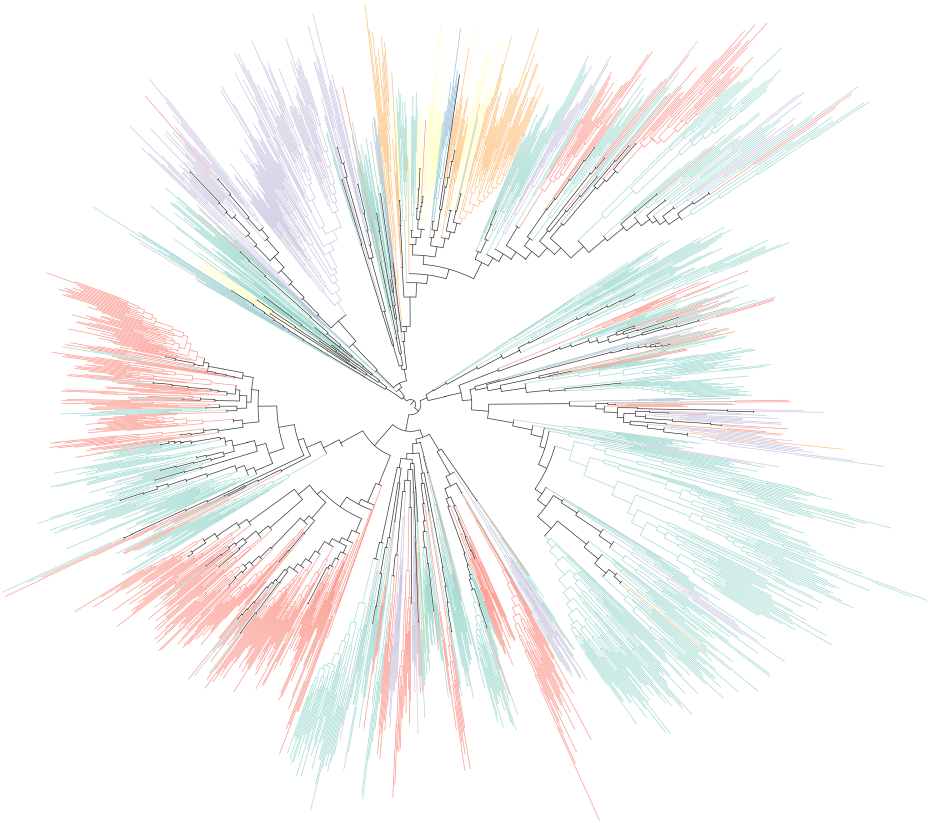

Large-scale screening of genomic data identifies novel mobile colistin resistance genes and reveals high over-representation in Pseudomonadota

Master's thesis in Biotechnology

ALICE SCHILLER

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Large-scale screening of genomic data identifies novel mobile colistin resistance genes and reveals high over-representation in Pseudomonadota

ALICE SCHILLER

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
Gothenburg, Sweden 2023

Large-scale screening of genomic data identifies novel mobile colistin resistance genes and reveals high over-representation in Pseudomonadota
ALICE SCHILLER

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

Master's Thesis 2023
Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Phylogenetic tree of predicted resistance genes colored by taxonomic class.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Large-scale screening of genomic data identifies novel mobile colistin resistance genes and reveals high over-representation in Pseudomonadota

ALICE SCHILLER
Department Mathematical Sciences
Chalmers University of Technology

Abstract

The emergence of antibiotic-resistant bacteria is a health problem of great concern. Antibiotic resistance development is driven by selection pressure and the environment is believed to be the origin of most antibiotic resistance genes, from where they can mobilize into pathogens. In order to be prepared when novel antibiotic resistance genes reach pathogens and prevent further transmission, early detection and knowledge about the spread is of high importance. One specific type of antibiotic that is of high interest to characterize is colistin. Colistin is an antibiotic that targets gram-negative bacteria and is sometimes seen as the last alternative to treat dangerous infections caused by multi-drug resistance gram-negative bacteria. The emergence of mobile colistin resistance genes hence threatens the efficiency of treating these types of infections. The aim of this thesis is to identify potential novel colistin resistance genes and evaluate them in terms of gene mobility and phylogeny.

In order to achieve this, a gene model optimized for colistin resistance genes has been created with fARGene. This model was then used to screen large-scale bacterial genomic data for potential novel colistin resistance genes. The predicted genes were analyzed in terms of mobility and phylogeny. This resulted in 680 257 predicted genes, over-represented in the class *Gammaproteobacteria* within the phylum *Pseudomonadota*, that could be summarized into 1611 clusters. Out of these clusters, 104 showed signs of mobility, and many were closely related to the already known mobile colistin resistance genes. Additionally, 13 clusters comprising potential mobile novel colistin resistance genes that are present, or at risk of ending up, in pathogenic hosts could be identified.

Keywords: antibiotic resistance, colistin, profile HMM, gene mobilization, horizontal gene transfer

Acknowledgements

First of all, I would like to thank my examiner Erik, supervisor David, and co-supervisor Anna. To Erik, thank you for welcoming me into your group and for giving me the opportunity to do this interesting thesis project. Your enthusiasm and interesting questions have inspired me along the way. To David, thank you for being such a pedagogic and patient supervisor. It has really meant a lot to me that your door always is open for questions and that you take your time to help me with all types of problems and guide me along the way. To Anna, thank you for your valuable input during our meetings. I would also like to thank you for your engagement in organizing and planning social events for the group - you should know that it means a lot.

Moreover, I would like to thank all of you, and the rest of the group for being an inspiration to what working life within academia can look like. I can truly say that my positive experience with doing my master's thesis in your group has been an important factor in my decision to continue with PhD studies.

Finally, I would like to thank my family and friends for always supporting and encouraging me in my work. Your love and support always make me feel better when I have doubted myself along the way.

Alice Schiller, Gothenburg, June 2023

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ARG	antibiotic resistance gene
Dtr	DNA transfer and conjugal replication
EptA	Lipid A phosphoethanolamine transferase
fARGene	Fragmented Antibiotic Resistance Gene idENtifiEr
HGT	horizontal gene transfer
HMM	Hidden Markov model
ICE	integrating conjugative element
IR	inverted repeats
IS	insertion sequence
LPS	lipopolysaccharide
MCR	mobile colistin resistance
Mpf	mating pair formation
PEA	Phosphoethanolamine

Contents

List of Acronyms **viii**

1 Introduction **1**
1.1 Aim . . . 2

2 Theory **3**
2.1 Colistin resistance . . . 3
2.1.1 Mode of action . . . 3
2.1.2 Resistance mechanism . . . 4
2.1.3 Mobile colistin resistance . . . 5
2.2 Gene mobilization . . . 6
2.2.1 Mobilization within the genome . . . 6
2.2.2 Horizontal gene transfer . . . 8
2.3 Hidden Markov models . . . 9
2.3.1 Profile Hidden Markov Model . . . 9
2.4 Fargene . . . 12

3 Methods **15**
3.1 Model creation . . . 15
3.1.1 Data . . . 15
3.1.2 Model optimization . . . 15
3.2 Screening and phylogenetic analysis . . . 16
3.3 Mobility analysis . . . 17

4 Results **19**
4.1 Gene models optimized for colistin resistance . . . 19
4.2 Screening and taxonomic over-representation . . . 20
4.3 Phylogentic relationship to known resistance genes . . . 21
4.4 Phylogentic and mobility analysis . . . 22
4.4.1 Clade A1 . . . 24
4.4.2 Clade A2 . . . 27
4.4.3 Clade A3 . . . 28
4.4.4 Clade A4 . . . 29
4.4.5 Clade A5 . . . 31
4.4.6 Horizontal gene transfer and pathogens . . . 32

5 Discussion **35**

5.1 Model creation and optimization 35
5.2 Analysis of phylogeny and mobility 36

6 Conclusion and Future Outlook **39**

Bibliography **41**

A Appendix A **I**

B Appendix B **III**

1

Introduction

Antibiotic resistance is a growing health problem all over the world, causing bacterial infections that are difficult, or in the worst case impossible, to treat [1]. According to the World Health Organization (WHO), even infections that today are easy to treat can become deadly in the future, thus putting humanity in a similar situation as before antibiotics were discovered. In 2019, antibiotic resistance was estimated to be associated with 4.95 million deaths whereas 1.27 million of these were directly caused by resistant bacterial infections [2].

Some of the most common infections caused by resistant bacteria are lower respiratory infections and bloodstream infections such as pneumonia and blood poisoning [2]. Up to 80 % of these cases were caused by only 6 different bacteria; *Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Streptococcus pneumoniae*, *Acinetobacter baumannii*, and *Pseudomonas aeruginosa*. Particularly alarming is multi-drug resistance gram-negative bacteria, which in the worst case can be resistant to almost all available antibiotics. Multi-drug resistance has been detected in common pathogenic gram-negative bacteria such as *Klebsiella pneumoniae*, *Acinetobacter baumannii*, and *Pseudomonas aeruginosa* [3],[4].

The last alternative antibiotic to treat some of these dangerous infections is colistin [3], [4]. Colistin is a polypeptide antibiotic that targets the membrane of gram-negative bacteria. It was used in clinics during the 1950s-1960s but was later replaced by other antibiotics due to its severe side effects. However, from the beginning of the 21st century, the emergence of multi-drug resistance gram-negative bacteria has required the use of colistin once again. Resistance to colistin has previously been restricted to mutations in chromosomal genes that can not be transmitted between bacteria, but in 2015 the first mobile colistin resistance (MCR) gene, MCR-1, was detected [5]. Since then, nine additional MCR genes (MCR-2 up to MCR-10) have been discovered [6], threatening the efficiency of colistin as an alternative to treat multi-drug resistant gram-negative bacterial infections.

Mobilization of resistance genes constitutes a major problem since it enables them to be transmitted between bacteria and eventually end up in pathogens. Due to the huge genetic reservoir found in the environment, most antibiotic resistance genes (ARGs) are believed to originate from environmental bacteria where the genes initially had other functions [7]. The genes are transferred from environmental bacteria into pathogens, either entirely within the environment or via the microbiota of animals or humans, accelerated by the selection pressure caused by the heavy use of

antibiotics. The transfer of mobile resistance genes into pathogens is a process of several steps. First, chromosomal encoded ARGs develop the ability to move within the genome. This is followed by horizontal gene transfer (HGT), which is the process of transferring genetic elements between bacteria in a population.

In order to limit this spread and be prepared when novel ARGs reach human pathogens, it is essential to identify possible future ARGs that risk mobilizing from the environment [7]. Thanks to the rapid development of next-generation sequencing methods, large amounts of bacterial genomic data are available for analysis. This enables effective computational methods to be used in comparison to classical laboratory analysis, which is hampered by difficulties with cultivating many environmental bacteria [8]. One such method is Fragmented Antibiotic Resistance Gene idENtifiEr (fARGene) which uses Hidden Markov models (HMMs) to identify novel ARGs from bacterial genomic data [8].

1.1 Aim

This project aims to perform a large-scale characterization of colistin resistance genes. The goal of the characterization is to identify potential novel colistin resistance genes and analyze them in terms of gene mobility and phylogeny. More specifically, the mobility analysis aims to investigate the risk of mobilization of novel colistin resistance genes and the goal of the phylogenetic analysis is to reveal which type of bacteria carry colistin resistance genes and how both known and predicted resistance genes are related to each other.

To achieve the aim of this project, probabilistic gene models optimized for colistin resistance genes will be constructed with fARGene and be used to screen databases with bacterial genomes for potential novel colistin resistance genes. Mobilization will be evaluated by searching for mobile genetic elements in the regions surrounding the resistance genes predicted by the model. A phylogenetic tree of the resistance genes will be created to investigate both the relationship between the predicted genes and their relationship to the already-known colistin resistance genes. In order to decide which type of bacteria harbor these types of genes, and if the predicted genes are found in pathogens, the bacterial host of the predicted colistin resistance genes will be determined.

2

Theory

The information given in this section will explain the mode of action of colistin and how the resistance mechanism works. It will also cover the status of the resistance today and how the current MCR genes are related to each other. The mechanisms behind mobilization, both within the genome and HGT, will also be explained. Lastly, the theory behind fARGene will be explained, with emphasis on probabilistic gene models based on profile HMMs.

2.1 Colistin resistance

2.1.1 Mode of action

Colistin, or polymyxin E, is a member of the polymyxin antibiotic family [4]. It is a cationic polypeptide antibiotic that targets the outer membrane of gram-negative bacteria. Gram-negative bacteria's cell envelope consists both of an inner membrane and an outer membrane, which are separated by a periplasmic space with peptidoglycan. The outer membrane is asymmetric and consists of an inner layer of phospholipids and an outer layer of lipopolysaccharide (LPS) [9]. LPS consists of lipid A, which is anchored to the outer membrane, as well as an oligosaccharide and an antigen. The structure of the outer membrane and LPS can be seen in Figure 2.1.

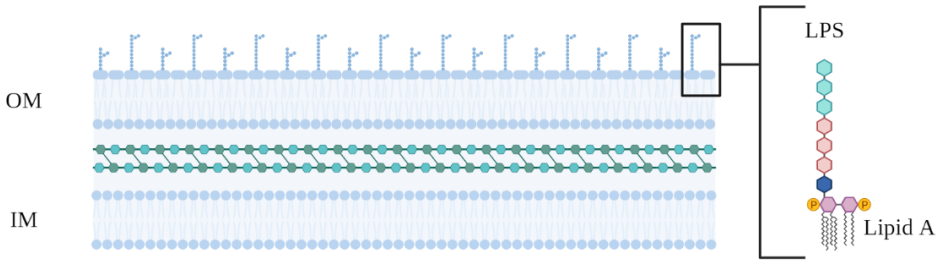

Figure 2.1: The cell envelope of gram-negative bacteria and the structure of LPS.

The lipid A part of LPS is the exact target of colistin. Colistin consists of one hydrophobic and one polar part [10]. The polar part of colistin targets the LPS on

the outer membrane through electrostatic interaction between the cationic colistin and the negatively charged phosphate groups on the lipid A domain of LPS. This interaction makes it possible for the hydrophobic part of colistin to interact with the fatty acid chains in the outer membrane and increase its permeability. This way, colistin can reach the inner cell membrane and eventually cause cell lysis and death.

2.1.2 Resistance mechanism

Common mechanisms of gram-negative bacteria used to adapt to the surrounding environment or protect from antibiotics is to modify the lipid A domain of the LPS [9]. One such mechanism that gives resistance to colistin is performed by the inner membrane enzyme lipid A phosphoethanolamine transferase (EptA). EptA modifies the LPS by adding phosphoethanolamine (PEA) to the phosphate group of lipid A [11], thus decreasing the negative charge. This way, the bacterium is protected from colistin since the electrostatic interaction can not occur.

EptA is regulated by two different two-component regulatory systems; PmrA/PmrB and PhoP/PhoQ (Figure 2.2) [11]. A two-component regulatory system consists of a sensor kinase that reacts to outer stimuli and a response regulator that controls gene transcription based on signals from the sensor kinase [9]. This way, bacteria can change their gene expression based on outer stimuli. For example, different ions, changes in pH, and cationic polypeptide antibiotics such as colistin stimulate PmrA/PmrB and PhoP/PhoQ [11]. Outer stimuli, together with mutations in the PmrA/PmrB and PhoP/PhoQ systems, upregulates EptA and can cause resistance to colistin.

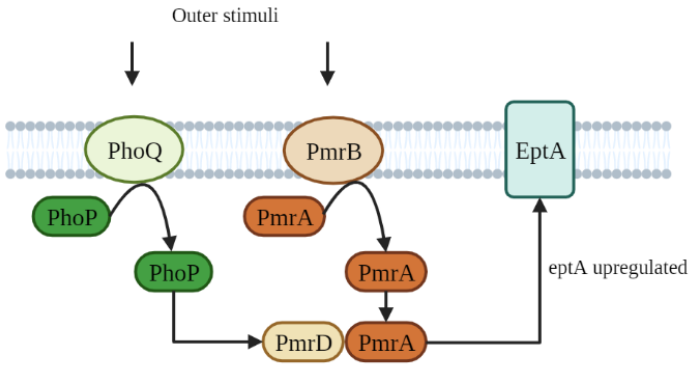

Figure 2.2: The PmrA/PmrB and PhoP/PhoQ two-component regulatory systems that control the expression of the chromosomal colistin resistance gene EptA. Here it can be seen that PhoP/PhoQ is indirectly controlling EptA via the activation of PmrD.

2.1.3 Mobile colistin resistance

The resistance mechanism used by EptA is also used by the more recently discovered MCR genes [12]. Ten different mobile colistin resistance genes are identified today (MCR-1 up to MCR-10), which all give resistance by adding PEA to lipid A. It is believed that all MCR proteins, as well as EptA, are structurally and functionally unified [12]–[14]. The enzyme consists of two domains with a catalytic site in the interface. In the catalytic site, a PEA is cleaved from a phosphatidylethanolamine (PE) lipid. Then, the PEA can be transferred to the lipid A domain of the LPS and change the charge of the membrane. Although the structures and functions of the enzymes are conserved between MCR proteins, the group is still very phylogenetically diverse, meaning that not all the sequences in the MCR family share high similarity to each other. The phylogenetic relationship between the MCR genes can be seen in Figure 2.3.

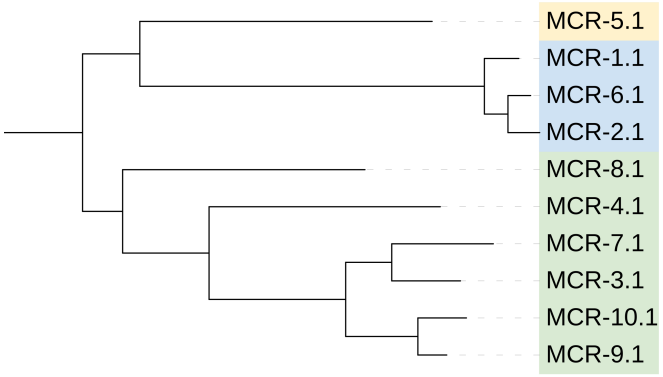

Figure 2.3: Phylogentic relationship of the ten MCR genes. Two distinct groups are marked in blue and green. MCR-5 is marked in yellow since it is very distantly related to both groups.

MCR-1, MCR-2, and MCR-6 (previously denoted MCR-2.2) form one group of phylogenetically close genes (marked in blue in Figure 2.3). It is believed that *Moraxella* species are a common progenitor for the genes [15], [16]. MCR-1 is the first MCR gene that was identified and was found in *Escherichia coli* [5]. Today, MCR-1 has been observed in many other pathogens such as *Klebsiella pneumoniae*, *Salmonella enterica*, and *Enterobacter* [17], [18]. MCR-2, which also was detected in *Escherichia coli*, shows 81 % sequence identity to MCR-1 [19]. Additionally, MCR-6 that was found in *Moraxella sp.* [20] shows above 80 % amino acid identity to both MCR-1 and MCR-2 [21]. There are also intrinsic colistin resistance genes (ICR) from *Moraxella* that are closely related to this group of MCR genes [22].

The other group of resistance genes comprises MCR-3, MCR-4, MCR-7, MCR-8, MCR-9, and MCR-10 (marked in green in Figure 2.3). Within this group, MCR-3 and MCR-7 are closely related with 70 % amino acid identity and are believed to originate from *Aeromonas* species [23], [24]. Further, MCR-9 and MCR-10 are phylogenetic close to each other with 83 % amino acid identity and are believed to have *Buttiauxella* species as progenitors [12]. There are also ICR genes from *Buttiauxella*

that are closely related to MCR-9 and MCR-10 [12]. MCR-4 and MCR-8 are also placed in this group of MCR genes. However, they are more distantly related. MCR-4 is most closely related to MCR-3, with 49 % amino acid identity [25]. MCR-4 was initially found in *Escherichia coli* and *Salmonella enterica* and its believed progenitor is *Shewanella* species. MCR-8, first discovered in *Klebsiella pneumoniae* [26], are most closely related to MCR-9 with 44 % amino acid identity and are believed to originate from *Stenotrophomonas* [27]. MCR-5 has been suggested also to be included in this group of resistance genes [14]. However, other studies place MCR-5 in a completely separate clade from all MCR genes [21], and the gene shows only around 35 % amino acid identity to MCR-1, MCR-2, MCR-3, and MCR-4 [28].

2.2 Gene mobilization

As mentioned in the introduction, bacteria possess the ability to transfer genetic material within a population [29]. Driven by selection pressure, chromosomal ARGs often located in environmental bacteria can move into pathogens [7]. This process can occur both within the environment and via the microbiota of humans and animals. The development from a chromosomal immobile ARG to a mobile ARG that can transfer between bacteria can be divided into two main steps. Initially, mobilization within the genome occurs which means that the ARG acquires the ability to move from one place in the genome to another. This ability is essential for the second step to occur; reallocation to a mobile element that can move between bacteria. When the ARG is located in a mobile element, such as a plasmid, it can be transferred between bacteria through HGT. The following sections will explain the two steps of mobilization, as well as HGT in more detail.

2.2.1 Mobilization within the genome

One genetic element that is essential for mobilization and involved in the spread of antibiotic resistance is transposons [30, p. 288]. A transposon is a DNA segment with the ability to move between positions in the genome. This process is called transposition and is used to move genetic elements within the genome [29, p. 321-325]. The transposon carries two elements that are required for transposition: the transposase enzyme and the inverted repeats (IR) flanking the transposon. The role of the transposase enzyme is to recognize the sequence of the IRs. IRs are DNA segments where the 5'-3' sequences of their opposite strands are almost the same. This makes it possible for one transposase enzyme to bind to each IR and also to each other, to form an excised transposon. Through this process, the transposon can "cut and paste" from one part of the genome to another. The process of transposition can be seen in Figure 2.4.

The smallest transposon is called insertion sequence (IS) element. IS elements consist only of the transposase and IRs and their only function is therefore to move within the genome [29]. However, a transposon can also carry passenger genes and therefore carry and move one or several ARGs. One way that ARGs inserts into the transposon is the formation of a composite transposon. A composite transpo-

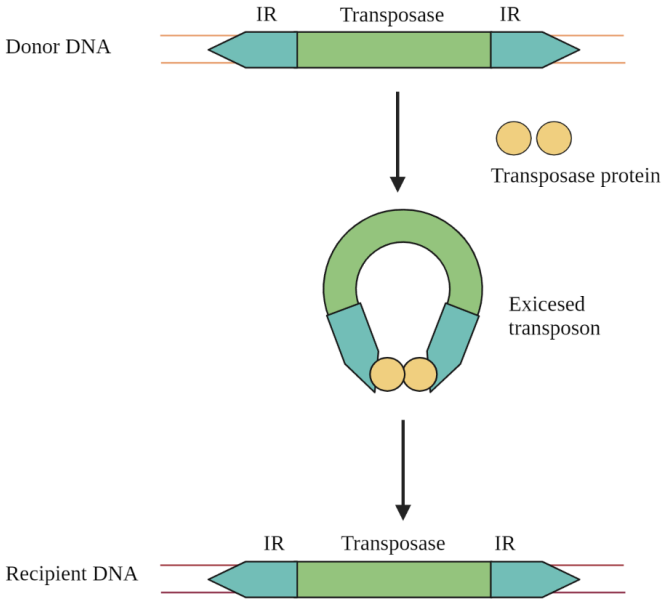

Figure 2.4: The process of transposition. Here, the IRs of the transposon bind to the transposase protein to form an excised transposon. This way, the excised transposon can insert into another part of the DNA and move between locations in the genome.

son consists of two IS elements that flank other genes and is created when two IS elements of the same type insert close to each other in the genome. The difference between an IS element and a composite transposon can be seen in Figure 2.5.

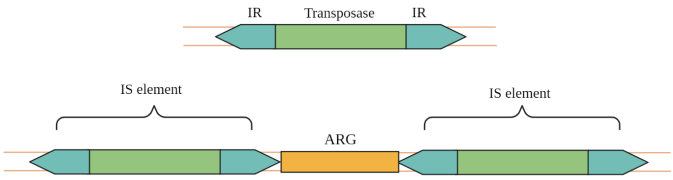

Figure 2.5: An IS element and a composite transposon that consists of an ARG flanked by two IS elements.

Another possibility for ARGs to integrate into a transposon is by the integration of gene cassettes [29]. A gene cassette is a circular, non-replicating, DNA segment that has excised from the genome. It consists of one or more ARGs and their attachment sites. The integration of a gene cassette requires that the transposon carries an integron. An integron is a DNA segment that consists of an integrase gene and an attachment site. The integrase recognizes the attachment site of the gene cassette, and through recombination with the attachment site of the integron, the gene cassette can be integrated into the transposon.

ARGs located in a transposon, either via the formation of a composite transposon

or by integration of gene cassettes, can reallocate to a mobile element through transposition. The mobile element can transfer between bacteria and thereby spread the ARGs within a population. There are two types of mobile elements mainly utilized in the transfer of ARGs between bacteria: plasmids and integrating conjugative elements (ICEs) [29, p.215-240]. ICEs, in contrast to plasmids, are not free from the chromosomal DNA but integrated into it. They can excise from the chromosomal DNA as free circular elements and transfer into other bacteria. The next section will explain how these elements can be transferred between bacteria through HGT.

2.2.2 Horizontal gene transfer

HGT is the transfer of DNA within a bacterial population and makes it possible for bacteria to share genetic material by other means than reproduction [29, p.215-240]. There are three types of HGT: transformation, transduction, and conjugation. Transformation means that bacteria take up free DNA from the environment such as genetic material from dead bacteria. Transduction is the process of transferring DNA through a bacteriophage between bacteria.

Out of these three mechanisms, conjugation is the most prevalent in the transfer of ARGs. Conjugation is the ability to transfer DNA between two bacteria through a protein channel [29, p.215-240]. The formation of the protein channel and other necessary transferring functions requires a conjugation system which can be found in ICEs and self-transmissible plasmids. A self-transmissible plasmid is a plasmid that contains all genes necessary to transfer itself to other bacteria. There are also mobilizable plasmids, that contain parts of the conjugation system and with help from a self-transmissible plasmid can conjugate to another bacteria.

The essential part of the conjugation system is the transfer (*tra*) genes that code for two components required for the conjugation; the DNA transfer and conjugal replication (Dtr) component and mating pair formation (Mpf) component [29, p.215-240]. Here, the Mpf component is responsible for the formation of the protein channel and for holding the donor and recipient cells together during the conjugation. This component also codes for a coupling protein whose function is to contact the Dtr component when the bacteria is ready for conjugation. The Dtr component is responsible for the preparation of the DNA that should be transferred in the donor cell as well as the ligation in the recipient cell. For this, the Dtr component includes the relaxase protein that cleaves the plasmid at the origin of transfer (oriT) which creates the DNA strand to be transferred through the channel [31]. It also contains one or more nicking proteins that ligate the DNA strand in the recipient cell. A mobilizable plasmid carries only the *tra* genes that code for the Dtr component [29, p.215-240]. Another name that is commonly used for the genes coding for the Dtr component is *mob* genes. In the case of a mobilizable plasmid, the self-transmissible plasmid is responsible for the connection to the recipient bacteria and the formation of the channel. The coupling protein of the self-transmissible plasmid then contacts the mobilizable plasmid that can prepare for conjugation through its own Dtr system.

When the donor and recipient cell are connected through the protein channel, DNA

from the plasmid, or ICEs, can be transferred [29, p.215-240]. A single strand is transferred from the donor to the recipient which then can act as a template during DNA replication in the recipient cell. If the DNA came from ICEs, the circular element can be integrated into the chromosome of the recipient cell. This way, ARGs can both transfer as plasmids from the donor bacteria as well as integrate into the chromosomal DNA of the recipient bacteria.

2.3 Hidden Markov models

This section is based on the theory about HMMs and profile HMMs described in [32]. A Hidden Markov model (HMM) is a probabilistic model. Saying that a model is probabilistic means that the model produces different outputs with certain probabilities. In this case, the output is protein sequences and the model can be seen as a probability distribution over all possible sequences which has its peak for sequences belonging to the protein family of interest.

An HMM consists of states and symbols. In the case of protein sequences, the states are positions in the sequence and symbols are amino acids at each position. Saying that the model is hidden, means that only the sequence of symbols can be observed, not the states visited to produce it. The sequence of states visited in the model is called a path and is denoted π. The probability of moving to state π_i from state π_{i-1} is called transition probability and is denoted

$$a_{kl} = P(\pi_i = l | \pi_{i-1} = k), \tag{2.1}$$

where the probability only is dependent on the previous state. Moreover, the probability of observing a certain symbol (amino acid) in a state is called emission probability and is denoted

$$e_k(b) = P(x_i = b | \pi_i = k) \tag{2.2}$$

, where $e_k(b)$ is the probability of observing the symbol b in state k.

2.3.1 Profile Hidden Markov Model

A Profile HMM is a specific type of HMM with the purpose to search databases for sequences that belong to a protein family. The model is built from a Multiple sequence alignment (MSA) which shows how sequences in a family are related to each other. In an MSA, each position of the sequence is represented by a column, where each row corresponds to a single sequence. An example of an MSA can be seen in figure 2.6.

Seq1	A	E	-	C	E
Seq2	A	E	-	C	-
Seq3	A	-	-	C	A
Seq4	A	E	D	-	E
Seq5	A	C	D	A	E
	*	*		*	*

Figure 2.6: A multiple sequence alignments of five sequences. The letters are amino acids and the lines are gaps. Columns marked with a star can be seen as more conserved between the sequences.

Protein families often have parts of the protein sequences that are more conserved. It is desirable to capture these parts in order to identify new sequences that belong to the family. To capture the variations of conservation three different states are needed in the Profile HMM; match states, insertion states, and deletion states. The conserved parts are used to build the match states and the number of columns from the MSA used for this is called the length of the model. In Figure 2.6, the conserved parts are marked with a star. A rule of thumb is to include columns where less than 50 % of the rows are gaps. However, this can also be decided by a dynamic programming algorithm that will not be covered here. Figure 2.7 shows the structure of the match-part of the profile HMM.

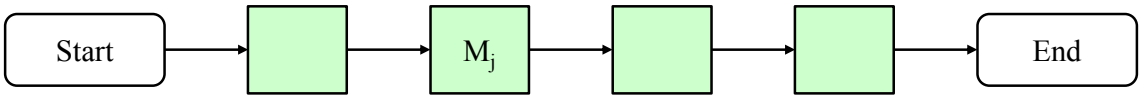

Figure 2.7: Match states in a profile HMM of length 4. The boxes correspond to states and the arrows are transition probabilities $a_{j-1,j}$.

The match part consists of match states and transition probabilities between matches. When only looking at the match-part, all transition probabilities are 1 since there are no other possible transitions than $M_{j-1} \rightarrow M_j$. Using only this part of the model, the probability of a sequence x is therefore the sum of the emission probabilities and can be written as

$$P(x) = \prod_{i=1}^{L} e_i(x_i) \tag{2.3}$$

, where L is the length of the model and $e_i(x_i)$ is the probability to observe symbol x_i in position i. However, more interesting is to evaluate the log-odds ratio. The log-odds ratio is the logarithm of the ratio of the probability of x being produced by the model to the probability of x being produced by a random model. Taking the logarithm of the product of the probabilities gives a sum that can be written as

$$S = \sum_{i=1}^{L} log \frac{e_i(x_i)}{q_{x_i}}, \tag{2.4}$$

where q_{x_i} is the probability of the symbol x_i being produced by a random model at position i.

As mentioned, only parts of the protein family are conserved. To handle parts of the sequence that do not match the match-part of the model, as well as gaps in the match-part, insertion, and deletion states are added. Insertions are handling gaps that correspond to part of the MSA that is not included in the match-part while deletions are handling gaps within the match-part of the model. The log-odds score for the insertions are

$$log(a_{M_j I_j}) + log(a_{I_j M_{j+1}}) + (k-1)log(a_{I_j I_j}) \tag{2.5}$$

where k is the length of the insertion. The cost corresponds to the transition $M_j \rightarrow I_j$, k-1 transitions $Ij \rightarrow Ij$, as well as the transition $Ij \rightarrow Mj+1$. There are emission probabilities within the insertion states as well, but these are assumed to be distributed the same way as a random model, thus the log-odds score for this is zero ($log\frac{q_{x_i}}{q_{x_i}} = 0$). The log-odds score for the deletions is the sum of transition probabilities $M \rightarrow D$, a number of $D \rightarrow D$ transitions, and one $D \rightarrow M$ transition. The deletion states have no emissions and can be seen as silent states between match states. The structure of a profile HMM with all states included can be seen in Figure 2.8.

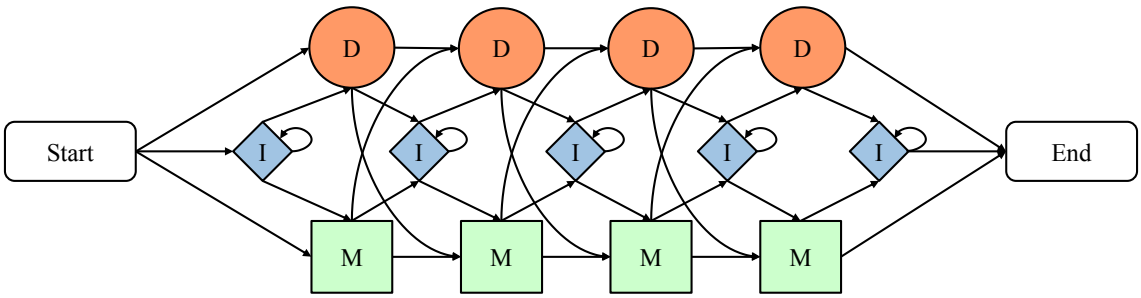

Figure 2.8: Complete profile HMM. The arrows correspond to transition probabilities between states. Match states are represented by squares, insertion states by diamonds, and deletion states by circles.

In order to peak the probability distribution of the model for proteins belonging to the protein family, the transition and emission probabilities need to be estimated. There are many possible ways to estimate the probabilities, and all of them will not be covered here. One possible way is simply to count the number of occurrences of each transition and emission given the protein sequences in the MSA. However, if the dataset is small, there is a risk that some transmission and emission never occur and therefore the probabilities will be zero. One approach to solve this is to add 1 pseudo count to each transition and emission. There are other methods to add pseudo-counts as well, that will not be covered here.

When the model is built, it can be used to decide if a new sequence x belongs to the protein family or not. As mentioned, the scoring is based on additive log-odds scoring that evaluates the match of sequence x to the model in comparison

to a random sequence model. In the profile HMM used in this project, the scoring considers all possible alignments of the sequence x to the model [33]. This gives a score based on the relative likelihood of each possible alignment of x to the profile in comparison to other scoring systems where the score only considers the optimal alignment.

2.4 Fargene

fARGene (Fragmented Antibiotic Resistance Gene idENtifiEr) is a tool that predicts ARGs from bacterial genomic input data [8]. The genomic input data can be either metagenomes or longer sequences. In this thesis, fARGene was used to predict colistin resistance genes from whole genomes, thus the theory will focus on how the tool can be used for longer sequences.

The overall workflow of fARGene can be seen in Figure 2.9. A probabilistic gene model is used to classify genes from the input data as potential ARG or non-ARG [8]. The probabilistic gene model used for the classification is a profile HMM that captures the patterns of the conserved parts of resistance genes. This means the gene model needs to be optimized specifically for each group of antibiotic resistance genes to capture the conserved sequences. In this thesis, the gene model will be optimized for colistin resistance genes.

Figure 2.9: The overall function of fARGene. Genes from bacterial genomic data are classified as ARG or non-ARG by an optimized Profile HMM.

There are two parameters that are important when optimizing the Profile HMM for the antibiotic of interest: sensitivity and specificity [8]. Sensitivity is the ability to correctly detect all functional ARGs. Specificity is defined as the ability to not classify evolutionary close genes, without the resistance mechanism, as ARGs. Thus, a model with high specificity minimizes the false positives.

Optimizing the model means that a log-odds threshold score is chosen based on estimated sensitivity and specificity [8]. Then, when screening the genomic data for resistance genes, sequences given a log-odds score above the threshold score will be classified as ARGs, and sequences with a log-odds score under the threshold will be classified as non-ARGs. The sensitivity of the model is estimated using leave-one-out cross-validation of the known ARGs. This means that one sequence at a time is removed from the set of known resistance genes and a profile HMM is created based

on the remaining sequences, followed by applying the model to the removed sequence. To estimate the specificity, the model is applied to a negative dataset consisting of genes evolutionary close to the ARGs but without the resistance mechanism.

The sensitivity and specificity are based on the fraction of correctly classified sequences [8]. Here, a correctly classified sequence is based on a significance threshold score (E-value), which is a predefined setting in fARGene. An optimal log-odds threshold score is then chosen so that both sensitivity and specificity are high. This way, the model can separate ARGs from closely related genes without the resistance phenotype.

3

Methods

3.1 Model creation

In the following section, the process of creating the Profile HMM will be described. First, the data used to build and estimate the performance of the model will be presented followed by the model optimization.

3.1.1 Data

In total, 49 sequences of known resistance genes were used to build the model. 43 out of these were colistin resistance genes retrieved from Comprehensive Antibiotic Resistance Database (CARD) (accessed January 2023) whereof 41 MCR genes and 2 intrinsic resistance genes from *Moraxella* (ICR-M) [34]. The additional 6 sequences were intrinsic resistance genes from *Buttiauxella* (ICR-B) that were retrieved from National Center for Biotechnology Information (NCBI) [35]. All 49 sequences come from *Gammaproteobacteria*. A full list of the resistance genes used to build the profile HMMs can be seen in Appendix A.

The negative dataset used to estimate the specificity consists of sulfatases from the protein family PEA transferases (accessed from InterPro January 2023) [36]. This family consists of around 21 000 proteins and to get a smaller negative dataset the sequences were clustered at 70% amino acid identity. This was done using USEARCH v.8.0.1445 with the parameters "-cluster_fast -id 0.7" [37]. This resulted in 1595 sulfatase sequences as the negative dataset.

3.1.2 Model optimization

The model optimization was performed with fARGene v.0.1 using the tool "fargene_model_creation" with the parameter "- -only-full-length" [8]. Since the resistance genes are phylogenetically separated into two groups, it was decided that both separating the resistance genes into two models and combining all sequences into one model should be tested. The separated Profile HMMs were based on the two groups of resistance genes: MCR-1, MCR-2, MCR-6, and ICR-M in one group, and MCR-3, MCR-4, MCR-7, MCR-8, MCR-9, MCR-10, and ICR-B in the other group. As mentioned in the theory, the placing of MCR-5 into one of the groups is not evident. Considering this, two different separations were tested; one where MCR-5 was grouped together with MCR-1, MCR-2, MCR-6, and ICR-M, and one where

MCR-5 was grouped with the other known resistance genes. In the combined Profile HMM all 49 resistance gene sequences were included. For all three models, the sensitivity was estimated with cross-validation and the specificity with the negative dataset, as described in the theory about fARGene. The same negative dataset was used to estimate the specificity of all models. A log-odds threshold score was then chosen so that both sensitivity and specificity were high.

In order to validate the different models further before performing the large-scale screening for resistance genes, a small subset was used in an initial screening. The subset consisted of 31 187 genomes from NCBI Reference Sequence Database (RefSeq) (accessed February 2023) [35]. RefSeq is a part of NCBI GenBank, which is the database used for large-scale screening.

There was no considerable difference in performance between the separated models and the combined model in terms of sensitivity and specificity. Additionally, the sum of the number of predicted genes from the separated models was almost the same as the total from the combined model, thus only the combined Profile HMM was used in the following large-scale screening.

3.2 Screening and phylogenetic analysis

1 018 179 genomes from NCBI GenBank (accessed January 2022) were screened for novel colistin resistance genes using the combined Profile HMM built with fARGene [35]. A cut-off score of 500 was used, which means that sequences given a score over 500 are classified as resistance genes, and sequences given a score less than 500 are classified as non-resistance genes. The genes classified as colistin resistance genes were clustered at 70% amino acid identity using USEARCH v.8.0.1445 with the parameters "-cluster_fast -id 0.7" [37]. The centroids of each cluster were aligned using MAFFT v.7.515 [38] and a phylogenetic tree was created with FastTree v.2.1.11 [39]. The tree was midpoint rooted and visualized with Interactive Tree of Life (iTOL) [40].

To understand which type of bacteria carries the predicted colistin resistance genes the taxonomy was determined. Each sequence predicted by the model has a contig accession number, representing the contig carrying the gene. The genome assemblies from NCBI GenBank consist of multiple contigs that have unique contig accession numbers. In order to map the sequences to their corresponding genome assembly, the contig accession numbers were mapped to assembly accession numbers. The taxonomic ID for each sequence was then retrieved from the assembly summary file at NCBI, where each assembly accession number is mapped to a taxonomic ID [41]. The function "getTaxonomy" of the R-package taxonomizr was then used to get full lineage taxonomy (kingdom, phylum, class, order, family, genus, and species) from the taxonomic ID [42], [43]. An example, that shows the flow from contig accession number to full linage taxonomy, can be seen in figure 3.1.

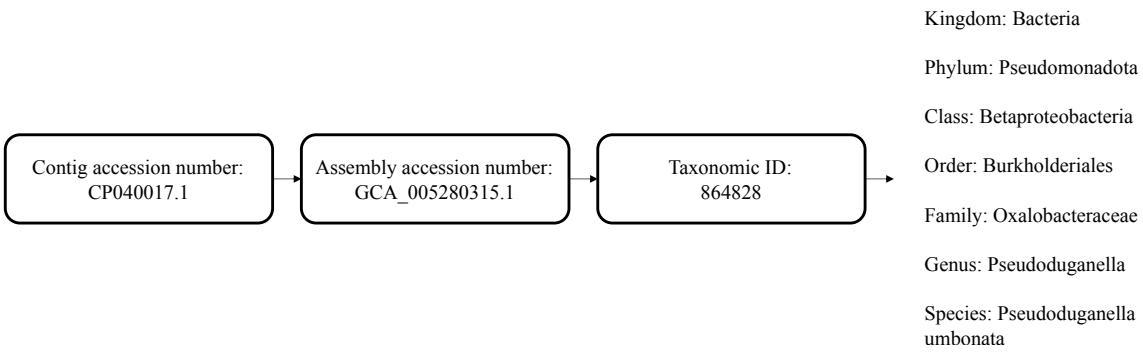

Figure 3.1: The flow from contig accession number to full linage taxonomy. Each contig accession number is mapped to its assembly accession number. Each assembly accession number is mapped to a taxonomic ID. The taxonomic ID is used to get full lineage taxonomy.

The taxonomy of the clusters was summarized based on taxonomic class. The dominating class of each cluster was used to represent the taxonomy of each cluster. From this, five clusters were distinctive since they consisted of sequences from three or more classes, while the others consisted of sequences from only one or two classes. However, only a few sequences in these clusters came from a divergent class than the dominating, thus it was decided to verify the taxonomy of these sequences. The verification was performed with MeTaxa2 which assigns taxonomy based on small and large rRNA subunits [44]. From the MeTaxa results, it could be concluded that some of the sequences from divergent classes were miss classified in NCBI and others were assumed to be contaminated since rRNA subunits from multiple classes were found within the same genome. Therefore, no clusters were assumed to include sequences from 3 or more classes and all clusters could be represented by its dominating class.

From the results of the taxonomy assignment, it seemed like the class *Gammaproteobacteria* was over-represented in bacteria that carry a resistance gene. To control this, Fischer's exact test was performed to investigate the over- and under-representation of *Alpha-*, *Beta-*, *Gamma-*, *Delta-*, and *Epsilonproteobacteria*. The over- and under-representation of the phyla *Pseudomonadota*, *Bacillota*, *Actinomycetota*, and *Bacteroidota* were also controlled. A test was considered significant if the p-value < 0.001.

To understand how the predicted resistance genes are related to the known resistance genes the clusters with the highest amino acid identity to the known resistance genes were determined. This was done using the tool "blastp" in Blast v.2.5.0 [45]. The resistance genes used to build the model as well as the chromosomal gene EptA were used in the blast database.

3.3 Mobility analysis

In order to analyze the risk of horizontal gene transfer of the predicted resistance genes, a mobility analysis was carried out. 10 000 base pairs upstream and down-

stream of the predicted resistance genes were screened for conjugative elements, IS elements, and integrons. The analysis was done for all clusters. However, if a cluster contained more than 1000 sequences, only 1000 random sequences were screened to reduce the time required for the analysis. First, GEnView v0.1.1 was used to retrieve and screen the genetic regions of 10 000 base pairs upstream and downstream of the predicted resistance genes [46]. GEnView searches for genes that are associated with mobile elements and integrons which can be used in further analysis.

To find conjugative elements, the genes found by GEnView were first translated in all six reading frames using EMBOSS Transeq v6.5.7.0 [47]. Then, HMMER v3.1b2 [48] was used to screen for conjugative elements using 124 different HMMs from MacSyfinder Conjscan v2.0 [49]. The HMMs represents *mob* genes, genes from eight different conjugating Mpf systems, and coupling proteins. The eight different Mpf systems are named B, C, F, FA, FATA, G, I, and T. The genes found by GEnView were also screened for IS elements and integrons. IS elements were searched for in the reference database ISFinder [50] using BLASTx v2.10.1 [45]. For a genetic element to be classified as an IS element >50 % coverage and >90 % amino acid identity to a known IS transposase is required. It also has to be located within 1000 base pairs from the predicted resistance gene. To find integrons Integron Finder v.1.5.1 was used [51].

4

Results

In this chapter, the results will be presented. They will then be further analyzed in the Discussion. Initially, Section 4.1 covers the results of the model creation and optimization. Section 4.2 will present the results from the large-scale screening and the taxonomy of the predicted genes. Section 4.3 will present the phylogenetic relationship between the identified genes and the known resistance genes. Last, Section 4.4 will present the phylogenetic and mobility analysis, first in general and then some interesting parts more thoroughly.

4.1 Gene models optimized for colistin resistance

From the 49 sequences of known colistin resistance genes, three different approaches to building the profile HMM were tested. Two different separated models where the difference was the placing of MCR-5 and one combined model with 49 sequences were created. The separated Profile HMMs were based on the two groups: MCR-1, MCR-2, MCR-6, and ICR-M in one group, and MCR-3, MCR-4, MCR-7, MCR-8, MCR-9, MCR-10, and ICR-B in the other group. The sensitivity and specificity as a function of the domain score for the combined model can be seen in Figure 4.1. For the separated models, the sensitivity and specificity as a function of the domain score can be found in Appendix B.

Table 4.1: Number of input sequences for each model together with sensitivity, and specificity at the optimized threshold. Here, G1 denotes MCR-1, MCR-2, MCR-6 and ICR-M and G2 denotes MCR-3, MCR-4, MCR-7, MCR-8, MCR-9, MCR-10, and ICR-B. The total number of genes predicted from the initial screening of RefSeq genomes can be seen in the right-most column.

Model	Input sequences	Sensitivity	Specificity	RefSeq
Combined	49	1.000	0.9975	12 382
G1 + MCR-5	21	1.000	1.000	12 363
G2	28	1.000	0.9994	
G1	19	1.000	1.000	12 160
G2 + MCR5	30	1.000	0.9975	

The sensitivity and specificity at the optimized threshold score for each model can be seen in Table 4.1. All models showed a sensitivity of 1.0 and a specificity very close

to 1.0 at the optimized threshold score, thus the models can successfully separate colistin resistance genes from closely related genes without a resistance mechanism. The specificity was slightly lower for the G2 part of the separated models than the G1 part, both with and without MCR-5. The combined model also showed a small decrease in specificity in comparison to the G1 part of the separated models.

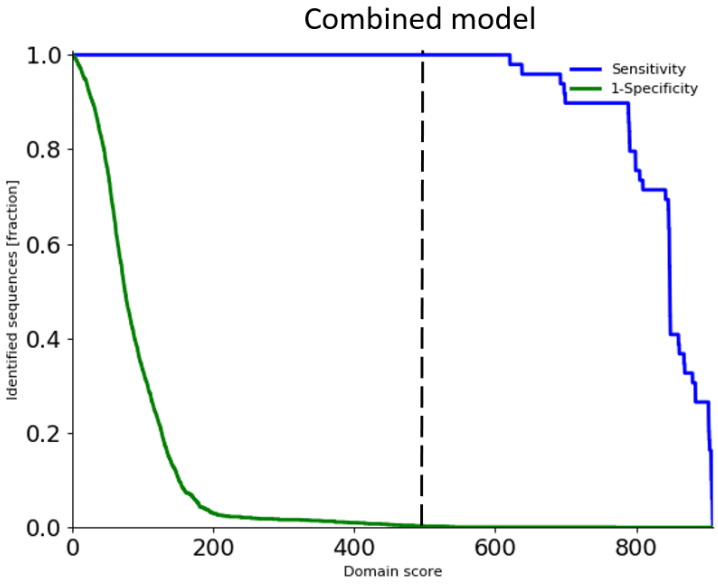

Figure 4.1: Sensitivity and 1 - specificity as a function of the domain score for the combined model with all 49 reference sequences. The sensitivity and specificity are based on the fraction of correctly classified sequences. The used threshold score of 500 is marked with a dashed line.

The initial smaller screening with 31 187 sequences from RefSeq gave almost the same sum of predicted genes from the separated models as the total from the combined model, as can be seen in the right-most column in Table 4.1. Additional results from this screening are not shown since RefSeq is a subset of GenBank used in the large-scale screening. Considering the small difference both in performance and predicted genes, together with the time requirements that come with large-scale screening with multiple models, only the combined model was used in the following large-scale screening.

4.2 Screening and taxonomic over-representation

From the screening of 1 018 179 NCBI GenBank genomes, 680 257 genes were classified as colistin resistance genes by the model. Clustering the identified genes at an amino acid identity of 70 % gave 1611 clusters. Interestingly, one of the clusters consists of 517 332 sequences which correspond to 76 % of all predicted sequences. The 680 257 predicted genes come from 593 447 unique genomes which means that a potential resistance gene could be identified in nearly 60 % of the NCBI GenBank genomes.

The taxonomic analysis showed that the phylum *Pseudomonadota* is significantly over-represented (p-value = 2.2×10^{-16}) and 99.67 % of the predicted genes are carried by bacteria from the phylum. The other phyla that were investigated (*Bacillota*, *Actinomycetota*, *Bacteroidota*) were all significantly under-represented (p-value = 2.2×10^{-16}). Among *Pseudomonadota*, the class *Gammaproteobacteria* is significantly over-represented (p-value = 2.2×10^{-16}) and the classes *Beta-*, *Alpha-*, *Epsilon-*, and *Deltaproteobacteria* are significantly under-represented (p-value = 2.2×10^{-16}). The log-transformed ratios of the over- and under-representation for the taxonomic classes can be seen in figure 4.2.

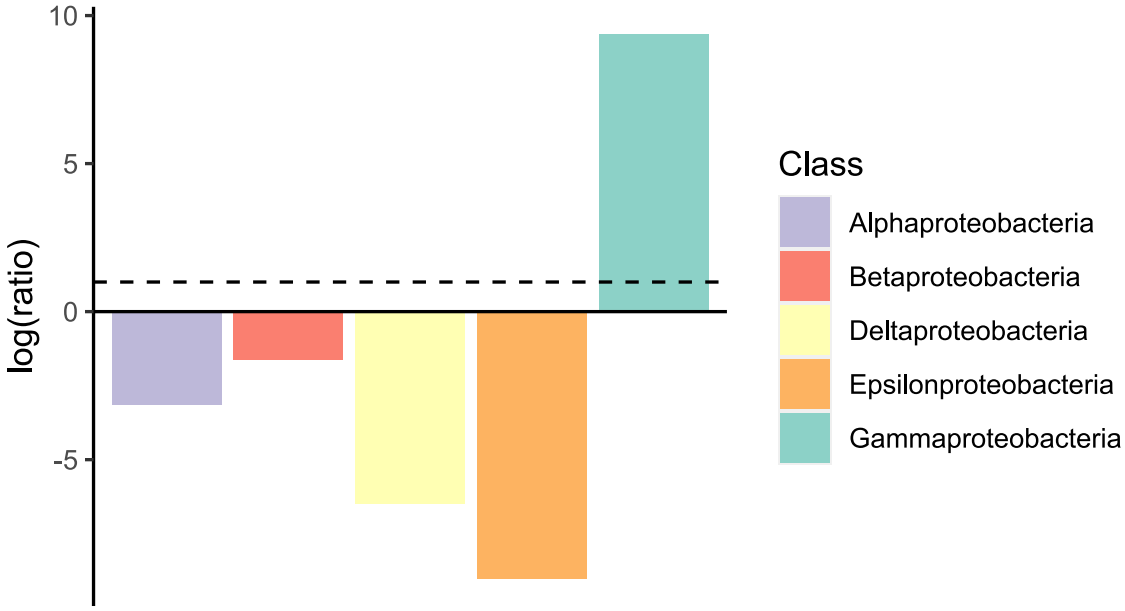

Figure 4.2: Over- and under-representation of taxonomic classes in the predicted colistin resistance genes. All results are significant with a p-value = 2.2×10^{-16}. A log(ratio) above zero indicates an over-representation and a ratio below zero indicates an under-representation.

4.3 Phylogentic relationship to known resistance genes

To investigate the phylogenetic relationship between the predicted genes and the known resistance genes, the amino acid identity of the 1611 clusters in relation to the known colistin resistance genes was determined. The clusters showing the highest amino acid identity to each of the resistance genes (closest homolog) can be seen in Table 4.2. The span of amino acid identity corresponds to the lowest and highest amino acid identity to the known resistance gene within the closest homolog cluster. The rightmost column shows the number of genes in the cluster that corresponds to the known resistance gene, assuming that $> 90\%$ amino acid identity is required.

Looking at the number of genes in each cluster that have an amino acid identity above 90 % of the known resistance gene, it can be seen that the model finds all known resistance genes except MCR-6 and MCR-7. It can also be seen that the model finds MCR-1 and MCR-9 to a greater extent than the other resistance genes. Additionally, the biggest cluster containing 517 332 (out of 680 257) sequences showed the highest amino acid identity to the chromosomal gene EptA and 156 426 sequences in the cluster had an amino acid identity >90 % to the gene.

Table 4.2: Closest homolog cluster to each resistance gene. The table shows the cluster name, cluster size, the span of amino acid identity to the known resistance genes, and the number of genes with an amino acid identity > 90%.

Known resistance gene	**Closest homolog**	**# genes in cluster**	**%AA identity to known resistance gene**	**# genes in cluster with %AA identity >90%**
MCR-1	cluster_23	4566	97 - 100 %	4563
MCR-2	cluster_23	4566	97 - 100 %	3
MCR-3	cluster_8	882	66 - 100 %	391
MCR-4	cluster_42	159	76 - 100 %	54
MCR-5	cluster_81	114	100 %	114
MCR-6	cluster_1478	1	63 %	0
MCR-7	cluster_9	831	64 - 80 %	0
MCR-8	cluster_157	77	96 - 100 %	77
MCR-9	cluster_7	4025	79 - 100 %	3883
MCR-10	cluster_7	4025	79 - 100 %	117
ICR-Mc	cluster_195	200	97 - 100 %	200
ICR-Mo	cluster_87	56	95 - 100 %	56
ICR-B	cluster_7	4025	79 - 100 %	14

4.4 Phylogentic and mobility analysis

A phylogenetic tree, visualizing the relationships between the predicted genes can be seen in Figure 4.3. The tree represents the centroid sequences of the 1611 clusters (70 % clustering of the predicted genes) and each leaf is colored based on the dominating taxonomic class within the cluster. The clusters which showed the highest similarity to genes used to build the model, as well as EptA, are marked with the name of the corresponding gene. Additionally, conjugative- and IS elements are annotated binary, where a black annotation means that an element could be found in the cluster. Smaller clades in the tree where the known resistance genes are located are marked with A1 - A5. Clusters, where both conjugative elements and pathogens are found, are marked with black circles. These clusters, as well as clade A1-A5, will be described more thoroughly later in this section.

As shown in Figure 4.3, the tree is separated into three major clades. The clusters with the highest similarity to the known resistance genes separate into three different clades in accordance with the two distinct groups of resistance genes. MCR-1, MCR-2, MCR-6, and ICR-M are located in the same clade while MCR-3, MCR-4, MCR-7, MCR-8, MCR-9, MCR-10, and MCR-B as well as EptA are in another clade. MCR-5 is located in a separate clade, which reinforces the uncertainty of grouping MCR-5. The known resistance genes are distributed over the tree which indicates a high diversity of the predicted colistin resistance genes. Additionally, it can be seen that sequences belonging to the same class cluster together, given the distinction between colors in the tree.

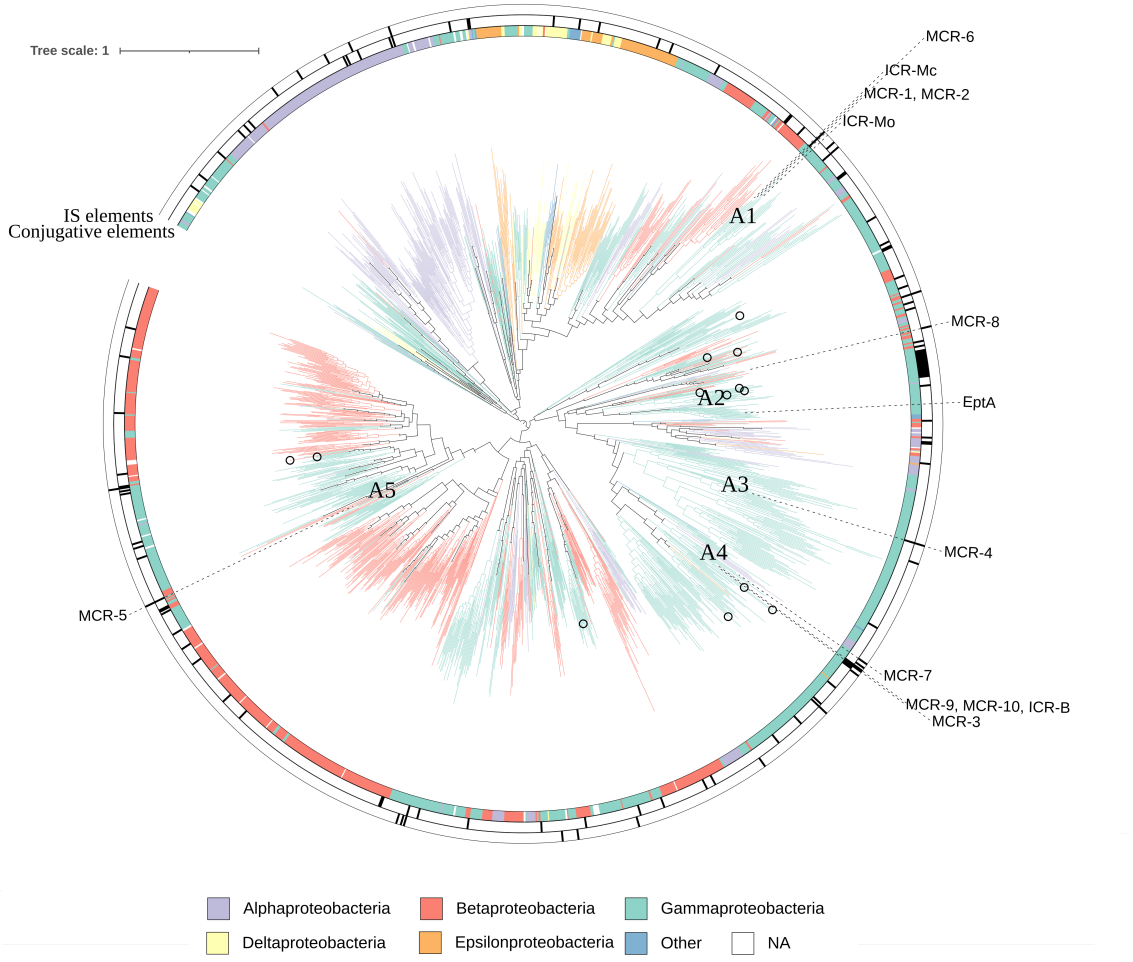

Figure 4.3: Phylogentic tree of centroids from each cluster of predicted resistance genes. The tree is annotated based on taxonomic class as well as the presence of mobile elements. The clusters with the highest similarity to the known resistance genes are marked, as well as clusters with conjugative elements and pathogens. A1-A5 represents clades where the known resistance genes are located. The black circles represent clusters comprising both conjugative elements and pathogens.

In order to identify clusters that potentially consist of mobile resistance genes, 10 000 base pairs downstream and upstream of each predicted gene were screened for conjugative elements, IS elements, and integrons. The conjugative elements searched for were *mob* genes, genes from eight different Mpf components (B, C, F, FA, FATA, G, I, and T) as well as coupling proteins. In total, conjugative elements could be found in 104 clusters and IS elements in 27 clusters. 12 of the clusters carry both conjugative elements and IS elements. In 2 of these 12 clusters, integrons are present as well. The conjugative elements identified were both *mob* genes, coupling proteins, and genes from Mpf systems. Remarkably, genes from Mpf_{FATA} and Mpf_{FA}, which are Mpf systems normally not found in *Pseudomonadota* [31], were found in 38 of the clusters. The other genes from Mpf system come from Mpf_{T}, Mpf_{F}, Mpf_{I}, and Mpf_{G} which are Mpf typically found in *Pseudomonadota*.

4.4.1 Clade A1

The tree structure of clade A1 can be seen in Figure 4.4. In general, sequences in this part of the tree do not have pathogenic hosts, except in the clade with the known resistance genes. In addition, *Moraxella sp.* was found in these clusters, which is expected since it is the progenitor of MCR-1, MCR-2, and MCR-6.

As expected, many different mobile elements are associated with MCR-1 and MCR-2 while no mobile elements are associated with the chromosomal resistance genes ICR-M. No mobile elements were found in the MCR-6 cluster. However, as can be seen in table 4.2, the sequence in this cluster only shares 63% amino acid identity with MCR-6. Further, mobile elements were found in seven clusters in addition to the cluster with MCR-1 and MCR-2. The hosts of these mobile elements were the non-pathogenic *Methylotenera sp.*, *Methylophliyus sp*, *Psychrobacter sp.*, and *Paraglaciecola sp.*. A trend that can be seen in these clusters is the presence of the conjugative elements from the systems Mpf_{FATA} and Mpf_{T}.

Table 4.3: Summary of mobile elements found in clade A1. t4cp denotes a coupling protein and virb4 denotes an ATPase that is a part of the conjugation system. The top three species in the cluster (or less if only one or two species) can be seen in the rightmost column. Pathogens are marked in bold.

Cluster ID	**Size**	**Conjugative elements**	**IS elements**	**Species**
cluster_1145	1	Mpf_{FATA} (1)	-	*Pseudomonadota bacterium*
cluster_516	3	Mpf_{T} (3)	-	*Methylotenera sp.*
cluster_127	9	Mpf_{FATA} (6)	-	*Methylophilus sp.*
cluster_23 (MCR-1, MCR-2)	4566	Mpf_{FA} (1) Mpf_{FATA} (1) Mpf_{F} (5) Mpf_{I} (4) Mpf_{T} (10) MOBH (3) MOBP1 (368) t4cp (66) virb4 (2)	ISApl1_IS30 (137) ISSpu2_IS630 (2) IS1294_IS91 (2) IS1S_IS1 (2) IS1X2_IS1 (2) ISKpn26_IS5_IS5 (1)	***Escherichia coli*** ***Salmonella enterica*** ***Klebsiella pneumoniae***
cluster_346	7	-	ISPssp1_IS3_IS3 (2)	*Psychrobacter sanguinis* *Psychrobacter sp.*
cluster_151	132	MOBF (1) Mpf_{T} (2)	ISPssp3_IS1595_IS1016 (1)	*Psychrobacter immobilis* *Psychrobacter sp.*
cluster_1385	1	Mpf_{T} (1)	-	*Paraglaciecola hydrolytica*
cluster_282	1	Mpf_{T} (1)	-	*Paraglaciecola sp.*

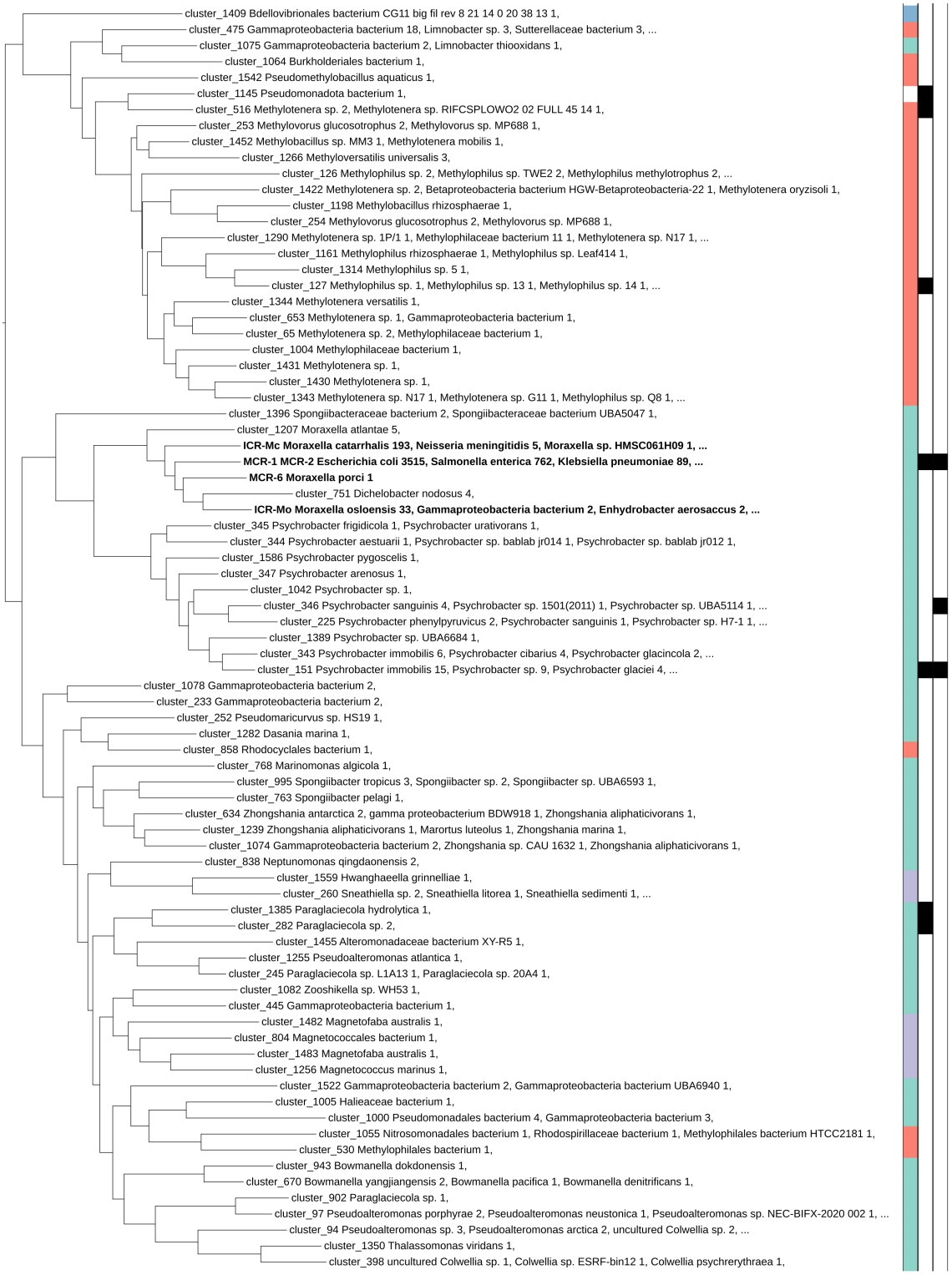

Figure 4.4: Structure of clade A1. The three dominating species of each cluster can be seen at the tips. The clusters corresponding to the known resistance genes are marked in bold. The same annotations (taxonomic class, conjugative elements, IS elements) as in the big tree can be seen to the right.

4.4.2 Clade A2

The tree structure of clade A2 can be seen in Figure 4.5. In order to get a better overview, this part stretches from MCR-8 to EptA. In general, this part of the tree does not contain sequences from pathogens, except in the MCR-8 and EptA clusters. As expected, *Klebsiella pneumoniae* is found in the MCR-8 cluster since it is the believed progenitor. Additionally, one cluster (cluster_24) close to the EptA cluster contains many sequences from the pathogen *Klebsiella pneumoniae*.

Clade A2 also includes a remarkable clade of 18 clusters where conjugative elements from Mpf_{FATA} and Mpf_{FA} could be found in all clusters. As mentioned, these elements are normally not found in the phylum *Pseudomonadota*. The mobile elements found in the other clusters are summarized in Table 4.4. In the cluster corresponding to MCR-8, two IS elements but no conjugative elements could be found. In the other clusters, elements from Mpf_T could be found in the hosts *Saezia sanguinis*, *Pelistega europaea*, and *Pectobacterium sp.*.

Table 4.4: Summary of mobile elements found in clade A2. The top three species in the cluster (or less if only one or two species) can be seen in the rightmost column. Pathogens are marked in bold.

Cluster ID	Size	Conjugative elements	IS elements	Species
cluster_157 (MCR-8)	77	-	ISEcl1_IS3_IS2 (11) ISKpn26_IS5_IS5 (9)	***Klebsiella pneumoniae*** ***Raoultella ornithinolytica*** ***Klebsiella quasipneumoniae***
cluster_1557	1	Mpf_T (1)	-	*Saezia sanguinis*
cluster_513	1	Mpf_T (1)	-	*Pelistega europaea*
cluster_109	302	Mpf_T (1)	-	*Pectobacterium brasiliense* *Pectobacterium versatile* *Pectobacterium carotovorum*

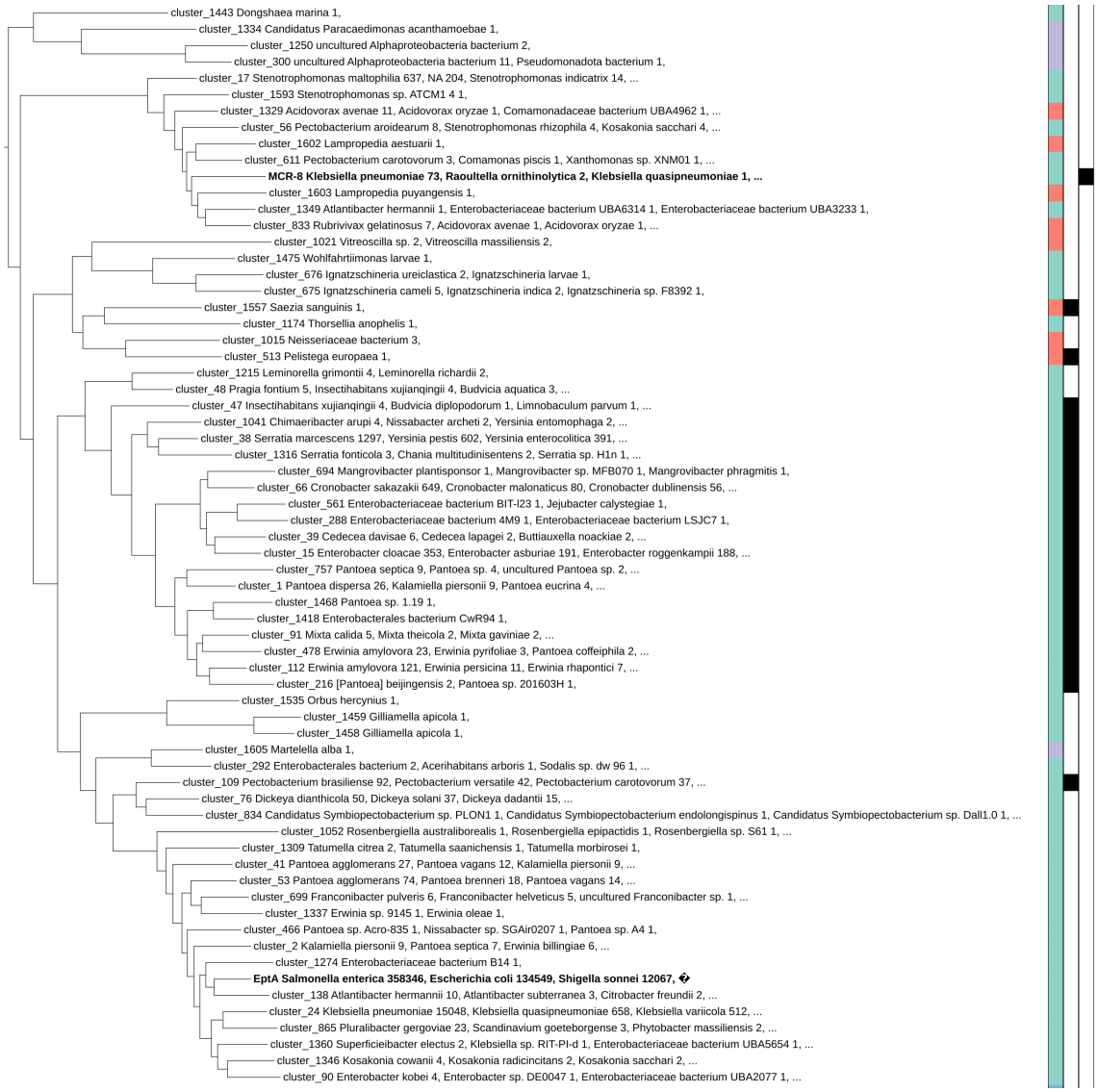

Figure 4.5: Structure of clade A2. The three dominating species of each cluster can be seen at the tips. The clusters corresponding to the known resistance genes are marked in bold. The same annotations (taxonomic class, conjugative elements, IS elements) as in the big tree can be seen to the right.

4.4.3 Clade A3

The structure of clade A3 can be seen in Figure 4.6. This clade nearly exclusively consists of sequences from the non-pathogenic genus *Shewanella*, except in the MCR-4 cluster where *Salmonella enterica* could be found as well. This correlates well with the literature since *Shewanella* is believed to be the progenitor of MCR-4. The mobile elements found in A3 are summarized in Table 4.5. In the MCR-4 cluster, multiple *mob* genes and IS elements could be found. IS elements could also be found in one additional cluster with *Shewanella* as host.

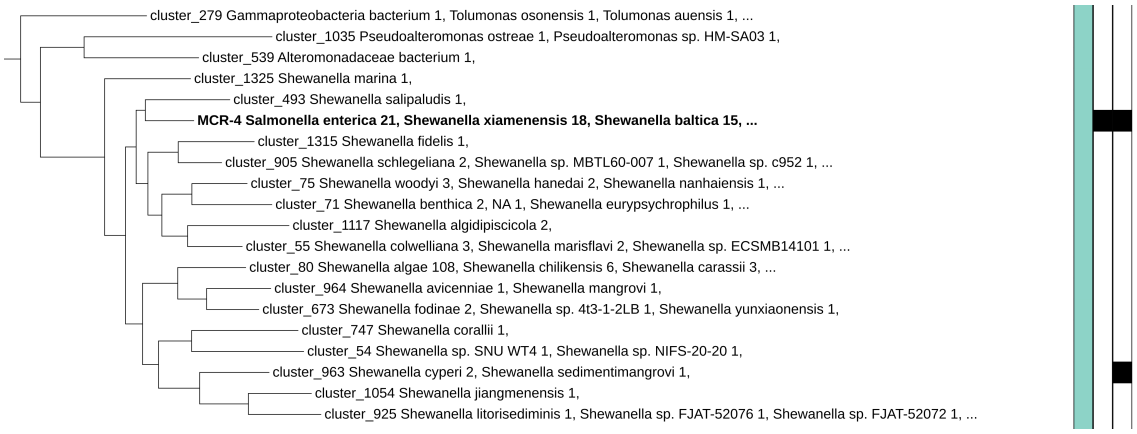

Figure 4.6: Structure of clade A3. The three dominating species of each cluster can be seen at the tips. The clusters corresponding to the known resistance genes are marked in bold. The same annotations (taxonomic class, conjugative elements, IS elements) as in the big tree can be seen to the right.

Table 4.5: Summary of mobile elements found in clade A3. The top three species in the cluster (or less if only one or two species) can be seen in the rightmost column. Pathogens are marked in bold.

Cluster ID	Size	Conjugative elements	IS elements	Species
cluster_42 (MCR-4)	151	MOBH (1) MOBP1 (1) MOBQ (36)	IS10R_IS4_IS10 (2) ISKpn26_IS5_IS5 (3) IS26_IS6_unknown (9) IS5_IS5_IS5 (1) ISAba16_IS66 (1) ISSod4_IS256 (1) ISSba11_IS21 (1)	***Salmonella enterica*** *Shewanella xiamenensis* *Shewanella baltica*
cluster_963	3	-	ISSham1_IS200/IS605_IS200 (1)	*Shewanella cyperi* *Shewanella sedimentimangrovi*

4.4.4 Clade A4

The structure of clade A4 can be seen in Figure 4.7. In contrast to the other clades, many pathogens are found in the clade. In the region of known resistance genes, pathogens from the genus *Aeromonas* as well as the pathogens *Salmonella enterica*, *Enterobacter hormaechei*, and *Escherichia coli* can be observed. The presence of *Aeromonas* species is expected since MCR-3 and MCR-7 are believed to originate from this genus. Additionally, the pathogen *Plesiomonas shigelloides* can be found in several clusters in clade A4.

The mobile elements in clade A4 are summarized in Table 4.6. Many different mobile elements, both conjugative elements and IS elements, are associated with the MCR-3 cluster and MCR-9/MCR-10/ICR-B cluster. In the MCR-7 cluster only IS elements could be found. However, it should be noted that the model only could find sequences with a maximum of 80 % amino acid identity to MCR-7, which can be seen in Table 4.2. In the other clusters IS elements, elements from Mpf_F, *mob* genes, and coupling proteins could be found. In cluster_998 and cluster_384, conjugative elements in pathogens could be found. This is an indication of novel MCR genes in pathogens with the ability to transfer horizontally to other bacteria.

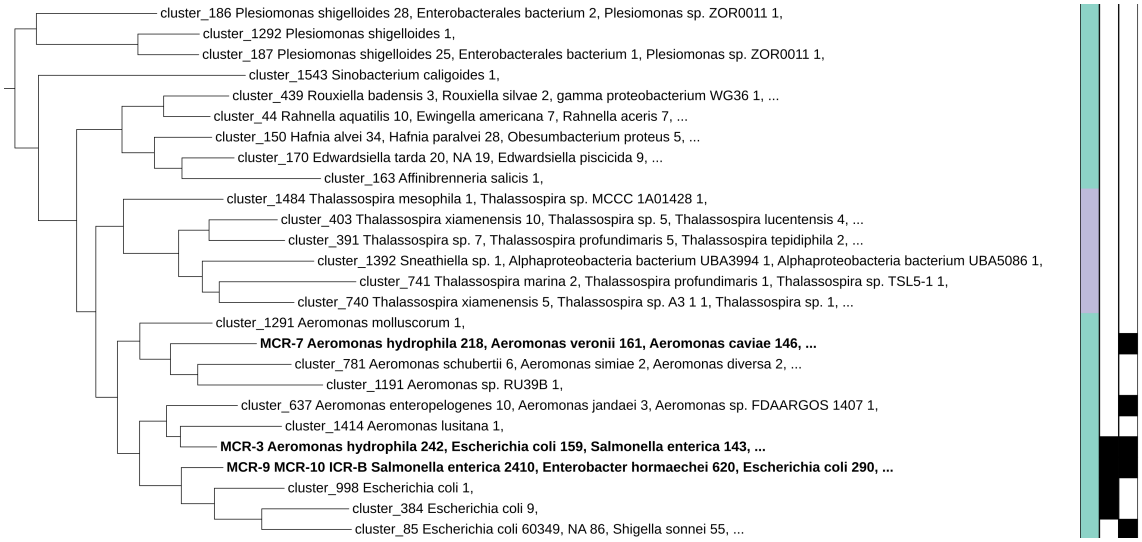

Figure 4.7: Structure of clade A4. The three dominating species of each cluster can be seen at the tips. The clusters corresponding to the known resistance genes are marked in bold. The same annotations (taxonomic class, conjugative elements, IS elements) as in the big tree can be seen to the right.

Table 4.6: Summary of mobile elements found in clade A4. t4cp denotes a coupling protein and virb4 denotes an ATPase that is a part of the conjugation system. The top three species (or less if only one or two species) can be seen in the rightmost column. Pathogens are marked in bold.

Cluster ID	**Size**	**Conjugative elements**	**IS elements**	**Species**
cluster_9 (MCR-7)	826	-	IS5D_IS5_IS5 (1) ISAeme10_IS1595_ISPna2 (2) ISAhy1_IS1595_ISPna2 (1)	***Aeromonas hydrophila Aeromonas veronii Aeromonas caviae***

cluster_637	14	-	ISApu2_IS4_ISH8 (1)	***Aeromonas enteropelogenes*** *Aeromonas jandaei* *Aeromonas sp.*
cluster_8 (MCR-3)		Mpf_F (2) Mpf_F (26) MOBC (1) MOBF (21) MOBP1 (3) MOBQ (2) t4cp (2) virb4 (21)	ISKpn40_IS3_IS150 (88) IS15DI_IS6 (15) IS15_IS6(4) IS4321_IS110_IS1111 (1) IS4321R_IS110_IS1111 (1) IS15DIV_IS6 (1) ISSen9_IS1 (1) IS5_IS5_IS5 (5) IS5D_IS5_IS5 (2) ISApu2_IS4_ISH8 (1) ISKpn26_IS5_IS5 (27) ISAs17_IS3_IS2 (9) ISAeca5_IS1595_IS1595 (1) ISKpn10_IS3_IS407 (1) ISAhy1_IS1595_ISPna2 (1) ISKpn15_IS66_ISBst12 (1) ISAs18_IS4_IS10 (2) ISAs7_IS3_IS51 (4) ISAs6_IS3_IS3 (1)	***Aeromonas hydrophila*** ***Escherichia coli*** ***Salmonella enterica***
cluster_7 (MCR-9, MCR-10, ICR-B)	1567	Mpf_F (1) Mpf_T (11) MOBF (3)	IS903B_IS5_IS903 (194) IS1B_IS1 (66) IS26_IS6 (81) IS1R_IS1 (4) ISEsp1_IS66 (1) ISEc36_IS3_IS2 (13) IS1A_IS1 (1) IS903_IS5_IS903 (1) ISKox1_IS66 (2)	***Salmonella enterica*** *Enterobacter hormaechei* ***Escherichia coli***
cluster_998	1	Mpf_F (1) t4cp (1)	-	***Escherichia coli***
cluster_384	9	MOBH (3)	-	***Escherichia coli***
cluster_85	60 502	-	IS1203_IS3_IS51(6) IS629_IS3_IS51(1) ISEc8_IS66 (1)	***Escherichia coli*** ***Shigella sonnei***

4.4.5 Clade A5

The structure of clade A5 can be seen in Figure 4.8. Pathogens can be found in the MCR-5 cluster as well as some clusters with species from the genus *Legionella*. However, the pathogenicity of the other clusters is difficult to verify since many bacteria

only are annotated based on class (for example *Gammaproteobacteria bacterium*). The mobile elements found in clade A5 can be seen in Table 4.7. Multiple *mob* genes and IS elements, as well as a coupling protein could be found in the MCR-5 cluster. Additionally, elements from Mpf_T, coupling protein, and the ATPase virb4 could be found in two other clusters.

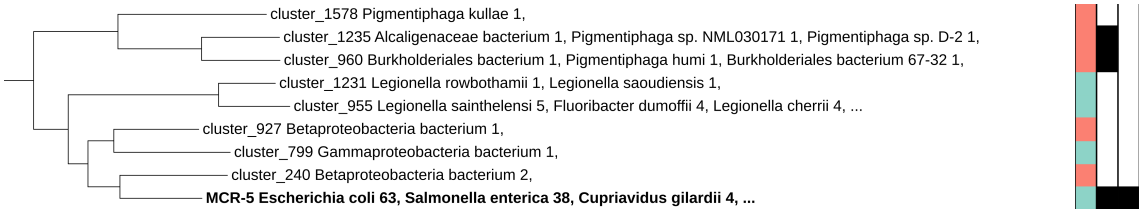

Figure 4.8: Structure of clade A5. The three dominating species of each cluster can be seen at the tips. The clusters corresponding to the known resistance genes are marked in bold. The same annotations (taxonomic class, conjugative elements, IS elements) as in the big tree can be seen to the right.

Table 4.7: Summary of mobile elements found in clade A5. t4cp denotes a coupling protein and virb4 denotes an ATPase that is a part of the conjugation system. The top three species (or less if only one or two species) can be seen in the rightmost column. Pathogens are marked in bold.

Cluster ID	Size	Conjugative elements	IS elements	Species
cluster_1235	3	Mpf_T (3) t4cp (3) virb4 (3)	-	*Alcaligenaceae bacterium* *Pigmentiphaga sp.* *Pigmentiphaga sp.*
cluster_960	3	Mpf_T (3) t4cp (3) virb4 (3)	-	*Burkholderiales bacterium* *Pigmentiphaga humi*
cluster_81 (MCR-5)	123	MOBF (5) MOBP1 (4) MOBQ (2) MOBV (5) t4cp (1)	IS26_IS6 (1) IS15_IS6 (1) ISPa96_IS5_IS5 (1)	***Escherichia coli*** ***Salmonella enterica*** ***Cupriavidus gilardii***

4.4.6 Horizontal gene transfer and pathogens

As a final summary of the results, the clusters comprising both conjugative elements and pathogens can be seen in 4.8. In total, 13 clusters (in addition to clusters corresponding to known resistance genes) could be found. The leaves of the clusters in 4.8 are marked with a black circle in the phylogenetic tree in Figure 4.3. Here, it should be noted that is not possible to say if the conjugative elements actually are located in pathogens or other non-pathogens in the same cluster, since this is a

limitation in the mobility analysis. However, these results still point out potential new MCR genes that are located in pathogens or risk transferring to pathogens in the future.

Additionally, cluster_998 and cluster_384 (which also can be seen in clade A4) consist solely of pathogens, thus here it is safe to say that potential new MCR genes that can transfer horizontally are found in pathogens. Moreover, four of the clusters, cluster_38, cluster_66, cluster_15, and cluster_1, are found in the clade with Mpf_{FATA} and Mpf_{FA} mentioned in the section about clade A2.

Table 4.8: Clusters comprising both pathogens and conjugative elements. The pathogens found in the clusters can be seen in the rightmost column.

Cluster name	**Size**	**Conjugative elements**	**Species**
cluster_508	5	MOBF (3)	*Vibrio cholerae*
cluster_4	4401	Mpf_{FA} (1) Mpf_{FATA} (1)	*Neisseria meningitidis* *Neisseria gonorrhoeae*
cluster_383	35	Mpf_{FATA} (1)	*Aggregatibacter actinomycetemcomitans*
cluster_38	3257	Mpf_{FA} (865) Mpf_{FATA} (865)	*Yersinia pestis* *Yersinia pseudotuberculosis* *Shigella boydii* *Yersinia enterocolitica* *Yersinia kristensenii*
cluster_66	841	Mpf_{FA} (823) Mpf_{FATA} (823)	*Cronobacter sakazakii* *Cronobacter turicensis* *Cronobacter malonaticus* *Cronobacter dublinensis* *Cronobacter universalis* *Cronobacter condimenti*
cluster_15	1646	Mpf_{FA} (987) Mpf_{FATA} (985)	*Enterobacter cloacae* *Escherichia coli* *Klebsiella pneumoniae* *Salmonella enterica*
cluster_1	124	Mpf_{FA} (987) Mpf_{FATA} (985)	*Enterobacter cancerogenus*
cluster_998	1	Mpf_{G} (1) t4cp (1)	*Escherichia coli*
cluster_384	9	MOBH (3)	*Escherichia coli*
cluster_11	241	Mpf_{FATA} (2) MOBF (1)	*Vibrio cholerae* *Vibrio parahaemolyticus*
cluster_246	226	Mpf_{T} (1)	*Pseudomonas fluorescens*
cluster_18	1329	Mpf_{T} (1)	*Pseudomonas aeruginosa*
cluster_215	116	Mpf_{T} (114) t4cp (115) virb4 (110)	*Burkholderia cenocepacia*

5

Discussion

In this thesis, a profile HMM optimized for colistin resistance genes was created with fARGene. The model performed well in terms of sensitivity and specificity, meaning that it can successfully separate resistance genes from closely related genes without a resistance mechanism. Large-scale genomic data were screened for novel resistance genes, which identified 680 257 genes, over-represented in *Pseudomonadota*. From these genes, a large cluster of 517 332 genes could be connected to the chromosomal gene EptA. Moreover, a mobility analysis was performed to evaluate the mobility of colistin resistance genes and investigate the risk of HGT. From the mobility analysis, it could clearly be seen that there is an ongoing spread of MCR genes and that there are signs of novel MCR genes detected in the screening. In this chapter, a more in-depth discussion about the results and how they can be interpreted will follow.

5.1 Model creation and optimization

Initially, three different models were created. However, despite the two phylogenetically diverse groups of resistance genes, the combined and separated model performed with similar sensitivity and specificity, which can be seen in Table 4.1 as well as in Figure 4.1 and Appendix B. Probably, this can be explained by the active site that is conserved throughout all MCR genes and the property of profile HMMs. Since the objective of profile HMMs is to capture the conserved parts of a group of proteins, it is likely that the active site is captured by the model which describes why the models perform equally, despite the phylogenetic difference between the known resistance genes.

Further, the sensitivity (ability to find all ARGs) can be analyzed by Table 4.2 which shows the closest homolog to the known resistance genes and how many genes with an amino acid identity >90 %. Here it could be noted that the model correctly finds all resistance genes except MCR-6 and MCR-7 and that MCR-1 and MCR-9 are found to a greater extent than the other genes. This could potentially be explained by the difference in the number of input genes, giving a model more sensitivity to certain genes. The model was built with 14 different variants of MCR-1 but only one MCR-9. Additionally, six ICR-B sequences were used in the model which is very similar to MCR-9, which could explain why the model finds MCR-9 to a greater extent. In comparison to this, only one sequence from MCR-6 and MCR-7 was used in the model. However, by arguing this way, the model would also identify MCR-3 to a greater extent (13 variants of MCR-3 in the input data) and potentially missing

MCR-8 with only one input sequence and which also is relatively phylogenetically far away from the other resistance genes. More likely, this can be explained by the prevalence of the different resistance genes in the database, rather than a lack of model performance. This theory is also supported by the spread of the different MCR genes globally [21]. It has been shown that MCR-1 and MCR-9 are disseminated to a greater extent than the other MCR genes and that MCR-6 only has been found once as well as MCR-7 which is only identified in a few places in the world.

5.2 Analysis of phylogeny and mobility

One interesting result is the large amount of identified genes. It could be seen that the model found a gene in nearly 60 % of the screened genomes. However, it is safe to say that all of these genes are not of clinical relevance and will not give a resistance phenotype, otherwise colistin resistance would be observed to a much larger extent. As mentioned in the results, one dominant cluster with 517 332 (76 % of the predicted genes) could be identified. This cluster could be connected to the chromosomal resistance gene EptA and 156 426 genes in the cluster showed an amino acid identity >90 % to the gene. Many of the genes in this cluster have *Salmonella enterica* and *Escherichia coli* from the phylum *Pseudomonadota* as hosts, which correlates well with the literature about EptA. As mentioned in the theory, modification of the outer membrane is a common defense mechanism of gram-negative bacteria to protect from antibiotics as well as other stresses in the surrounding environment, so it is expected to find these EptA-like genes in many of the genomes. However, it should be noted that finding these many genes is not a sign of a lack of performance of the model. EptA also has the conserved active site and we want the model to find these genes to not hamper in sensitivity to find the clinically relevant MCR genes with similar conserved sites.

However, despite that these genes most probably do not give resistance phenotype due to low expression, this is still alarming. The mobility analysis reveals several mobile elements distributed in the phylogenetic tree, indicating that mobilization of these EptA-like chromosomal genes can be a future problem. Additionally, MCR-9 has been identified in a colistin-susceptible strain (strain without colistin resistance phenotype) and MCR-10 was first identified from colistin-susceptible *Enterobacter* strains [12], which indicates that the resistance can develop silently and cause resistance phenotype first when exposed to colistin or other outer stimuli.

Another interesting result is the strong over-representation of hosts from the phylum *Pseudomonadota* and the class *Gammaproteobacteria*, which is represented in Figure 4.2. One possible explanation for this is that all sequences used to build the model come from *Gammaproteobacteria* and that the model, therefore, is more sensitive for detecting sequences from *Gammaproteobacteria*. However, it is also likely that the colistin resistance mechanism is more prevalent in *Gammaprotebacteria* due to some biological reason and that this explains the over-representation both in the sequences used to build the model and the predicted genes. This could for example be membrane properties that are unique for *Gammaprotebacteria* or that *Gammaprotebacteria* are more prevalent in some environments with high selection pressure for

colistin resistance. Here, a further literature study about the difference in outer membrane properties of different bacteria within the phylum *Pseudomonadota* as well as additional studies including the environmental factor are needed to better understand these results.

Moreover, it can be seen that the phylogeny of the tree corresponds well with the literature about the MCR genes. MCR-1, MCR-2, and MCR-6 are found in one part of the tree while MCR-3, MCR-4, MCR-7, MCR-8, MCR-9, and MCR-10 are found in another part. Here, MCR-3 and MCR-7, together with MCR-9 and MCR-10 form a tighter group, and MCR-4 and MCR-8 are more diverse, as expected from the literature. In addition, MCR-5 is distant from the other genes, which also correlates well with the uncertainty of grouping this gene.

In addition, the species found in the clusters with the highest similarity to the known MCR genes confirm what is known from the literature. MCR-1, MCR-2, and MCR-6 are found in *Moraxella* species. MCR-3 and MCR-7 are found in *Aeromonas* and many of the clusters in their clade (A4) include genes from this genus. MCR-9 and MCR-10 are found in *Buttiauxella* species from where they are believed to originate. MCR-4 is found in its believed progenitors *Shewanella* and almost all clusters in this clade (A3) comprises genes from this genus. MCR-8 is found in *Klebsiella pneumoniae*, where it was first detected, and in neighboring clusters, its believed progenitor *Stenotrophomonas* can be found. The origin of MCR-5 is not known, but it has been proven that closely related sequences to MCR-5 have been found in *Legionella* and *Burkholderiales* [27], which also is found in clade A5. Moreover, multiple sequences from *Cupriavidus* are found in the MCR-5 cluster, which indicates that MCR-5 might originate from these species. However, MCR-5 has been identified in *Cupriavidus gilardii* before, but from this no conclusion about the origin of MCR-5 could be made [28].

As pointed out in the results, one unexpected result was the presence of Mpf_{FATA} and Mpf_{FA} since they are not typically found in *Pseudomonadota*. It is difficult to tell why these elements were found, but they could potentially be false positive results. In many of the clusters comprising both pathogens and conjugative elements (Table 4.8), Mpf_{FATA} and Mpf_{FA} are present, thus it will be important to look more into this in the future. In general, it would be interesting to do a more thorough mobility analysis to get a more comprehensive understanding of the mobility of colistin resistance genes. Here, it would be interesting to expand the search to more than 10 000 base pairs upstream and downstream since it is likely that additional mobile elements will be found. For example, a plasmid can be much longer than 10 000 base pairs.

The mobility analysis clearly shows that there is an ongoing spread of colistin resistance genes and that many mobile elements are associated with the genes. Looking into the clades (A1-A5) it could be seen that conjugative elements were found close to most of the known resistance genes, indicating that there are novel MCR genes with the ability to transfer horizontally, both in non-pathogenic and pathogenic hosts. Additionally, multiple clusters with IS elements could be found. This is a sign of ongoing mobilization of colistin resistance genes, that risks ending up on conjugative element and possess the ability to transfer horizontally in the future.

Particularly interesting are the 13 clusters that comprise both conjugative elements (that are required for HGT) and pathogenic hosts. Potentially, these clusters correspond to novel MCR genes with the ability to transfer between bacteria and already are present in pathogenic hosts or risk ending up in pathogens in the future. Interestingly, most of these clusters are located in the big clade with MCR-3, MCR-4, MCR-7, MCR-8, MCR-9, and MCR-10, and some are located in the MCR-5 clade, while no such clusters could be found in the neighborhood of MCR-1, MCR-2, and MCR-6. Here, it would be interesting to see if the reservoir for MCR-1, MCR-2, and MCR-6 type of resistance genes for example are more connected to non-pathogenic environments, in contrast to the other MCR genes. To evaluate this, an analysis of the isolation source of the samples, or developing a model for metagenomes would be a suitable next step. In addition to this, it would be interesting to verify if the conjugative elements are found in pathogens, or non-pathogens in the same cluster.

6

Conclusion and Future Outlook

This project aimed to perform a characterization of colistin resistance genes with the main goal to identify potential novel colistin resistance genes and evaluate them in terms of phylogeny, taxonomy, and mobility. A profile HMM was optimized using fARGene and through a large-scale screening, potential colistin resistance genes could be identified and analyzed. This project could identify several clusters of potential novel MCR genes and pointed out 13 clusters of particular interest, with genes possessing the ability to transfer between bacteria that at the same time are associated with pathogens. Identifying novel resistance genes, both already mobilized and future threats is of great importance to understanding and preventing the spread of antibiotic resistance. This knowledge can for example be used in the clinic to detect resistant infections at an early stage or in the development of new effective treatments.

It should however be noted, that the potential novel MCR genes found in this thesis are computational predictions. To confirm the resistance to colistin this needs to be evaluated in a wet lab. In addition to this, another important step in the future is to investigate which type of environments harbor these types of genes. This enables us to further understand colistin resistance and can answer questions such as how the genes transfer from the environment to pathogens and which actions that are necessary to prevent further spread.

Bibliography

[1] World Health Organization, *Antibiotic resistance*, Feb. 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance.

[2] C. J. L. Murray, K. S. Ikuta, F. Sharara, *et al.*, "Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis.," *The Lancet*, vol. 399, no. 10325, pp. 629–655, 2022, ISSN: 0140-6736. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edselp&AN=S0140673621027240&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[3] J. Li, R. L. Nation, J. D. Turnidge, *et al.*, "Colistin: the re-emerging antibiotic for multidrug-resistant Gram-negative bacterial infections," *Lancet Infectious Diseases*, vol. 6, no. 9, pp. 589–601, Sep. 2006. DOI: 10.1016/S1473-3099(06)70580-1.

[4] M. E. Falagas, S. K. Kasiakou, and L. D. Saravolatz, "Colistin: The Revival of Polymyxins for the Management of Multidrug-Resistant Gram-Negative Bacterial Infections," *Clinical Infectious Diseases*, vol. 40, no. 9, pp. 1333–1341, May 2005. DOI: 10.1086/429323.

[5] Y. Y. Liu, Y. Wang, T. R. Walsh, *et al.*, "Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study," *The Lancet Infectious Diseases*, vol. 16, no. 2, pp. 161–168, Feb. 2016. DOI: 10.1016/S1473-3099(15)00424-7.

[6] C. Wang, Y. Feng, L. Liu, L. Wei, M. Kang, and Z. Zong, "Identification of novel mobile colistin resistance gene mcr-10," *Emerging Microbes and Infections*, vol. 9, no. 1, pp. 508–516, Jan. 2020. DOI: 10.1080/22221751.2020.1732231.

[7] D. G. Larsson and C. F. Flach, "Antibiotic resistance in the environment," *Nature Reviews Microbiology*, vol. 20, no. 5, pp. 257–269, May 2022. DOI: 10.1038/s41579-021-00649-x.

[8] B. Fanny, Ö. Tobias, B. Fredrik, M. Nachiket P., L. D. G. Joakim, and K. Erik, "Identification and reconstruction of novel antibiotic resistance genes from metagenomes.," *Microbiome*, vol. 7, no. 1, pp. 1–14, 2019, ISSN: 2049-2618. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edsdoj&AN=edsdoj.766b724af9364dd683b3f3b846b0d214&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[9] B. D. Needham and M. S. Trent, "Fortifying the barrier: The impact of lipid a remodelling on bacterial pathogenesis.," *Nature Reviews Microbiology*, vol. 11, no. 7, p. 467, 2013, ISSN: 1740-1526. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edsgao&AN=edsgcl.335627034&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[10] T. Velkov, P. E. Thompson, R. L. Nation, and J. Li, "Structureactivity relationships of polymyxin antibiotics," *Journal of Medicinal Chemistry*, vol. 53, no. 5, pp. 1898–1916, 2010, PMID: 19874036. DOI: 10.1021/jm900999h. eprint: https://doi.org/10.1021/jm900999h. [Online]. Available: https://doi.org/10.1021/jm900999h.

[11] A. Samantha and A. Vrielink, "Lipid a phosphoethanolamine transferase: Regulation, structure and immune response," *Journal of Molecular Biology*, vol. 432, pp. 5184–5196, 18 Aug. 2020, ISSN: 10898638. DOI: 10.1016/j.jmb.2020.04.022.

[12] W. Chengcheng, F. Yu, L. Lina, W. Li, K. Mei, and Z. Zhiyong, "Identification of novel mobile colistin resistance gene mcr-10.," *Emerging Microbes Infections*, vol. 9, no. 1, pp. 508–516, 2020, ISSN: 22221751. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=151677786&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[13] Y. Xu, W. Wei, S. Lei, J. Lin, S. Srinivas, and Y. Feng, "An evolutionarily conserved mechanism for intrinsic and transferable polymyxin resistance," *mBio*, vol. 9, 2 Mar. 2018, ISSN: 21507511. DOI: 10.1128/mBio.02317-17.

[14] H. Zhang, Z. Zong, S. Lei, *et al.*, "A genomic, evolutionary, and mechanistic study of mcr-5 action suggests functional unification across the mcr family of colistin resistance," *Advanced Science*, p. 1 900 034, Apr. 2019, ISSN: 2198-3844. DOI: 10.1002/advs.201900034.

[15] L. Poirel, N. Kieffer, J. F. Fernandez-Garayzabal, A. I. Vela, Y. Larpin, and P. Nordmann, "Moraxella Species as Potential Sources of MCR-Like Polymyxin Resistance Determinants.," *Antimicrobial agents and chemotherapy*, vol. 61, no. 6, Jun. 2017, ISSN: 10986596. DOI: 10.1128/AAC.00129-17. [Online]. Available: https://doi.org/10.1128/AAC.00129-17.

[16] L. Poirel, N. Kieffer, J. F. Fernandez-Garayzabal, A. I. Vela, Y. Larpin, and P. Nordmann, "MCR-2-mediated plasmid-borne polymyxin resistance most likely originates from Moraxella pluranimalium," *Journal of Antimicrobial Chemotherapy*, vol. 72, no. 10, pp. 2947–2949, Jul. 2017, ISSN: 0305-7453. DOI: 10.1093/jac/dkx225. eprint: https://academic.oup.com/jac/article-pdf/72/10/2947/21587567/dkx225.pdf. [Online]. Available: https://doi.org/10.1093/jac/dkx225.

[17] M. Doumith, G. Godbole, P. Ashton, *et al.*, "Detection of the plasmid-mediated mcr-1 gene conferring colistin resistance in human and food isolates of Salmonella enterica and Escherichia coli in England and Wales," *Journal of Antimicrobial Chemotherapy*, vol. 71, no. 8, pp. 2300–2305, Apr. 2016, ISSN: 0305-7453. DOI: 10.1093/jac/dkw093. eprint: https://academic.oup.com/jac/article-

pdf/71/8/2300/6776098/dkw093.pdf. [Online]. Available: https://doi.org/10.1093/jac/dkw093.

[18] K.-j. Zeng, Y. Doi, S. Patil, X. Huang, and G.-B. Tian, "Emergence of the plasmid-mediated mcr-1 gene in colistin-resistant enterobacter aerogenes and enterobacter cloacae," *Antimicrobial Agents and Chemotherapy*, vol. 60, no. 6, pp. 3862–3863, 2016. DOI: 10.1128/aac.00345-16. eprint: https://journals.asm.org/doi/pdf/10.1128/aac.00345-16. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/aac.00345-16.

[19] B. B. Xavier, C. Lammens, R. Ruhal, *et al.*, "Identification of a novel plasmid-mediated colistin-resistance gene, mcr-2, in escherichia coli, belgium, june 2016," *Eurosurveillance*, vol. 21, no. 27, 30280, 2016. DOI: https://doi.org/10.2807/1560-7917.ES.2016.21.27.30280. [Online]. Available: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2016.21.27.30280.

[20] M. AbuOun, E. J. Stubberfield, N. A. Duggett, *et al.*, "mcr-1 and mcr-2 (mcr-6.1) variant genes identified in Moraxella species isolated from pigs in Great Britain from 2014 to 2015," *Journal of Antimicrobial Chemotherapy*, vol. 72, no. 10, pp. 2745–2749, Aug. 2017, ISSN: 0305-7453. DOI: 10.1093/jac/dkx286. eprint: https://academic.oup.com/jac/article-pdf/72/10/2745/25182515/dkx286.pdf. [Online]. Available: https://doi.org/10.1093/jac/dkx286.

[21] Z. Ling, W. Yin, Z. Shen, Y. Wang, J. Shen, and T. R. Walsh, "Epidemiology of mobile colistin resistance genes mcr-1 to mcr-9," *Journal of Antimicrobial Chemotherapy*, vol. 75, no. 11, pp. 3087–3095, Jun. 2020, ISSN: 0305-7453. DOI: 10.1093/jac/dkaa205. eprint: https://academic.oup.com/jac/article-pdf/75/11/3087/33893062/dkaa205.pdf. [Online]. Available: https://doi.org/10.1093/jac/dkaa205.

[22] W. Wei, S. Srinivas, J. Lin, *et al.*, "Defining icr-mo, an intrinsic colistin resistance determinant from moraxella osloensis," *PLOS Genetics*, vol. 14, e1007389, 5 May 2018, ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1007389.

[23] W. Yin, H. Li, Y. Shen, *et al.*, "Novel plasmid-mediated colistin resistance gene mcr-3 in escherichia coli," *mBio*, vol. 8, no. 3, e00543–17, 2017. DOI: 10.1128/mBio.00543-17. eprint: https://journals.asm.org/doi/pdf/10.1128/mBio.00543-17. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/mBio.00543-17.

[24] Y.-Q. Yang, Y.-X. Li, C.-W. Lei, A.-Y. Zhang, and H.-N. Wang, "Novel plasmid-mediated colistin resistance gene mcr-7.1 in Klebsiella pneumoniae," *Journal of Antimicrobial Chemotherapy*, vol. 73, no. 7, pp. 1791–1795, Apr. 2018, ISSN: 0305-7453. DOI: 10.1093/jac/dky111. eprint: https://academic.oup.com/jac/article-pdf/73/7/1791/25034381/dky111.pdf. [Online]. Available: https://doi.org/10.1093/jac/dky111.

[25] A. Carattoli, L. Villa, C. Feudi, *et al.*, "Novel plasmid-mediated colistin resistance mcr-4 gene in salmonella and escherichia coli, italy 2013, spain and belgium, 2015 to 2016," *Eurosurveillance*, vol. 22, no. 31, 30589, 2017. DOI: https://doi.org/10.2807/1560-7917.ES.2017.22.31.30589. [Online].

Available: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2017.22.31.30589.

[26] X. Wang, Y. Wang, Y. Zhou, *et al.*, "Emergence of a novel mobile colistin resistance gene, mcr-8, in ndm-producing klebsiella pneumoniae," *Emerging Microbes & Infections*, vol. 7, no. 1, pp. 1–9, 2018, PMID: 29970891. DOI: 10.1038/s41426-018-0124-z. eprint: https://doi.org/10.1038/s41426-018-0124-z. [Online]. Available: https://doi.org/10.1038/s41426-018-0124-z.

[27] M. B. Khedher, S. A. Baron, T. Riziki, *et al.*, "Massive analysis of 64,628 bacterial genomes to decipher water reservoir and origin of mobile colistin resistance genes: Is there another role for these enzymes?" *Scientific Reports*, vol. 10, 2020, PMID: 29970891. DOI: 10.1038/s41598-020-63167-5. [Online]. Available: https://doi.org/10.1038/s41598-020-63167-5.

[28] M. Borowiak, J. Fischer, J. A. Hammerl, R. S. Hendriksen, I. Szabo, and B. Malorny, "Identification of a novel transposon-associated phosphoethanolamine transferase gene, mcr-5, conferring colistin resistance in d-tartrate fermenting Salmonella enterica subsp. enterica serovar Paratyphi B," *Journal of Antimicrobial Chemotherapy*, vol. 72, no. 12, pp. 3317–3324, Sep. 2017, ISSN: 0305-7453. DOI: 10.1093/jac/dkx327. eprint: https://academic.oup.com/jac/article-pdf/72/12/3317/21899367/dkx327.pdf. [Online]. Available: https://doi.org/10.1093/jac/dkx327.

[29] T. M. Henkin, L. Snyder, and J. E. Peters, *Snyder and Champness molecular genetics of bacteria.* (ASM books). Wiley Blackwell, 2020, ISBN: 1555819753.

[30] J. Lewis, D. Morgan, B. Alberts, *et al.*, *Molecular biology of the cell.* Garland Science, 2015, ISBN: 0815344643.

[31] C. Smillie, M. P. Garcillán-Barcia, M. V. Francia, E. P. C. Rocha, and F. de la Cruz, "Mobility of plasmids," *Microbiology and Molecular Biology Reviews*, vol. 74, no. 3, pp. 434–452, 2010. DOI: 10.1128/MMBR.00020-10. eprint: https://journals.asm.org/doi/pdf/10.1128/MMBR.00020-10. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/MMBR.00020-10.

[32] R. Durbin, *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* 2006, ISBN: 978-0-521-62971-3.

[33] S. R. Eddy, "Hmmer user's guide: Biological sequence analysis using profile hidden markov models," 2020.

[34] A. G. McArthur, N. Waglechner, F. Nizam, *et al.*, "The comprehensive antibiotic resistance database," *Antimicrobial Agents and Chemotherapy*, vol. 57, no. 7, pp. 3348–3357, 2013. DOI: 10.1128/aac.00419-13. eprint: https://journals.asm.org/doi/pdf/10.1128/aac.00419-13. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/aac.00419-13.

[35] D. A. Benson, M. Cavanaugh, K. Clark, *et al.*, "GenBank," *Nucleic Acids Research*, vol. 41, no. D1, pp. D36–D42, Nov. 2012, ISSN: 0305-1048. DOI: 10.1093/nar/gks1195. eprint: https://academic.oup.com/nar/article-pdf/41/D1/D36/3680750/gks1195.pdf. [Online]. Available: https://doi.org/10.1093/nar/gks1195.

[36] T. Paysan-Lafosse, M. Blum, S. Chuguransky, *et al.*, "Interpro in 2022," *Nucleic acids research*, vol. 51, no. D1, pp. D418–D427, Jan. 2023, ISSN: 0305-

1048. DOI: 10.1093/nar/gkac993. [Online]. Available: https://europepmc.org/articles/PMC9825450.

[37] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Aug. 2010, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq461. eprint: https://academic.oup.com/bioinformatics/article-pdf/26/19/2460/48857155/bioinformatics_26_19_2460.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btq461.

[38] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, ISSN: 0305-1048. DOI: 10.1093/nar/gkf436. eprint: https://academic.oup.com/nar/article-pdf/30/14/3059/9488148/gkf436.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkf436.

[39] M. N. Price, P. S. Dehal, and A. P. Arkin, "Fasttree 2 – approximately maximum-likelihood trees for large alignments," *PLOS ONE*, vol. 5, no. 3, pp. 1–10, Mar. 2010. DOI: 10.1371/journal.pone.0009490. [Online]. Available: https://doi.org/10.1371/journal.pone.0009490.

[40] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL) v4: recent updates and new developments," *Nucleic Acids Research*, vol. 47, no. W1, W256–W259, Apr. 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gkz239. eprint: https://academic.oup.com/nar/article-pdf/47/W1/W256/28879775/gkz239.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkz239.

[41] P. A. Kitts, D. M. Church, F. Thibaud-Nissen, *et al.*, "Assembly: A resource for assembled genomes at ncbi," *Nucleic Acids Research*, vol. 44, pp. D73–D80, D1 Jan. 2016, ISSN: 0305-1048. DOI: 10.1093/nar/gkv1226.

[42] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/.

[43] S. Sherrill-Mix, *Taxonomizr: Functions to work with ncbi accessions and taxonomy*, R package version 0.10.2, 2023. [Online]. Available: https://CRAN.R-project.org/package=taxonomizr.

[44] J. Bengtsson-Palme, M. Hartmann, K. M. Eriksson, *et al.*, "Metaxa2: Improved identification and taxonomic classification of small and large subunit rrna in metagenomic data," *Molecular Ecology Resources*, vol. 15, no. 6, pp. 1403–1414, 2015. DOI: https://doi.org/10.1111/1755-0998.12399. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12399. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12399.

[45] 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, ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283605803602.

[46] S. Ebmeyer, R. D. Coertze, F. Berglund, E. Kristiansson, and D. G. J. Larsson, "GEnView: a gene-centric, phylogeny-based comparative genomics pipeline for

bacterial genomes and plasmids," *Bioinformatics*, vol. 38, no. 6, pp. 1727–1728, Dec. 2021, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab855. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/6/1727/49008775/btab855.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btab855.

[47] F. Madeira, Y. m. Park, J. Lee, *et al.*, "The EMBL-EBI search and sequence analysis tools APIs in 2019," *Nucleic Acids Research*, vol. 47, no. W1, W636–W641, Apr. 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gkz268. eprint: https://academic.oup.com/nar/article-pdf/47/W1/W636/28879917/gkz268.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkz268.

[48] S. R. Eddy, "Accelerated profile hmm searches," *PLOS Computational Biology*, vol. 7, no. 10, pp. 1–16, Oct. 2011. DOI: 10.1371/journal.pcbi.1002195. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1002195.

[49] S. S. Abby, J. Cury, J. Guglielmini, B. Néron, M. Touchon, and E. P. C. Rocha, "Identification of protein secretion systems in bacterial genomes," *Scientific Reports*, vol. 6, p. 23 080, 1 Mar. 2016, ISSN: 2045-2322. DOI: 10.1038/srep23080.

[50] P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, and M. Chandler, "ISfinder: the reference centre for bacterial insertion sequences," *Nucleic Acids Research*, vol. 34, no. suppl$_1$, pp. D32–D36, Jan. 2006, ISSN: 0305-1048. DOI: 10.1093/nar/gkj014. eprint: https://academic.oup.com/nar/article-pdf/34/suppl_1/D32/3923468/gkj014.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkj014.

[51] J. Cury, T. Jové, M. Touchon, B. Néron, and E. P. Rocha, "Identification and analysis of integrons and cassette arrays in bacterial genomes," *Nucleic Acids Research*, vol. 44, no. 10, pp. 4539–4550, Apr. 2016, ISSN: 0305-1048. DOI: 10.1093/nar/gkw319. eprint: https://academic.oup.com/nar/article-pdf/44/10/4539/19695676/gkw319.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkw319.

A

Appendix A

Name	Accession number	Name	Accession number
MCR-1.1	AKF16168.1	MCR-3.9	AST36144.1
MCR-1.2	ANR95875.1	MCR-3.10	ATQ63376.1
MCR-1.3	ANJ15621.1	MCR-3.11	AUN87920.1
MCR-1.4	APM87143.1	MCR-3-12	AVZ47168.1
MCR-1.5	APM84488.1	MCR-4.1	ASR73329.1
MCR-1.6	AQK48217.1	MCR-4.2	AVK94777.1
MCR-1.7	AQQ11622.1	MCR-4.3	AUI38915.1
MCR-1.8	AQY61516.1	MCR-4.4	AVK94779.1
MCR-1.9	AVA31022.1	MCR-4.5	AVK94778.1
MCR-1.10	ASK49940.1	MCR-5.1	ASK40551.1
MCR-1.11	ATM29809.1	MCR-5.2	AVM85875.1
MCR-1.12	BBB21811.1	MCR-6.1	ASK49942.1
MCR-1.13	AVM85874.1	MCR-7.1	AUR80098.1
MCR-1.33	UGY30527.1	MCR-8.1	AVX52225.1
MCR-2.1	SBV31106.1	MCR-9.1	WP_001572373.1
MCR-2.2	ASK49941.1	MCR-10.1	QDO66747.1
MCR-3.1	ASF81896.1	ICR-Mc	EGE18576.1
MCR-3.2	OYN70668.1	ICR-Mo	WP_082741435.1
MCR-3.3	ASU10319.1	ICR-Ba	WP_034495833.1
MCR-3.4	SBZ31568.1	ICR-Bb	WP_064558897.1
MCR-3.41	MBA2799562.1	ICR-Bf	WP_064546189.1
MCR-3.5	ASU04896.1	ICR-Bg	WP_064511805.1
MCR-3.6	AST36140.1	ICR-Bi	WP_120062886.1
MCR-3.7	AST36141.1	ICR-Bn	WP_064554336.1
MCR-3.8	AST36143.1		

Table A.1: Protein names and accession numbers for the resistance genes used to build the Profile HMM.

B

Appendix B

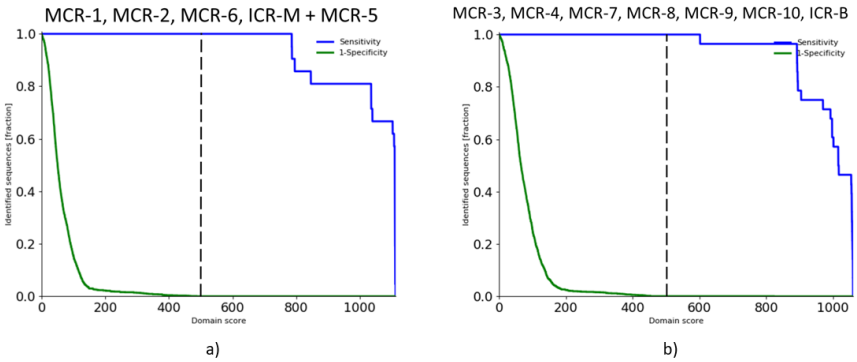

Figure B.1: Sensitivity and 1 - specificity as a function of the domain score for the separated model where MCR-5 is grouped with MCR-1, MCR-2, MCR-6, and ICR-M. The sensitivity and specificity are based on the fraction of correctly classified sequences. The used threshold score of 500 is marked with a dashed line.

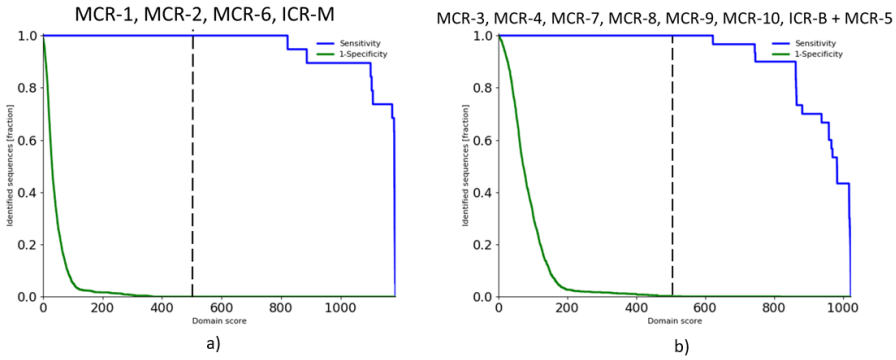

Figure B.2: Sensitivity and 1 - specificity as a function of the domain score for the separated model where MCR-5 is grouped with MCR-3, MCR-4, MCR-7, MCR-8, MCR-9, MCR-10, and ICR-B. The sensitivity and specificity are based on the fraction of correctly classified sequences. The used threshold score of 500 is marked with a dashed line.

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

CHALMERS
UNIVERSITY OF TECHNOLOGY