2, pp. 195–203, 2005.

[27] M. Bassetti and E. Righi, "Multidrug-resistant bacteria: What is the threat?" *Hematology 2013, the American Society of Hematology Education Program Book*, vol. 2013, no. 1, pp. 428–432, 2013.

[28] K. R. Kumar, M. J. Cowley, and R. L. Davis, "Next-generation sequencing and emerging technologies," in *Seminars in thrombosis and hemostasis*, Thieme Medical Publishers, vol. 45, 2019, pp. 661–673.

[29] N. Auslander, A. B. Gussow, and E. V. Koonin, "Incorporating machine learning into established bioinformatics frameworks," *International journal of molecular sciences*, vol. 22, no. 6, p. 2903, 2021.

[30] X. Wang, B. Yu, A. Ma, C. Chen, B. Liu, and Q. Ma, "Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique," *Bioinformatics*, vol. 35, no. 14, pp. 2395–2402, 2019.

[31] D.-H. Le, N. Xuan Hoai, and Y.-K. Kwon, "A comparative study of classification-based machine learning methods for novel disease gene prediction," pp. 577–588, 2015.

[32] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[33] L. Breiman, *Manual–setting up, using, and understandingnbsp; random forests v4.0*, 2003.

[34] B. H. Menze *et al.*, "A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC bioinformatics*, vol. 10, pp. 1–16, 2009.

[35] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[36] A. G. McArthur *et al.*, "The comprehensive antibiotic resistance database," *Antimicrobial agents and chemotherapy*, vol. 57, no. 7, pp. 3348–3357, 2013. DOI: //doi.org/10.1128/aac.00419-13.

[37] D. A. Benson *et al.*, "Genbank," *Nucleic acids research*, vol. 46, no. Database issue, p. D41, 2018. DOI: https://doi.org/10.1093%2Fnar%2Fgkx1094.

[38] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[39] O. Tange, *Gnu parallel 20200622 ('floyd')*, Zenodo. DOI: https://doi.org/10.5281/zenodo.3903853.

[40] C. L. Schoch *et al.*, "Ncbi taxonomy: A comprehensive update on curation, resources and tools," *Database*, vol. 2020, baaa062, 2020.

[41] C. Theune, *Pycountry*, Accessed 2024-01-09, PyPI. [Online]. Available: https://pypi.org/project/pycountry/.

[42] BV-BRC, *Bacterial and viral bioinformatics resource center*, Accessed 2024-01-09. [Online]. Available: https://www.bv-brc.org/.

[43] L. Breiman, A. Cutler, A. Liaw, and M. Wiener, *Randomforest: Breiman and cutler's random forests for classification and regression, version 4.6-14*, 2022.

[44] X. Robin *et al.*, "Proc: An open-source package for r and s+ to analyze and compare roc curves," *BMC Bioinformatics*, vol. 12, p. 77, 2011.

[45] W. Kuhn, M. Weston, and S. Jed, "Building predictive models in r using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008. DOI: 10.18637/jss.v028.i05. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

[46] D. Lund *et al.*, "Extensive screening reveals previously undiscovered aminoglycoside resistance genes in human pathogens," *Communications Biology*, vol. 6, no. 1, p. 812, 2023.

[47] Z. Zhang *et al.*, "Assessment of global health risk of antibiotic resistance genes," *Nature communications*, vol. 13, no. 1, p. 1553, 2022.

[48] A.-N. Zhang *et al.*, "An omics-based framework for assessing the health risk of antimicrobial resistance genes," *Nature communications*, vol. 12, no. 1, p. 4765, 2021.

[49] M. Zhuang *et al.*, "Distribution of antibiotic resistance genes in the environment," *Environmental pollution*, vol. 285, p. 117 402, 2021.

[50] EU, "Surveillance atlas of infectious disease," *European Centre for Disease Prevention and Control*, 2023.

[51] A. H. Siddiqui and J. Koirala, "Methicillin resistant staphylococcus aureus," *Statpearls*, 2018.

[52] *Meticillinresistenta gula stafylokocker (mrsa) – sjukdomsstatistik*, Accessed 2023-, Folkhälsomyndigheten. [Online]. Available: https://www.folkhalsomyndigheten.se/folkhalsorapportering-statistik/statistik-a-o/sjukdomsstatistik/meticillinresistenta-gula-stafylokocker-mrsa/.

[53] EU, "Antimicrobial consumption dashboard (esac-net)," *European Centre for Disease Prevention and Control*, 2023.

[54] EU, "Special eurobarometer 522 - antimicrobial resistance," *Directorate-General for Health and Food Safety - Public Health*, 2022.

[55] O. Jonas and A. Irwin, "Drug-resistant infections: A threat to our economic future," *Washington, DC: World Bank*, 2017.

[56] M. Prioux, "Samhällsekonomiska konsekvenser av antibiotikaresistens," *Folkhälsomyndigheten*, 2013.

[57] A. N. Poudel *et al.*, "The economic burden of antibiotic resistance: A systematic review and meta-analysis," *Plos one*, vol. 18, no. 5, e0285170, 2023.

[58] K. E. Thorpe, P. Joski, and K. J. Johnston, "Antibiotic-resistant infection treatment costs have doubled since 2002, now exceeding 2*billionannually*," *Health Affairs*, vol. 37, no. 4, pp. 662–669, 2018.

[59] W. Huang *et al.*, "In-hospital medical costs of infections caused by carbapenem-resistant klebsiella pneumoniae," *Clinical Infectious Diseases*, vol. 67, no. suppl_2, S225–S230, 2018.

[60] H. Jia *et al.*, "The attributable direct medical cost of healthcare associated infection caused by multidrug resistance organisms in 68 hospitals of china," *BioMed research international*, vol. 2019, 2019.

[61] X. Meng *et al.*, "Risk factors and medical costs for healthcare-associated carbapenem-resistant escherichia coli infection among hospitalized patients in a chinese teaching hospital," *BMC infectious diseases*, vol. 17, no. 1, pp. 1–9, 2017.

[62] X. Zhen, C. Stålsby Lundborg, X. Sun, N. Zhu, S. Gu, and H. Dong, "Economic burden of antibiotic resistance in china: A national level estimate for inpatients," *Antimicrobial Resistance & Infection Control*, vol. 10, pp. 1–9, 2021.

[63] K. Iskandar *et al.*, "Economic burden of urinary tract infections from antibiotic-resistant escherichia coli among hospitalized adult patients in lebanon: A prospective cohort study," *Value in health regional issues*, vol. 25, pp. 90–98, 2021.

[64] K. M. Krause, A. W. Serio, T. R. Kane, and L. E. Connolly, "Aminoglycosides: An overview," *Cold Spring Harbor perspectives in medicine*, vol. 6, no. 6, 2016.

[65] Y. Wang *et al.*, "Integrated metagenomic and metatranscriptomic profiling reveals differentially expressed resistomes in human, chicken, and pig gut microbiomes," *Environment international*, vol. 138, p. 105 649, 2020.

[66] L. Ma *et al.*, "Metagenomic assembly reveals hosts of antibiotic resistance genes and the shared resistome in pig, chicken, and human feces," *Environmental science & technology*, vol. 50, no. 1, pp. 420–427, 2016.

[67] D. J. Larsson and C.-F. Flach, "Antibiotic resistance in the environment," *Nature Reviews Microbiology*, vol. 20, no. 5, pp. 257–269, 2022.

[68] B. Berglund, "Environmental dissemination of antibiotic resistance genes and correlation to anthropogenic contamination with antibiotics," *Infection ecology & epidemiology*, vol. 5, no. 1, p. 28 564, 2015.

[69] W. G. Weisburg, S. M. Barns, D. A. Pelletier, and D. J. Lane, "16s ribosomal dna amplification for phylogenetic study," *Journal of bacteriology*, vol. 173, no. 2, pp. 697–703, 1991.

[70] J. K. Sussman, E. L. Simons, and R. W. Simons, "Escherichia coli translation initiation factor 3 discriminates the initiation codon in vivo," *Molecular microbiology*, vol. 21, no. 2, pp. 347–360, 1996.

[71] R. L. Vellanoweth and J. C. Rabinowitz, "The influence of ribosome-binding-site elements on translational efficiency in bacillus subtilis and escherichia coli in vivo," *Molecular microbiology*, vol. 6, no. 9, pp. 1105–1114, 1992.

[72] I. S. Panicker, G. F. Browning, and P. F. Markham, "The effect of an alternate start codon on heterologous expression of a phoa fusion protein in mycoplasma gallisepticum," *PloS one*, vol. 10, no. 5, e0127911, 2015.

[73] S. F. Bailey, L. A. Alonso Morales, and R. Kassen, "Effects of synonymous mutations beyond codon bias: The evidence for adaptive synonymous substitutions from microbial evolution experiments," *Genome biology and evolution*, vol. 13, no. 9, evab141, 2021.

[74] N. Brieger Noack, "Kunskaper och åsikter om antibiotikaförskrivning i slutenvården," *Folkhälsomyndigheten*, 2015.

[75] B. af Ugglas, "Hospital bed capacity and emergency department crowding – the impact on patient safety," *SNS Research Brief*, vol. 76, 2021.

A

Appendix

1 obs	5 obs	10 obs	20 obs	50 obs	100 obs	250 obs
gene length	number of pathogens	number of pathogens	number of countries	number of countries	number of countries	number of countries
GAG	gene length	number of continents	number of pathogens	number of pathogens	number of pathogens	number of continents
GTG	mean GC-content	number of countries	number of continents	number of continents	number of continents	number of pathogens
TTG	GGC	mean GC-content	mean GC-content	number of species	number of species	number of species
GGC	GTG	gene length	gene length	mean GC-content	Rmax,min	Rmax,min
TAT	ATA	GAA	number of species	TTG	R90,10	R90,10
number of pathogens	TAT	ATA	TAT	gene length	IQR	std
GC-content of gene	GAT	GTG	GAT	R90,10		mean GC-content
GGG	CGA	GAG	AAG	GTG		IQR
AAG	GAG	CCA	ATA	Rmax,min		
AAA	CAT	GGC	GTG	CCG		
TGG	CTT	MAD	GGC	GCC		
CGA	ACG	CAT	CAT	ATA		
GAT	number of countries	CCG		ACC		
GCA	GC-content of gene	ACC				
ATC	CTG	AAG				
CAT	ACC	GC-content of gene				
GCC	CCA	max taxonomic distance				
CTG	Rmax,min					
ACC	AAG					
GGT	GAA					
AGC						
AAT						

Figure A.1: *Features with a MDA > 1.5% for each data set.*

Table A.1: *The total success score and individual scores for the top 100 scored ARGs. For the full list of ARGs, please e-mail the authors.*

ARG	Success score	GC_S	T_S	G_S	P_S
APH(6)-Id	7,346	1,479	9,065	5,627	10,388
sul2	7,111	1,406	10,364	4,789	9,433
tet(A)	7,081	1,529	10,565	4,969	9,194
sul1	6,000	1,612	7,472	4,717	7,282
AAC(6')-Ib7	5,642	1,953	6,034	5,250	7,282
APH(3')-Ia	5,424	1,922	6,436	4,412	10,627
tet(M)	5,145	1,279	7,117	3,173	9,433
AAC(3)-Ib	4,858	2,065	5,323	4,394	6,565
mdtF	4,698	0,998	7,689	5,459	3,937
ErmB	4,587	1,616	6,467	2,706	8,716
aadA6	4,201	1,485	5,431	3,747	4,654
ANT(3")-IIa	4,168	1,148	4,194	4,143	4,892
acrD	4,047	1,851	4,875	5,376	3,220
oleB	3,964	1,321	27,558	6,381	1,547
tet(B)	3,918	0,957	2,880	4,957	5,370
floR	3,758	1,165	5,632	3,101	4,415
mdtC	3,661	0,810	4,581	5,166	2,742
aadA2	3,601	1,420	3,777	3,424	5,609
qacE	3,321	1,377	2,679	3,963	4,654
tet(C)	3,303	2,737	3,870	2,275	12,300
oqxB	3,256	1,143	4,009	4,202	2,503
msrE	3,237	1,107	3,529	2,945	3,937
AAC(6')-Ie-APH(2")-Ia	3,232	1,273	4,117	2,347	6,087
PC1-blaZ	3,220	1,167	3,483	2,957	4,892
qacEdelta1	3,070	1,978	2,679	3,460	6,087
catI	3,048	1,742	2,277	3,819	7,521
dfrA1	3,048	1,206	2,169	3,927	4,654
acrB	3,031	1,126	3,081	5,334	2,981
msbA	3,013	1,001	3,761	5,334	2,264
AAC(3)-IId	2,967	1,268	2,509	3,424	3,698
APH(3")-Ib	2,940	1,690	1,720	4,161	6,087

Continued on next page

Table A.1: *The total success score and individual scores for the top 100 scored ARGs. For the full list of ARGs, please e-mail the authors.* (Continued)

ARG	Success score	GC_S	T_S	G_S	P_S
mphE	2,919	0,995	2,893	2,945	3,698
mdsB	2,856	0,485	4,643	5,376	1,069
aadA	2,796	2,060	1,952	3,532	4,654
H-NS	2,766	0,571	3,746	5,376	1,786
smeE	2,758	0,857	6,158	3,490	2,025
tet(W)	2,716	1,974	4,333	0,838	3,459
catB3	2,699	1,049	2,045	3,352	3,937
CRP	2,695	0,533	2,648	5,250	2,742
Mrx	2,652	1,237	1,736	3,568	4,176
evgS	2,637	0,967	2,292	5,292	2,981
dfrA12	2,583	1,006	2,138	3,029	4,176
mdtH	2,555	0,915	1,890	5,208	3,220
OXA-1	2,515	0,442	1,463	3,568	3,937
tlrC	2,512	0,835	22,842	3,239	1,786
mdtB	2,494	1,613	2,246	4,789	2,742
cpxA	2,487	0,909	2,710	5,166	2,264
TEM-116	2,477	2,802	2,153	1,593	7,999
APH(3')-IIIa	2,477	1,575	2,571	2,383	6,565
tet(O)	2,474	1,559	3,390	1,449	5,848
oleC	2,435	1,030	23,275	3,322	1,547
smeB	2,422	0,952	7,426	2,820	2,025
TEM-245	2,418	0,884	0,793	3,951	4,176
aadA16	2,408	1,247	1,823	2,993	3,220
AcrF	2,406	1,040	2,787	5,376	2,025
ErmC	2,393	1,462	3,050	1,736	5,370
YojI	2,386	0,470	2,509	5,250	2,264
bacA	2,363	0,595	2,462	5,334	2,264
emrR	2,363	0,552	2,462	5,208	2,264
AcrE	2,332	0,794	2,401	5,250	2,264
ANT(2")-Ia	2,312	1,846	1,844	2,778	3,937
SAT-2	2,282	0,966	1,535	3,029	3,698

Continued on next page

Table A.1: *The total success score and individual scores for the top 100 scored ARGs. For the full list of ARGs, please e-mail the authors.* (Continued)

ARG	Success score	GC_S	T_S	G_S	P_S
dfrA14	2,270	0,936	0,921	3,604	3,698
Ecol-mdfA	2,270	0,602	1,798	5,292	2,742
aadA9	2,257	1,628	1,426	2,886	5,131
eptA	2,236	0,411	2,447	5,250	2,025
MexI	2,232	1,017	4,179	3,155	1,308
Ecol-acrA	2,228	0,472	2,432	5,292	2,025
MdtK	2,227	-0,343	4,101	5,837	0,352
mdtE	2,213	0,285	2,401	5,334	2,025
kdpE	2,213	0,362	2,401	5,292	2,025
qacL	2,201	1,706	1,875	2,526	3,459
BRP(MBL)	2,188	1,221	1,921	2,455	3,220
Paer-CpxR	2,184	1,045	4,349	3,322	0,113
marA	2,179	0,443	3,050	5,208	1,308
tet(L)	2,172	0,850	2,895	1,449	4,892
MuxB	2,148	0,824	3,251	2,987	1,308
MexD	2,139	0,998	3,483	3,281	0,830
TolC	2,132	0,415	2,478	5,250	1,786
APH(3')-VIa	2,125	1,346	1,318	2,903	3,220
mdtN	2,120	0,598	2,215	3,676	2,025
ceoB	2,117	0,796	2,447	3,281	1,786
emrB	2,109	-0,095	2,432	5,208	1,786
emrA	2,109	0,201	2,432	5,166	1,786
srmB	2,106	1,238	16,007	2,903	1,308
mecA	2,093	0,606	1,921	2,275	2,264
AAC(3)-IIe	2,065	1,149	0,839	4,286	2,981
arr-3	2,052	0,987	1,829	2,275	3,937
TEM-34	2,037	1,081	0,854	2,993	3,698
TEM-1	2,027	1,061	0,622	2,993	4,654
APH(3')-IIa	2,019	2,518	1,457	1,521	4,415
AAC(6')-Ib-cr5	2,006	1,127	0,754	2,886	2,981
QnrS1	1,996	0,430	1,035	2,957	3,459

Continued on next page

Table A.1: *The total success score and individual scores for the top 100 scored ARGs. For the full list of ARGs, please e-mail the authors.* (Continued)

ARG	Success score	GC_S	T_S	G_S	P_S
cmlA5	1,986	1,840	0,651	2,131	2,981
TEM-208	1,983	1,403	0,360	2,562	3,220
MexF	1,982	1,363	2,602	4,244	1,308
aad(6)	1,972	1,157	1,813	2,131	4,892
aadA5	1,959	1,068	0,777	2,850	4,176
AAC(6')-Ie-APH(2")-Ia	1,939	1,926	0,963	1,952	3,220
CTX-M-15	1,931	0,642	0,468	3,532	3,220

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY