NEIGHBORHOOD CLUSTERING TO ANALYSE ANTIMICROBIAL RESISTANCE IN BACTERIAL GENOMES

by

CHANDANA NAVANEKERE RUDRAPPA

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
August 2021

I would like to dedicate my thesis to my parents.

Table of Contents

List of Tables

List of Figures

Abstract

Valuable insight into gene function and evolution can be obtained by analysing the order of genes in prokaryotic genomes, as neighboring genes often share related functions and evolutionary histories. Obtaining precise functional predictions is particularly important in the case of antimicrobial resistance (AMR) genes, as subtle differences in similarity patterns can reflect the potential for an organism to be treatable or resistant to one or more antibiotics. Databases such as the Comprehensive Antibiotic Resistance Database (CARD) provide high-quality predictions, but there is a significant gray area ("Loose hits" according to CARD) where genes differ in sequence from the reference sequence and may or may not confer AMR.

We introduce an approach to compare the genomic neighborhoods of AMR genes in genomes with different degrees of relatedness, to provide additional insight into their potential function. Our approach uses a technique to identify candidate AMR, then applies novel similarity measures and application of the UPGMA, MCL and DBSCAN graph-clustering techniques to identify patterns of similarity among gene neighborhoods. This analysis is complemented by phylogenetic analysis to assess the similarity of identified genes as well as their neighborhoods. We also provide a graphical tool to visualize the gene content in sets of neighborhoods.

AMR gene neighborhoods were observed to be very similar within closely related members of species including *Salmonella* Heidelberg. The proximity of some Loose hits to other AMR genes in many neighborhoods provided additional evidence for their function, whereas in other cases the CARD Loose hits were isolated and likely not associated with AMR. We also considered a set of genomes that encompassed several enteric pathogens. In this set, we found cases where seemingly poor Loose predictions were associated with clusters of AMR genes, and instances where gene order was surprisingly similar across distantly related genomes which may indicate recent transmission of AMR genes between pathogenic organisms. Our method provides new insights into the function of candidate AMR genes, and these refined predictions can be used to predict resistance and identify candidate evolutionary events.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Robert Beiko for believing in me and offering continuous support, valuable advise and guidance throughout my thesis work and also my entire Master's studies.

I would like to take this opportunity to thank my friends who supported me in all walks of life.

I am grateful for the Department of Computer Science for providing me this opportunity, and funding from Genome Canada, Research Nova Scotia, and NSERC to support my project.

I would also like to acknowledge the help of every member of Beiko Labs for their insightful suggestions and unparalleled support.

A special thanks to Keerthi, Harshitha and Abhishek for their invaluable support, confidence and motivation.

Chapter 1

INTRODUCTION

As the population grows and even though humans use antibiotics to fight infectious diseases, bacteria continue to be responsible for many diseases and deaths. Evident to this fact is that about one-quarter of all deaths worldwide each year result from infectious diseases caused by microbial pathogens and such bacterial infections are treated using antibiotics [75]. Antibiotics that are designed to attack bacterial infections have converted countless diseases from being considered as severe and lethal into simple and quick to vanquish. However, the combination of bacteria's ability to evolve and adapt and the overuse of antibiotics has led to a troubling problem: many types of bacteria are increasingly able to defeat antibiotics. The ability of bacteria to defeat the antibiotics that are designed to kill them is known as antimicrobial resistance (AMR) and this AMR is escalating to alarmingly high levels in all parts of the world. The causes of this increase include natural resistance, gene mutations, and transmission of genes between different strains and species of bacteria. The selective pressure from the use of antibiotics may provide an advantage for the mutants by increasing their resistance.

The US Centers for Disease Control and Prevention (CDC) reported that at least 2 million people are infected with antibiotic-resistant bacteria and more than 23,000 people die annually as a consequence of bacterial infections [1]. Global rise in AMR poses a threat to the human community and it is vital that we understand the structure, movement, functions, and genes in the bacterial genomes to ultimately understand and control the spread of AMR. The National Institute of Allergy and Infectious Diseases (NIAID) believes that a better understanding of the fundamental biology of microbes, their ability to block antimicrobial drugs, and host-pathogen interactions can help scientists identify novel drug targets and develop novel diagnostics and vaccines [89]. These novel diagnostics and vaccines can be used to mitigate the spread of AMR.

Genome-based research can reveal critical insights for battling antimicrobial resistance. The specific mechanisms of resistance and its origin can be determined by isolating and analyzing the genomes of the same microbial species from different human populations or various geographical locations [47]. For instance, one of the dataset chosen for this work is comprised of *Salmonella* isolates from different parts of Canada extracted from specific parts of food and animals.

1.1 Background on genes, genomes and gene functions

1.1.1 Genes and genomes

All living organisms have genomes (Figure 1.1) that consist of molecules called DNA. Genes (Figure 1.1) are the segments of DNA that contain instructions for building the molecules that are responsible for the workings of the organism (usually, but not always, proteins). DNA is the chemical chain that contains the genes that code for different kinds of proteins. These genes can be transmitted from one organism to another from parent to offspring or between organisms through the process of lateral gene transfer (LGT).

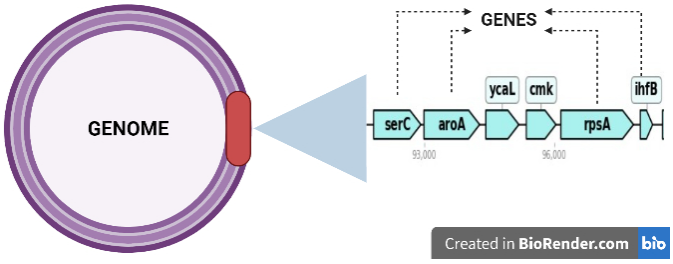

Figure 1.1: Representation of bacterial genes and genome. When a part of the large circular genome is expanded, it consists of genes (turquoise arrows) that are segments of DNA that encode proteins.

When compared with other life forms, bacterial genomes are typically smaller and there exists less variation in genomic size within and between species (Figure 1.2). Within bacteria, there is a strong correlation between the number of protein-coding genes and size of the genome. The size and content of the bacterial genome can be influenced by gene acquisition via LGT, gene duplication, genome reduction via loss of genetic material, and genomic rearrangement. One common feature of bacterial genomes is the relatively small spacing between adjacent genes. Several theories have

been formulated to explain the pattern of genomic size evolution amongst bacteria, including selective pressure on genome size to ensure faster replication, selection in favor of deletion of genomic material, mutation and genetic drift [73].

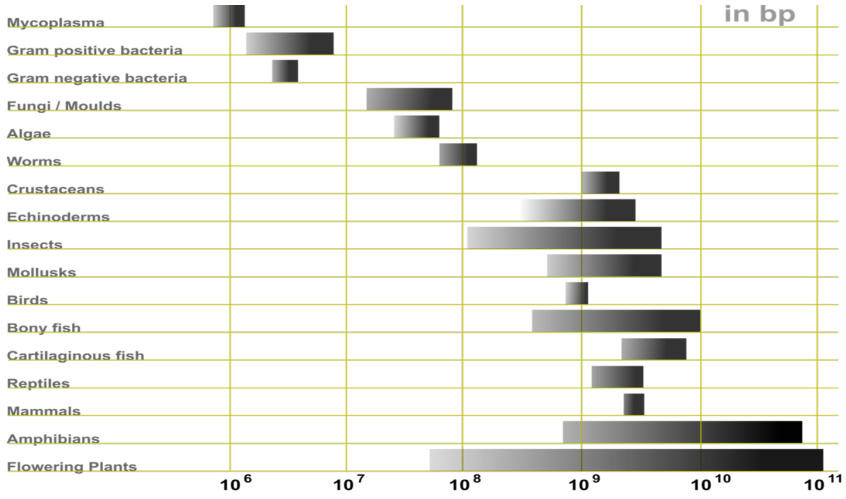

Figure 1.2: Variation in genome sizes in base pairs of various life forms [73]

1.1.2 Gene functions and function prediction

Bacteria vary greatly in properties such as metabolic capabilities including fermentation, environmental adaptations, and AMR. Even though a plethora of bacterial genomes and their corresponding gene and protein sequences are available, the function of many proteins remains unknown in spite of extensive experimental research. Computational prediction of gene function is a vital step in annotating newly sequenced genomes and predicting their capabilities. A common approach to predicting protein function is to compare the protein sequences encoded by the newly sequenced genomes to proteins of known function in reference databases; however, this process is complicated by the existence of multiple genes with similar sequences in many genomes. The most widely used method for characterizing newly sequenced proteins is through sequence similarity searches. These searches detect genes that have high levels of sequence similarity that likely reflects shared ancestry. Two genes are said to be homologous if they share a common ancestor and these homologs can be used to predict characterized proteins [97].

In the course of evolution, new species are formed when a group separates from the other members of the species and develops unique features and this is referred

to as a speciation event. Homologous genes in different genomes that diverged from a speciation event are termed *orthologs*. By contrast, *paralogs* are genes in a single genome whose relationships can be traced back to a gene duplication event where a single ancestral gene was copied twice and both copies were retained (Figure 1.3). The relationships between gene products and sequence divergence associated with the events of gene duplication or speciation can be explained with the help of orthologs and paralogs [60]. Orthologs tend to evolve more slowly when compared to paralogs and hence are favored while making function predictions from homologous genes [117]. Paralogs evolve more quickly and are more likely to mutate and take on other functions [17]. Due to the obvious necessity of duplication events in the process of generating paralogs, orthologs tend to share a slightly higher functional similarity than paralogs - the "ortholog conjecture" hypothesis [86].

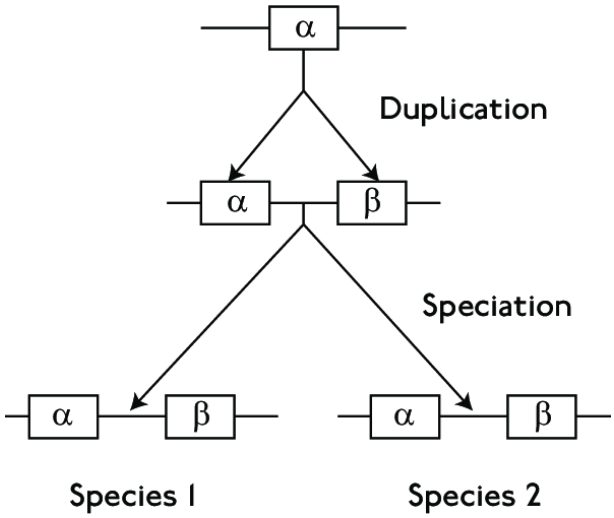

Figure 1.3: Orthologs and Paralogs. Gene α is duplicated in an ancestral genome and β is the duplicate copy of α. Copies of α and β are inherited by both descendant species as a result of a speciation event. These copies are related and are termed as "orthologs". Duplication causes α and β to be related and are termed as "paralogs" [40].

Predicting the functional annotations of gene products such as proteins is key to understanding disease mechanisms and related functions such as AMR. Recent advances in biotechnology have given rise to high-throughput experiments which accelerate and reduce the cost of obtaining distinct functional information about gene products [109]. Many solutions have been suggested to predict protein functions in

the last few decades [15] [108] [127]. However, it is difficult to decide which tools are best suited for predicting function because the process of protein function prediction is an open research problem [58].

Many types of data can be used for predicting the function of proteins such as protein sequence comparisons [81], protein-protein interaction networks [121], microarrays [54] , evolutionary relationships and genomic context [38]. These methods improve the ability to characterize the proteins and guide many biological experiments. The Critical Assessment of Functional Annotation (CAFA) [133] is a worldwide initiative aimed at analyzing, evaluating and improving the protein function prediction methods that have advanced functional prediction and novel annotations. As a contrast to traditional prediction methods and as a part of the CAFA challenge, a system that uses text from biomedical literature as a source of features to predict function was developed by [130]. This study concluded that when compared to a baseline classifier that uses sequence similarity alone, a text-based classifier performs better, and combining these text features with other types of features can potentially improve the performance of prediction methods.

Many bacterial genes are organized into groups of co-regulated genes called operons [92]. Operons are comprised of two or more genes that are adjacent to each other in the genome, are co-regulated, and whose protein products perform related tasks in the cell. Genes within operons tend to have related functions and they are conserved by vertical (parent-to-offspring) inheritance across species [99]. Hence, the functional relatedness of two adjacent genes and the fact that genes cluster together in multiple organisms suggests that they may belong to the same operon, which may provide an additional clue about their shared function. However, the task of inferring the function of a gene especially in a newly sequenced genome is still challenging. Strong protein-level similarities between a novel gene and a gene of previously known function can provide the best initial evidence [128]. Many new methods and theories are being proposed to interpret protein-based functional interactions based on the genomic context [55] [84]. By combining the methods of gene order conservation with gene fusion, the co-occurrence of genes in operons, and the co-occurrence of genes across genomes also known as phylogenetic profiling, significant context information

can be obtained for many genes. Genomic context can also be used to predict functional interactions between genes which serves as complementary to homology-based function predictions [65].

Given the tendency of genes with similar functions to cluster in the genome, gene order is an informative property as more and more genomes are sequenced and analyzed [119]. Many factors influence gene order including the organization of prokaryotic genomes into operons, LGT and hidden gene paralogy. The degree of gene order conservation correlates strongly with how closely the species are related; however, this conservation tends to be lost over time [119]. This loss can remove genes from an operon or even completely wipe it out. Recombination occurs when any two DNA molecules exchange a part of their genetic material with each other and events that cause this recombination can move genes within and between genomes; bacterial genes that are close to each other in the DNA are usually transferred together, a tendency reflected in the concept of genetic linkage [33]. Hence, that transfer of multiple nearby genes can lead to surprising levels of similarity in gene sequence and gene order among distantly related genomes.

1.2 Mobile genetic elements as agents of LGT

Bacterial strains that are multidrug resistant are a major cause of healthcare-associated infections around the world. The emergence of multiple antibiotic-resistant bacteria is driven in part by LGT (Figure 1.4). Diverse genes in a bacterial population collectively form a gene pool and bacteria can acquire pre-existing resistance determinants from this gene pool to gain antimicrobial resistance [96]. This DNA movement in prokaryotes is driven in part by activities of mobile genetic elements (MGEs). MGEs are segments of DNA that encode proteins and enzymes. The genes carried by these MGEs facilitate the movement of DNA within genomes or between two bacterial cells. MGEs are diverse in their nature, size, gene content, structure and mechanisms of transfer among genomes.

MGEs, in addition to genes involved in their mobility, frequently carry genes that are beneficial to the host organism, which contributes to the success of their transfer and maintenance in the recipient cell. MGEs also play a significant role in the acquisition and spread of resistance genes, and play a substantial role in the evolution of many bacterial genomes [29]. ** The vital roles of different types of MGEs

such as plasmids, genomic islands, integrative conjugative elements and integrons in the acquisition and transmission of AMR are briefly explained in the next sections.

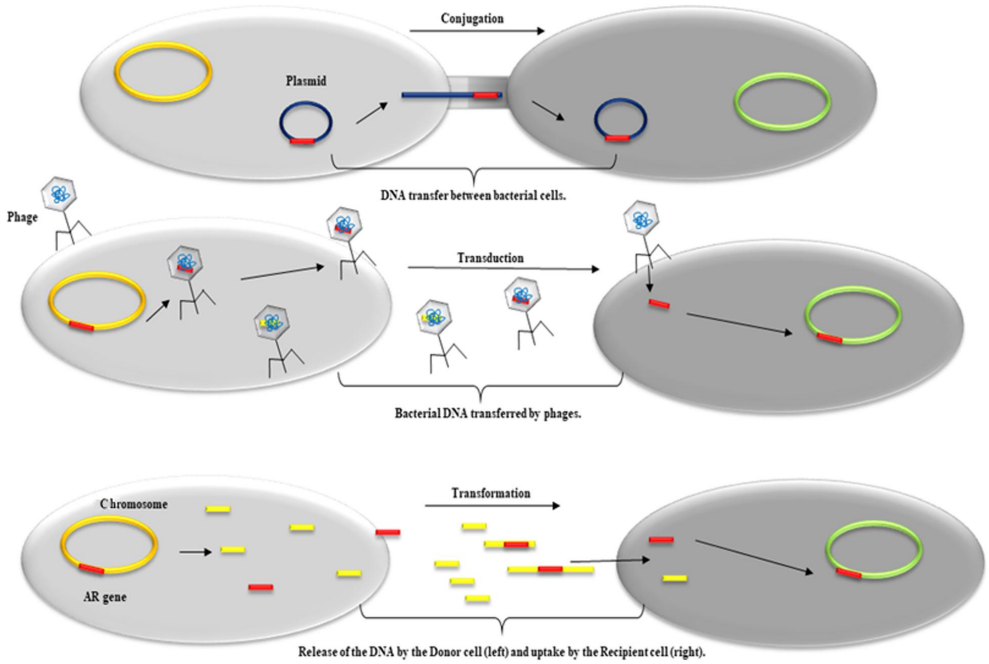

Figure 1.4: Various LGT mechanisms involved in the transfer of bacterial genetic material. The left cells represents the donor and recipient cells, respectively. The top event represents a plasmid-mediated DNA transfer between donor and recipient - conjugation. The second event shows the transfer of DNA between donor and recipient with the help of phages - transduction. The last event shows the direct uptake of DNA by a recipient once the genetic material is released by the donor - transformation. [19].

1.2.1 Plasmids

A plasmid is a circular, double-stranded DNA molecule that is spatially distinct from the cell's chromosomal DNA. Bacteria can acquire genetic advantages such as antibiotic resistance from the genes carried by plasmids. Plasmids can also carry other MGEs and have the ability to transfer both resistance and virulence determinants from one bacterium to another (sometimes to another species) via the process of conjugation (Figure 1.4) [124]. Unlike other MGEs, plasmids have a stretch of DNA (the origin of replication) that allows them to be replicated by the host bacterium. This explains the fact that plasmids can copy themselves independently of the bacterial chromosome, which can result in numerous copies of plasmids within a single bacterial cell. In spite of carrying fewer genes than the chromosome, plasmids can have a strong impact on the fitness of their host bacterium by conferring AMR, allowing the production of compounds that aid the host to kill other types of bacteria and digesting unwanted substances [115].

1.2.2 Genomic Islands

In bacterial genomes, a genomic sequence region extended across a number of orthologous genes that are co-arranged within another genome is called a synteny block [74]. Many of the accessory genes acquired by lateral transfer form syntenic blocks called genomic islands (GIs). GIs often carry important genes that can influence genome evolution via recombination (Figure 1.4) and LGT [16], pathogenesis [49] and antibiotic resistance [50]. Hence, the task of identifying such islands has now become a vital part of microbial genome analysis. Resistance islands are a class of GIs that contain multiple AMR genes, while genomic islands that contain virulence factors are often called pathogenicity islands. Bioinformatics studies have shown that GIs tend to carry more 'novel' genes (i.e. those that do not have orthologs in other species) than the rest of the genome [53]. The demand for those genome analysis methods that rely on regions that encode significant adaptations is increasing [66], but more work needs to be done to understand the mechanism of GI transfer and to make more accurate GI predictions in genomes [12].

1.2.3 Other classes of elements

Integrative and conjugative elements also known as conjugative transposons are MGEs that play a key role in bacterial adaptation [59]. These elements make up a large part of MGEs. The genes encoding key components of the ICE life cycle are often grouped into functional modules which may be exchanged among ICEs as well as with other mobile elements. Like plasmids, ICEs are self-transmissible by conjugation, but unlike plasmids, they must integrate into the host chromosome and be replicated as part of it. In addition to the core modules that mediate ICE integration, excision, conjugation and regulation, ICEs routinely encode a range of accessory functions, including virulence factors and resistance proteins for antibiotic and heavy metal resistance. ICEs have an important role in the dissemination of antibiotic resistance genes in pathogens such as *Vibrio cholerae*, which causes cholera in humans. The SXT^{MO10} ICE was described in *V. cholerae* O139, and it carries the genes that encode resistance to four antibiotics: sulfamethoxazole, trimethoprim, chloramphenicol, and streptomycin [125]. Thus, ICEs combine features of other classes of MGEs, such as bacteriophages which can integrate and excise from the host chromosome but do not transmit via conjugation; transposons that integrate the same way as ICEs but are not transferred horizontally; and plasmids which transfer between cells by conjugation but can replicate independently of the bacterial chromosome [131].

Integrons

Bacteria tend to replace their pre-existing genetic material with highly similar genes or fragments of genes. This process is known as homologous recombination [32]. Integrons use homologous recombination to transfer AMR and other types of genes between defined sites and exchange specific DNA elements called gene cassettes[51]. Integrons are composed of three major elements: *intI*, a gene that encodes an integrase responsible for insertion, deletion and rearrangement of gene cassettes; *attI*, a recombination site where the cassettes are inserted; and the means to express acquired gene cassettes [22]. An integron by itself lacks the ability to independently move and hence integrons as a whole cannot be considered as mobile. However, the gene cassettes harbored by integrons are considered mobile. Integrons are distributed in multiple copies in various locations of a genome, this helps integrons to facilitate the exchange of sequences between identical or related segments via homologous

recombination [134].

Several integron classes have been identified based on their IntI amino acid sequence. Among them, class 1 integrons (CL1s) are highly prevalent in antibiotic-resistant Gram-negative bacteria which are often embedded in plasmids and transposons which facilitates lateral transfer into a wide range of pathogens [41]. CL1s are often associated with Quaternary ammonium compound (QAC) resistance [4]; ampicillin, tetracycline and sulfamethoxazole-trimethoprim resistant genes in *E.coli* [64]; and streptomycin, tetracycline and sulfisoxazole resistance in *Salmonella* Typhimurium. [103].

1.3 Key genes in pathogens

Pathogen and pathogen function

The micro-organisms which possess the ability to cause disease to their host are referred to as pathogens and the severity of the disease symptoms is called virulence. The four most common types of pathogens associated with the disease are viruses, bacteria, fungi, and parasites. All living organisms are affected by pathogens including bacteria which are attacked by specific viruses called bacteriophages or simply phages [100].

There are a variety of ways through which a pathogen can cause illness to its hosts such as damaging the host cell walls during a replication event by the production of toxins [11]. When the cell walls are damaged due to toxins, it makes it easier for the pathogen to replicate. Colonizing the host, identifying the nutritionally compatible spots in the host, suppressing the defensive host response, duplication by utilizing host support system and disseminating to a new host are essential abilities of a pathogen that determine its survival and viability [2].

Virulence Factors

One of the major interests in microbiology and infection biology is comprehending which bacterial characteristics contribute most to a disease. Pathogenic bacteria produce molecules known as virulence factors (VFs) that can attack the host at the cellular level [90]. Bacterial pathogens possess a number of such VFs that determine the ability to cause various types of damage or diseases. Such factors include:

1. Surface components encoded on plasmids that allow the bacterium to invade

host cells [25].

2. Structural features called fimbriae and pili that help the pathogens attach to host cells [104] [68].

3. Secreted products like endotoxins, exotoxins and enzymes that degrade host tissue by causing inflammation and lethal shock [69].

4. Siderophores, small molecules secreted by bacteria that facilitate iron transport across cell membranes.

Siderophores are considered a major virulence factor during infection because they help pathogens acquire iron and damage the host [82]. As iron is necessary for many enzymatic reactions, a fierce battle for iron acquisition arises between host and pathogen during infection.

The virulence factor database (VFDB) provides in-depth information on major VFs present in various bacterial pathogens. The current version of VFDB is a repertoire of details on 16 important bacterial pathogens, virulence-associated genes, protein structural features, functions, mechanisms and important literature [24]. To improve the understanding of host–pathogen interactions, a web-based VF database - Victors was created. Victors provides a comprehensive, curated database of human, animal and zoonotic pathogen VFs [110].

AMR genes

Mutations and acquisitions of novel genes can induce resistance in a sensitive bacterium. Bacteria sensitive to a specific antibiotic must have a target region (e.g., an essential protein) for the antibiotic to act on and a mechanism that transfers the antibiotic into the cell before it is activated. Bacteria can develop resistance either through mutations in previously acquired genes or through LGT [14]. Restricting the overuse of drugs, modifying the target of the antibiotic attack, drug inactivation and efflux activation are a few of the control measures developed to fight AMR. [105]. Examples of AMR genes that confer different types of resistance to specific types of antibiotics include:

1. Fluoroquinolones are associated with multi-resistance and are frequently used for treating salmonellosis. Resistance to fluoroquinolones can be conferred by

mutations in the proteins that are targeted by antibiotics [123]. Drug efflux, which pumps toxic substances out of the cell, is another important mechanism of resistance that is conferred by proteins known as efflux pumps. Excessive expression of multidrug efflux pumps such as AcrAB-TolC and plasmid mediated quinoline resistance genes (qnr) are recently identified resistance mechanisms to fluoroquinolones [9].

2. Resistance in fosfomycin-susceptible bacteria is conferred by (i) mutations that occur in the proteins involved in transporting fosfomycin across the cell membrane, (ii) production of the fosfomycin inactivation enzyme FosA, and (iii) amino acid substitution in the active sites which decreases the fosfomycin binding affinity [57].

3. Aminoglycosides are particularly active against aerobic, Gram-negative bacteria such as members of the Enterobacteriaceae family, including *Escherichia coli*, *Klebsiella pneumoniae* and others. These antibiotics target the bacterial ribosome and terminate protein synthesis. There are three types of mechanisms of resistance against aminoglycosides: resistance due to efflux pumps (OprM), altering the target ribosome (16S rRNA) and enzymatic inactivation of the antibiotic molecule by enzymes such as AAC(3) and AAC(6) [106].

4. One of the three largest and important classes of antibiotics are β-lactamases which attack penicillins, monobactams, carbapenems and cephaloshporins.

   Genes encoding β-lactamases can be located on chromosomes, plasmids, integrons and transposons that inhibit the synthesis of the bacterial cell wall [106]. Plasmids can contain multiple β-lactamases of different classes, which confers broad and high-level β-lactam resistance. These plasmids and the corresponding resistance can spread amongst many bacterial species, making them resistant to most of the known β-lactam antibiotics [20].

5. The genes found in gene cassettes extracted from class 1 integrons present in plasmids or other mobile regions are extremely diverse but include many genes that confer resistance to antibiotics, including aminoglycosides, β-lactams,

chloramphenicol, trimethoprim, streptothricin, rifampin, lincomycin, and erythromycin [30].

Several tools have been developed to predict the occurrence of AMR genes in bacterial genomes. The open-source, manually curated database ARDB (http://ardb.cbcb.umd.edu/) provides information on antibiotic-resistant genes (ARGs), their gene annotations and resistance profiles, related proteins, and external links to other protein and gene databases. NCBI AMRFinderPlus has a browser that identifies acquired genes and mutations in both protein and nucleotide sequences (www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/). The Comprehensive Antibiotic Resistance Database (CARD) allows the user to either browse or download a curated collection of sequences and mutations underlying AMR. A key feature of CARD is the Antibiotic Resistance Ontology (ARO) which provides a hierarchical organization of AMR genes, antibiotics, mechanisms, and other information. The Resistance Gene Identifier (RGI) is both an online and a standalone tool that predicts AMR genes in bacterial genomes according to user specifications. RGI uses CARD as its central database to predict AMR genes (https://card.mcmaster.ca/analyze/rgi), and divides its predictions into three major categories. The Perfect hits are detected when AMR proteins are a 100% match to the reference sequence in CARD, while 'Strict' hits vary from the CARD reference sequence within the curated cut-off similarity score, and are useful for identifying variants of AMR genes that are previously unknown or antibiotic targets that are altered via mutation. The 'Loose' hits are found when the model cut-offs are very low. The "Loose" criteria allows detection of novel, predictive threats and more distantly related AMR gene homologs, with the risk of identifying homologous sequences and partial hits that may not contribute to AMR . Analysis of Loose hits supports novel AMR gene function prediction and research [3].

1.4 Literature review

1.4.1 Conservation of gene order

Analyzing the gene location and gene associations is an important area of genetics. Several methods have been proposed for comparing gene order across multiple genomes and detecting local gene order conservation. The amount of gene insertion/deletion and local rearrangements allowed are the two factors on which these

various methods depend [129].

Using the concept of orthologs is one of the well-known methods to compare gene order. The Clusters of Orthologous Groups (COGs) protein database [120] is based on all vs all comparisons between reference proteins from multiple genomes; a recent application was used where functional analysis of the COGs in 86 *Elizabethkingia* genomes revealed information on unique gene families associated with "information storage and processing". This analysis proved that *Elizabethkingia* has shown adaptive evolution to environmental change [72].

One of the most common approaches to identify orthologs is by using the reciprocal best hit (RBHs) method. Consider two proteins a and b on two different genomes A, B; if gene a from genome A finds has gene b on genome B as its best BLAST match and vice versa, they are RBHs and are considered as orthologs. The most commonly used program for finding sequence matches that uses RBHs is BLASTP [21]. Although BLAST searches are highly effective at finding homologous matches, they often assign the highest score to a protein that is not the closest evolutionary neighbor (e.g., an ortholog) of the query sequence. The potential pitfall due to this is that orthologs could be missed when using the RBH approach [63].

Operons are sets of adjacent genes that are expressed together and have related functions. Due to their close physical and functional linkage, gene order within operons is more highly conserved than between them [126]. Bork and coworkers came up with the concept of 'über-operon' which refers to a pair of genes whose functional and regulatory contexts remain to be intact even after an operon has undergone many rearrangements on itself and its individual genes [67]. The 'über-operon' concept was extended by examining the conservation of gene pairs in at least three genomes to minimize the probability that such pairs were conserved during evolution and not just shared by chance [107].

The set of genes that are physically close to a given gene in a chromosome or plasmid are termed its *neighborhood*. The gene neighborhood can include genes in both the upstream (before the gene start point) and downstream (after the gene end) directions. There is no fixed number of genes in a neighborhood, and the choice of neighborhood size is up to the researcher. A deeper understanding of the evolutionary relationships between genomes, gene-function prediction and detection of potentially

interacting proteins can be obtained by analyzing the neighborhood of genes of interest in a genome [129] [119]. For instance, conserved gene clusters in distantly related species may point to the occurrence of operons or otherwise functionally related gene groups [85] [93].

1.5 Gene order analysis and visualization tools

Several tools such as CGUG, CoreGenes, GeneOrder and Synteny Portal have been developed to study the conservation of gene order [78] [70]. GeneOrder relies on the BLAST sequence alignment algorithm where the scores are divided into three levels: high, medium and low, but there exists a limitation on the size of the genomes that can be provided as input [23]. OrthoCluster implements many variable constraints including the maximal percentage of mismatched genes. This application is flexible for the identifying the synteny blocks among species that have different evolutionary distances [132]. In contrast to OrthoCluster, an automated suite of programs was developed to explore the conservation of gene order that allows definitive identification of orthologs which can be used to evaluate the prokaryotic gene order conservation independent of their taxonomic distance [71].

1.6 Caveats of gene-order comparison

Although gene order can provide vital information about AMR genes and their flanking genes, comparing neighborhoods between genomes is not trivial, for several reasons in addition to the orthology problem introduced above:

1. There exist cases where the order of orthologs is not maintained even within similar species with same prokaryotic lineage. One such instance was found shortly after the sequencing of the genomes of *E. coli* and *Haemophilus influenzae* [85].

2. Functional annotations of bacterial genomes are not completely reliable and are prone to numerous errors. Incorrect assignment of start and stop of translated proteins, false prediction of genes, missed genes and frame-shifts are a few of the errors that affect gene order [116], [31].

3. Pseudogenes are segments of the bacterial genome that are very similar to functional genes but have become non-functional due to the accumulation of mutations and are very difficult to identify. The genome of *Mycobacterium leprae*

has over 1000 pseudogenes which results in uncertainty in the process of identification of orthologs necessary for gene-order comparison [28].

4. Gene order can be quite different because of a large number of genomic rearrangements [91]. The processes of insertion, deletion, and translocation result in changes in gene copy number, orientation, and position [98]. These events increase the fluidity of the bacterial genomes and make it very difficult to analyze the gene order between distantly related species.

1.7 Introducing distance measures to compare gene neighborhoods

Early observations showed that gene order is usually disrupted when the average protein sequence identity of orthologs shared between two genomes is <50% [56], hence introducing parameters during sequence comparisons to identify orthologs is necessary. Application of the BLAST criteria where the homologs have the expected rate of false positives of a certain statistical expectation value, limiting the length of the sequences to be compared and ensuring that both sequences are the best hits to each other (reciprocal confirmation) are a few of the parameters that can be introduced to identify the right orthologs [62].

In order to overcome the more serious problems of gene-order comparison of distantly related species, there is a need for distance measure that can be used to compare and cluster groups of similar neighborhoods. To handle the gene-order distortion and to quantify the conserved patterns observed, a measure called the neighborhood disruption frequency (NDF) was used. This NDF score between the genomes ranged between 0 which indicated complete conservation of gene order and 1 which corresponded to complete rearrangement. Using this measure, the study was able to gain insights into the rate of disruption of gene order and the genes that were responsible for genome rearrangement [118]. Another computational method - SNAP (Similarity-Neighborhood Approach) which used similarity scores as distance measure was developed for finding functionally related gene sets from genomic context. SNAP relies on a similarity-neighborhood graph (SN-graph) constructed from the chains of similarity and neighborhood relationships between orthologous genes in different genomes and adjacent genes (neighborhood) in the same genome. However, this approach requires computationally heavy resources and can be applied only to a

limited number of genomes [61].

1.8 Clustering techniques

Pairwise similarity or distance scores among a set of gene neighborhoods can be aggregated using a range of clustering techniques, which can reveal interesting patterns and information on AMR genes. There are many clustering techniques that apply various types of distance measures while grouping the genes of multiple prokaryotes. Affinity propagation that takes preference and kernel radius as parameters [45], clusterONE which accepts number of clusters and size of the threshold as variables [87], K-means and modified DBSCAN-based methods [35] are some noteworthy clustering methods. Many conventional clustering algorithms suffer from problems when analyzing gene data. Specifying the number of clusters, providing strength against noise, and poor handling of embedded and intersected clusters are a few of the limitations [94]. In such conditions, many biologists prefer hierarchical clustering methods that generate a set of divisions based on cluster hierarchy in the form of an output dendrogram.

The gene neighborhoods computed can also be considered as a biological network or a key-value pair graph. For such types of networks, nodes are represented using genes or proteins whereas the pre-computed similarities or functional linkages between the genes serve as edges. Self-organizing maps which are a recent addition to clustering gene data [77] and Markov clustering [10] are the two well-known graph based clustering techniques. These network-based clustering techniques are useful in understanding phenomena such as protein-protein interactions. As proteins tend to function in groups, studying these interaction networks can help in protein function prediction [122].

1.9 Thesis objectives and organization

High performance DNA sequencing and bioinformatic tools have drastically improved the process of investigating AMR by identifying the genes that confer resistance [18]. There is a need for new, fast and reliable methodologies that can provide more insights into gene-order conservation, gene function, and AMR propagation to better understand the risks posed by the transmission of AMR. The main objective of this work is to find AMR genes of interest in a reference genome, efficiently retrieve the sequence homologs from bacterial genomes, then compare the neighborhoods of AMR

genes in the genomes of closely and distantly related species. These neighborhoods can provide additional insights on various evolutionary events that may have contributed to AMR transmission. In the first part of the thesis we examine different approaches to compute and use gene-order information in the analysis of AMR genes. The second part of the thesis concentrates on the use of gene order to examine the more-uncertain Loose-hit predictions of the CARD. Phylogenetic analysis and clustering of the neighborhoods are used to predict whether CARD "Loose" hits are likely to confer antimicrobial resistance or the hits are indeed false positives.

We find that AMR gene neighborhoods are largely conserved within closely related members of species including (*Salmonella*), with notable exceptions that can be highlighted using our method. Examination of the neighborhoods of Loose hits provided more information on their probable functions. We generate a graphical representations of the genomic neighborhood surrounding the target AMR gene and the corresponding regions for its homologs in each comparison genome. Our approach helps to identify gene orthologs and potential functional gene clusters, and functional inference from clusters of Loose hit AMR genes. We identified specific cases where the neighborhoods of AMR gene models were affected by the insertions, deletions and lateral gene transfers amongst the members of *Salmonella* serovars. Our methods provided substantial evidence on AMR properties of candidate Loose hits of CARD which was supported by cluster dendrograms, gene order visualizations and phylogenetic trees.

The remainder of the thesis is organized into 3 sections. Chapter 2 concentrates on giving a detailed view of methods used for the analysis and also reporting results and conclusions. Chapter 3 includes the details of analysis when an additional AMR hit criterion in included in the analysis. The last chapter concludes the thesis by providing a brief summary and also future perspectives.

Chapter 2

CLUSTERING OF AMR GENE NEIGHBORHOODS

2.1 Motivation

The major objective of this chapter is to introduce and illustrate a new approach for the comparison of gene neighborhoods in a set of genomes. Our comparative approach is based on the assumption that gene order tends to be highly conserved among closely related bacteria.

Several mechanisms have been proposed to explain the extent and importance of gene clustering in understanding bacterial genome organization. Genes organized into clusters rather than the uniform distribution of conserved genes is a key feature of many bacterial genomes. Assuming that functional relevance and conservation of gene clusters are correlated, the gene-clustering property has been used to predict the function of genes through the annotations of their neighborhoods [42]. The 'guilt by association' principle states that the function of a neighboring gene in a cluster, whether from the same operon or merely adjacent, can be proposed if the function of one gene in a conserved cluster is known [7].

The chapter is divided into two major parts: the first part concentrates on providing a step-by-step detailed explanation on the methods performed (Figure 2.1) and to list the various datasets highlighting the outbreaks of many *Salmonella* strains over the recent years to understand why their analysis is vital. The second half focuses on applications of the methodology to test datasets. The results include detailed visualization of gene clusters, statistics, gene order and comments about AMR gene ortholog neighborhoods.

2.2 Methods

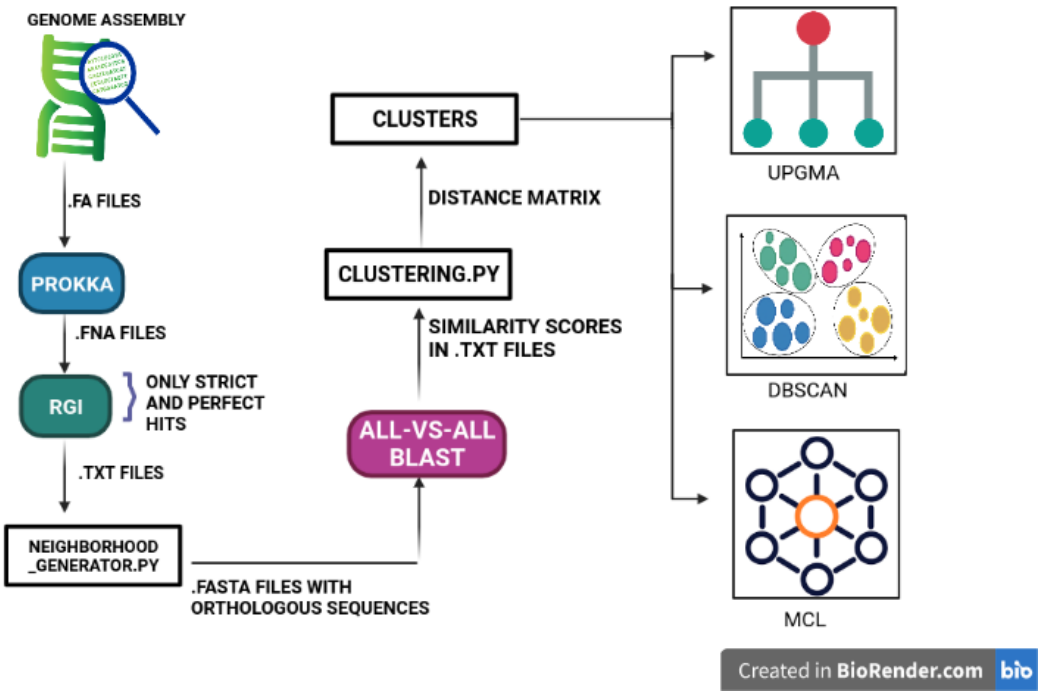

Figure 2.1: Steps involved in the approach including bioinformatic tools used (curved rectangles of blue, green and purple) and python scripts (.py extended names inside black rectangles) and the results obtained (pictorial representations inside a square). The output at every stage is shown using labelled emerging arrows. The first steps of the approach involve the collection of raw genome assemblies (FASTA-formatted files of genomes containing nucleotide sequences) and preprocessing those assemblies to extract two types of input files required for further analysis. Prokka - a program that annotates bacterial genomes and generates standard output files [112] and RGI were used for annotation and recognition of AMR genes within the genomes. Python scripts Neighborhood_Generator.py and Clustering.py were used to identify and divide the data based on various RGI models and construct their neighborhoods using the input files. All-vs-All BLAST scores of the neighborhood genes were used to compare multiple genomes and construct their corresponding neighborhood similarity matrix. The distance measure was applied on the similarity matrix to convert and generate the distance matrix that represented differences between each neighborhood. The distance and similarity scores were used as input for several clustering measures whose results were evaluated. The neighborhoods were visualized in a graphical format using libraries to understand the gene order of the neighborhoods.

2.2.1 Annotation and comparative analysis

Raw sequence assemblies extracted from three different datasets that we have chosen for our analysis were annotated to obtain the labelled relevant features using the command-line prokaryotic annotation tool Prokka 1.14.6 with the default specifications. Prokka was installed using bioconda from https://github.com/tseemann/prokka [48]. Out of various output files generated by Prokka, annotations were retrieved in GenBank and FASTA (.fna) format. The GenBank entries contain metadata and detailed information about the organism. Various details regarding important characteristics of the entry's sequence such as the presence of coding sequences, proteins, etc are provided by the "features" part of the GenBank entry [13]. Prokka generates corresponding protein sequences of the input nucleotide sequences in the FASTA format. For each genome, these are the two input data files that were used for the rest of the analysis.

Resistance genes present in each genome were identified using the CARD-RGI. RGI 4.2.2 was installed using bioconda from https://card.mcmaster.ca/download. The protein sequence file produced by Prokka was passed as the input to RGI with default specifications: all Loose hits of 95% identity or more were automatically listed as strict, regardless of alignment length. We included only high-confidence predictions (Strict and Perfect hits) from RGI in this chapter and address Loose hits separately in Chapter 3. Basic Local Alignment Search Tool (BLAST) 2.9.0 was used to identify homologous sequences and the output was generated in the tabular output format 6. BLAST provides high scoring segment pairs (HSPs) for each hit. The number of HSPs was limited by setting the threshold of max_hsps parameter to 1 so that only the matches with highest bit-score are returned. This choice was made to reduce the redundant BLAST entries and hence increase the computation time required to compare two neighborhoods.

2.2.2 GenBank and RGI data parsing

Data frames from the Python pandas package were used as data structures to extract, store, and manipulate the tabular formatted RGI output data. RGI predictions, which were obtained as outputs from the previous section includes the coordinates and orientation of the gene and the score with respect to a given AMR gene model. The

GenBank data was handled with the help of Biopython (https://biopython.org/) which provides methods for parsing the GenBank sequences. Bio.SeqIO provides an interface to handle input and output of sequence file formats including multiple sequence alignments but it considers the sequences only as SeqRecord objects (https://biopython.org/wiki/SeqIO). The SeqIo.parse class, which uses records and qualifiers, was used to extract important gene information such as contig details, locus tag, gene start, stop, orientation, product, and translation of the protein. One of the problems of gene annotation methods as discussed in the introduction is that there can be entries where the gene details are missing. Such genes were labelled as UID (unidentified) for the purpose of analysis and visualization to differentiate them from genes with predicted functions.

2.2.3 Neighborhood visualization using DNA feature viewer

To plot the sequence annotations from GenBank or General Feature Format (GFF) records, many tools are made available and DNA sequence visualization has become a common requirement in bioinformatics. We used DNA Feature Viewer [135] which is a tool that annotates sequences from GFF or GenBank format by converting them into a graphical format. The command-line version of DNA Feature Viewer was installed and modified as per the requirements of the project. Graphic records visually define the features of genes in each neighborhood annotation. One neighborhood corresponds to one graphic record and the distance between the first upstream gene to the last downstream gene is considered as the length of one graphic record. The DNA_Feature_viewer provides an option for specifying the gene start, stop, orientation, name, colour, and gene name information for each graphic feature to obtain the gene order visualization. The colour codes of predicted AMR genes corresponded to the category assigned by RGI: green for Perfect hits, yellow for Strict hits, and orange for Loose hits. DNA Feature Viewer provides an option of a translator. The translator is a set of style tags which can be defined once but can be used to ensure style consistency across annotation plots throughout a project. This translator was used to maintain the gene color, font style of labels, and highlight the AMR gene.

Figure 2.2 illustrates how DNA Feature Viewer automatically generates the visual elements of a graphic record to improve conciseness and readability.

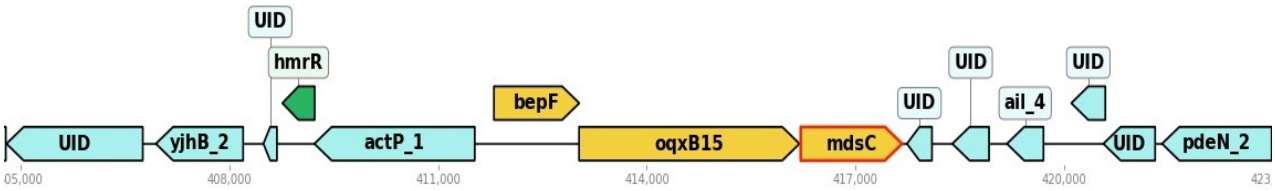

Figure 2.2: Gene order visualization using DNA Feature Viewer. Each gene is represented as a turquoise arrow. Feature labels which are either gene names or the "contig ends" are displayed directly inside their corresponding feature arrow, and the font color is automatically selected (as black or white) to fit the feature's background color. Labels which do not fit inside a feature arrow are displayed above it (*hmrR* and *ail_4* in this example). Finally, all features and label texts are organized along different vertical levels to avoid collisions. This ensures that the resulting plot remains readable irrespective of the figure's width. The relative positions of the genes in that contig is denoted by the indices shown below the genes based on start and stop indices. The thick red border is used to highlight the RGI gene of interest.

2.2.4 Identification of orthologs

For every target AMR gene model identified by the RGI, a python script that identified the orthologs and their corresponding neighborhoods was used. For each of the AMR gene models recognised by RGI, orthologs present in other genomes of the dataset were identified. In these analyses, a given CARD AMR gene was analyzed if it was predicted to be in more than 35% of the total genomes of the dataset. The start and stop information of the AMR gene was used as a reference to obtain the 10 upstream and 10 downstream genes which were considered as the neighborhood. We fixed the length of neighborhood to 10 in our analysis to avoid the exclusion of many genomes with shorter contigs. The orientation details of the AMR gene were used to align and orient the AMR gene within its neighborhood. One way to select 'suitable' homologs is to apply an E-value or percent identity cut off. In our approach, two protein sequences are treated as homologs for comparative purposes when they share more than 70% overall sequence identity.

The contig information played a vital role in the identification of neighborhood genes. As the GenBank data for each genome was large and hard to process, it was divided into separate sections based on the available contig information to ensure that the orthologs and the neighborhood genes on the same contig genes were identified. In some cases, the GenBank assembly lacked the entry for a specific contig on which

an ortholog was identified; these cases were ignored. In cases where ten upstream or downstream genes were absent due to truncated contigs, the contig ends in the neighborhood of the target AMR gene were marked in the visualization (Figure 2.3).

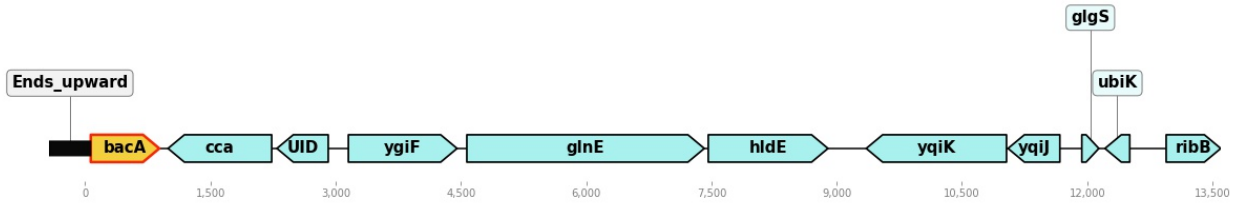

Figure 2.3: A neighborhood visualization demonstrating a contig end. The neighborhood of the AMR gene *bacA* (yellow) consists of 10 downstream genes whereas the contig ends with no genes in the upstream denoted using the "Ends_upward" tag.

2.2.5 Sequence similarity and distance matrix

The scores obtained from the alignment algorithms such as BLAST can be used to measure sequence similarity. In our approach, all the AMR genes along with their neighborhood genes were compared using All-vs-All BLAST, with each query sequence from a neighborhood compared with all the subject sequences from the other neighborhoods. The second part of the model involved reading the BLAST outputs in the tabular format into data frames. The details of gene identifier, alignment start, alignment end, and BLAST scores such as E-value, percent identity and bit-score were stored in the tabular format in the form of .txt files. These parameters were used to in the construction of neighborhood similarity matrix.

We used the BLAST bit-score to represent the similarity between pairs of genes. In addition to the E-value, the bit-score also can be used for statistical indication in a BLAST output . The bit-score provides measurement of sequence similarity without depending on length of the query sequence and database size, it is also normalized based on the pairwise alignment score. The bit score (BS) can be determined using the formula:

$$BS = \frac{\lambda S - lnK}{ln2}$$

where λ refers to the Gumbel distribution constant which is an extreme value

distribution, S refers to raw alignment score, and K refers to a constant associated with the applied scoring matrix. *BS* is proportional to the raw alignment score (S) and hence, higher the bit score, the better the match according to the scoring matrix (https://www.ncbi.nlm.nih.gov/BLAST). As the bit-score values vary over a broad range, normalization was used to eliminate redundant data and ensure that good quality clusters are generated and to improve the efficiency of clustering algorithms. The notation that was used to normalize the bit-scores of a ALL-vs-ALL comparisons is mentioned below:

$$NBS(g1, g2) = \frac{BS(g1, g1)}{BS(g1, g2)}$$

where NBS refers to normalized bit-score of a comparison of gene1(g1) with gene2(g2), BS(g1,g1) is the bit-score value obtained when the gene1 is compared with itself and BS(g1,g2) is the bit-score value obtained when the gene1 is compared with gene2. Using this notation helped us to normalize the bit-score values between 0 and 1.

The results of ALL-vs-ALL BLAST of the homologous pairs were used to compute the similarity matrix that shows the level of conservation in each neighborhood. Similarity scores between two neighborhoods(N1,N2) were computed based on two main criteria:

1. If neither N1 and N2 included contig ends, the similarity score of N1 and N2 was the aggregation of normalized bit-scores of the matched hits in the neighborhood. The maximum score between N1 and N2 would be 21 which occurs when the 10 upstream and 10 downstream along with target AMR gene of N1 have a 100% match with no bit-score differences to the 10 upstream and downstream genes of N2.

2. If either of the neighborhoods (N1 for instance) included a contig end and the difference in neighborhood was because of genes missing only due to the contig end, the similarity score of N1 and N2 was modified with the presumption that the missing genes were the matches to the genes in the neighborhood N2 and it was not penalised.

We then transformed the neighborhood similarity matrix into a distance matrix. Our methodology assumes that a distance matrix D, can be defined, whose element

$D_{i,j}$ represents the dissimilarity point at row i and column j. This is simply a measurement of how dissimilar AMR gene neighborhoods are to each other. The notation used to convert an entry of the similarity matrix into an entry in the distance matrix is:

$$D_{i,j} = 1 - \frac{SM_{i,j}}{max(SM)}$$

where $SM_{i,j}$ is the similarity matrix score at point i,j and the max(SM) represents the maximum similarity score for that complete neighborhood. The distance matrix depicts how each neighborhood relates with respect to the other based on factors such as bit-scores and neighborhood length. A symmetric distance matrix is obtained as the difference between N1 and N2 is the same as the difference between N2 and N1 and the diagonal elements were zero as every neighborhood is completely identical to itself.

2.2.6 Clustering

Once distance matrices were constructed for each pair of neighborhoods, we applied three types of clustering: UPGMA, MCL and DBSCAN to obtain clusters of similar neighborhoods. Many clustering algorithms rely on proximity or similarity between data objects and hence measuring distance between data objects acts as the foundation for clustering. Our approach builds a distance matrix based on neighborhood comparisons to construct the cluster hierarchies. Genes with similar roles in the cell often cluster together and hence we believe clustering the distance matrix generated by comparing the neighborhoods can reveal interesting patterns and information on AMR genes.

UPGMA

Hierarchical clustering is one of the most common methods employed for classification in field of biology. Classifying the organisms of different populations or species based on gene order, finding sequence similarity between genes or proteins and identification of genes with matching profiles are a few of the common uses of hierarchical clustering [8].

The unweighted pair group method with arithmetic mean (UPGMA) approach is used frequently in bioinformatics and ecology. We adopted UPGMA as it uses the

similarity across all data points and hence considered to be the most robust amongst other single - linkage hierarchical clustering methods [76]. UPGMA follows the hierarchical procedure of iteratively clustering similar genes from the given dissimilarity matrix. The algorithm investigates the structure of the pair-wise distance matrix to construct a rooted tree. The two nearest clusters are combined into a higher-level single cluster at each step. The distance between any two clusters m and n is calculated as the average of all distances between pairs of objects "u" in m and "v" in n, that is the mean distance between points of each cluster. The d[m,n] entry corresponds to the distance between cluster m and n. At each iteration, the UPGMA algorithm updates the distance matrix to reflect the distance of the new cluster - m with the remaining cluster. The distance is calculated using the equation:

$$d(u, v) = \sum_{m,n} \frac{d(u[m], v[n])}{(|m| * |n|)}$$

for all points u and v where $|m|$ and $|n|$ are the cardinalities of clusters m and n, respectively (scipy.cluster.hierarchy.linkage). The output of UPGMA is always a tree which is also known as a dendrogram. The UPGMA clustering algorithm was applied from the python scipy package cluster.hierarchy and the dendrogram was plotted using figure factory of plotly library. A pre-computed distance matrix for the neighborhoods was given as input to cluster similar neighborhoods which were represented using a dendrogram.

Graph-Based clustering

Graph and network based analysis techniques provides a way in which a biological entity can be analysed based on its local neighborhood in the graph and also the network as a whole entity. Conventional clustering techniques such as K-means generally follow a pairwise approach where they consider only the individual relationship between two biological entities rather than incorporating the higher-order interactions with their neighbours [44].

The Markov clustering (MCL) algorithm is designed to perform well, specifically while clustering the simple or weighted graphs. A single parameter, inflation, controls the extent of output clustering. We used the graph-based Markov clustering from the Markov-clustering module (https://github.com/guyallard/markov_clustering). A

similarity matrix was provided as the input for MCL and graphical network of each neighborhood with the distribution of the most significant clusters was generated by MCL. Similarity scores were used for clustering as MCL interprets the matrix entries or graph edge weights as similarities, and works well for undirected input graphs as suggested in MCL documentation.

Density-based clustering

Density-based spatial clustering (DBSCAN) is a widely used clustering technique where a density threshold is associated with the linked region. The size of the neighborhood (epsilon) and the minimum points (min-points) within the given cluster are the two predefined parameters which direct the DBSCAN and determine the quality of clusters. Using these two parameters, DBSCAN divides the input data points into core points (input data points which satisfy a minimum density requirement), border points (points in cluster that are not core points) and outlier categories. DBSCAN chooses a random point that has not been assigned to a cluster or been designated as an outlier and computes its neighborhood to determine if it's a core point. If true, it starts to cluster around this point or label the point as an outlier otherwise. Once a core point or cluster is identified, DBSCAN expands the cluster by adding all the points that are reachable to the cluster. All density-reachable points are calculated and are added to the cluster. A point's status is updated to border point if an outlier is added. These steps are repeated until all the points in the input are either assigned to a cluster or marked as an outlier (medium.com/dbscan).

DBSCAN was used from a popular python machine learning library Scikit-Learn as their implementation was found to be scalable and well-tested. The generated distance matrix was fed as input to DBSCAN that generates the clusters visualized in a 2 dimensional scatter plot. The denser cluster signifies the most significant neighborhood and the genomes that fall into that neighborhood which was not that informative for smaller datasets. The epsilon parameter was varied using hyperparameter tuning to determine the optimal number of clusters to understand whether the similarity of the neighborhood depends on the size of the clusters.

2.3 Genomic datasets

We focused on the genus *Salmonella* in the analyses described in this chapter. *Salmonella* is subdivided into *serotypes* which have the same type and number of surface antigens [5]. Studies prove that along with other properties, gene order is preserved extensively in closely related species, owing to this, only the genomes of different types of *Salmonella* serovars are used in the first part of the analysis. *Salmonella* is a pathogenic bacterium whose clinical manifestations range from common gastroenteritis (diarrhea, abdominal cramps, and fever) to enteric fevers that fall into the categories of life-threatening febrile systematic illness [46]. Recently, drug-resistant *Salmonella* has been associated with a considerable number of outbreaks in the U.S; given its importance and the potential for recent evolution we focused on this group. The table below provides a few important cases of such outbreaks related to *Salmonella* serovars used in this study. The details regarding the various outbreaks are extracted from the Centers for Disease Control and Prevention (CDC).

***Salmonella* serovars**	**Year**	**Outbreak-mode**	**Resistant to**
S. Typhimurium	2018	Dried coconut	Ampicillin, azithromycin
S. I 4,(5),12:i:	2015	Pork products	Ampicillin, treptomycin, sulfisoxazole, tetracycline
S. Enteritidis	2015	Raw, frozen and stuffed chicken	Ampicillin, tetracycline
S. Heidelberg	2014	Chicken	Three or more classes of antibiotics
S. Hadar	2011	Turkey Burgers	Ampicillin, amoxicillin, cephalothin, tetracycline

Table 2.1: Table of recent *Salmonella* serovar outbreaks

The *Salmonella* spp. used in this study are divided into three categories of datasets. The first two datasets 15_S.Heidelberg and 100_S.Heidelberg were extracted as a subset of 2500 genome sequences from poultry collected along the poultry production continuum (farm, retail) through the Canadian Integrated Program for Antimicrobial Resistance Surveillance (CIPARS) and the Canadian Food Inspection Agency's 2013 National microbiological poultry baseline survey. The genomes for the 15_Diverse dataset were extracted from NCBI as a part of the project that focused on

Antibiotic Resistance and diversity of *Salmonella enterica* serovars associated with broiler chickens. These *Salmonella* genomes encompassing 6 serotypes were isolated from poultry farms to study the development of machine learning method for predicting AST(Antimicrobial susceptibility testing) from genomic datasets and development of experimental AMR gene-detection methods.

Dataset Name	**Number of genomes**	**Year, country of extraction**	**Isolates**
15_S.Heidelberg	15	2005-2017,Canada	S. Heidelberg
100_S.Heidelberg	100	2005-2017,Canada	S. Heidelberg
15_Diverse	15	2005-2008,Canada	S. Heidelberg(3), S. Typhimurium(1), S. Hadar(4), S. 4,5,12(3), S. Enteritidis(3), S. Kentucky(1)

Table 2.2: Details of various datasets used in the project

2.4 Results

2.4.1 Dataset 2.1: 15 S.Heidelberg genomes

The model was first applied on the smaller dataset 15_S.Heidelberg to analyze the distribution of AMR gene neighborhoods across very closely related genomes of the same species. As this dataset consisted of very closely related genomes, we expect to find highly conserved AMR gene neighborhoods with few variations. CARD's Resistance Gene Identifier (RGI) 4.2.2 with the CARD database 3.0.1 was run with default settings for all isolates to predict resistance phenotypes. Perfect and Strict hits predicted by RGI were included, where a Perfect hit is an exact match to the curated reference sequences, and a Strict hit is a previously unknown variant of known AMR genes that matches a reference sequence at or above a stringent threshold.

The initial step of the model provides a brief summary of the RGI genes present in each genome. In this dataset, a total of 15 genomes were analyzed and 41 unique gene models were identified. A specific AMR gene model was chosen for analysis only if it was present in more than 20% of the genomes of this dataset. Figure 2.4 shows the detailed statistics of the number and type of AMR gene present per genome and these statistics support our hypothesis of expecting highly conserved neighborhoods within genomes of closely related species.

Figure 2.5 shows the distribution of average similarity scores of neighborhoods of 41 gene models. As per the histogram, almost 20% of the total 41 gene models had extremely similar neighborhoods with average similarity scores between 20 and 20.5 and 31% of the gene model neighborhoods had very similar neighborhoods with similarity scores between 19 to 20.25 and only 6% showed minor variations (2 to 3 gene mismatches) with average similarity score of 18.5. For this dataset of 15 genomes with 41 unique gene models, the highest average similarity score observed was 20.75 and the lowest score was 18. A High degree of conservation was observed in the neighborhoods of AMR gene models identified in this dataset as expected as there were no neighborhoods with scores less than 18 according to the histogram.

Amongst all the gene models, the neighborhood of *Haemophilus influenzae PBP3 conferring resistance to beta-lactam antibiotics* (condensed to *Hi_PBP3*) which exhibits antibiotic target alteration resistance mechanism against cephalosporin, cephamycin, and penam (a class of beta-lactams), was highly conserved with average

(A)

Genome	Loose	Strict	Perfect
SA20153980	0	40	3
SA20153983	0	39	1
SA20160170	0	40	2
SA20160172	0	40	2
SA20160190	0	44	2
SA20160197	0	40	2
SA20160200	0	41	2
SA20160201	0	40	2
SA20160203	0	40	2
SA20160209	0	40	2
SA20160210	0	40	2
SA20160228	0	40	2
SA20160323	0	44	2
SA20160326	0	40	2
SA20171155	0	39	1

(B)

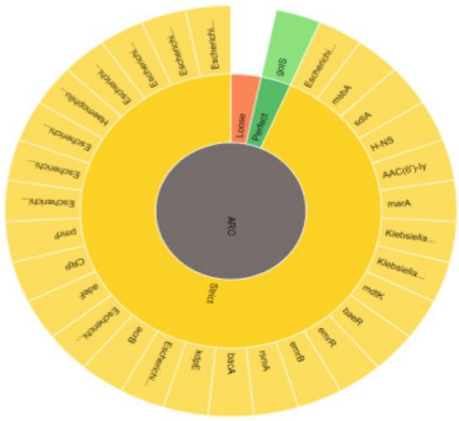

Figure 2.4: AMR gene statistics of 15 S.Heidelberg genomes analyzed. (A) Details of Perfect, Strict and Loose hits identified by CARD for each of 15 genomes. (B) Illustration of AMR classification for genome ID "SA20153983", sorted RGI results by AMR Gene Family (obtained from CARD's RGI web interface).

similarity score of 20.975. Figure 2.6 shows the conservation of gene order and conserved neighborhoods of the *Hi_PBP3* AMR gene. *Hi_PBP3* gene had 15 orthologs in all 15 genomes with no unidentified genes in the neighborhood. Only one neighborhood lacks all 10 upstream and downstream genes as its corresponding AMR gene was present on a short contig with few genes. The gene mismatches caused due to shorter contigs were not penalized in our scoring system and therefore had maximal similarity scores.

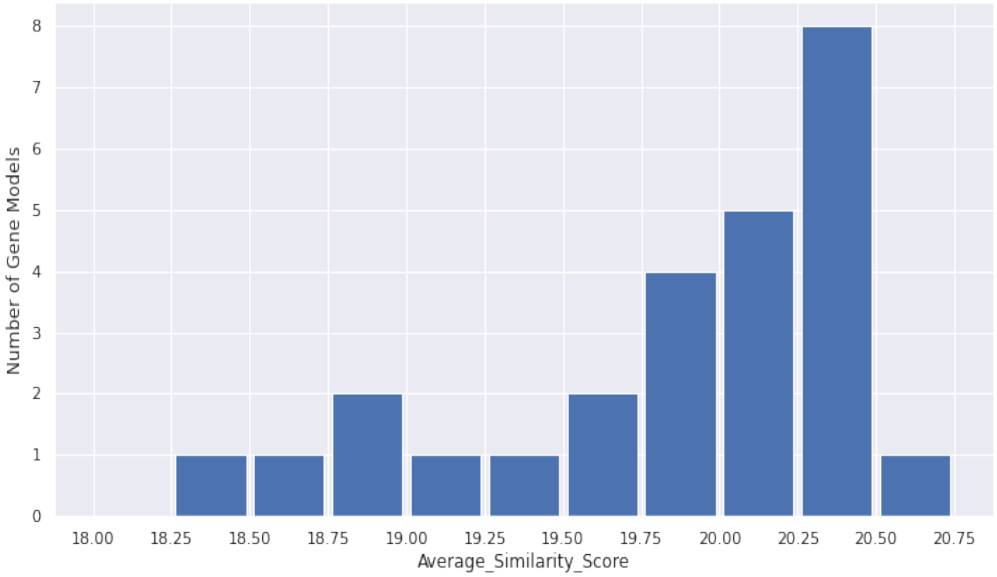

Figure 2.5: Histogram showing distribution of average similarity scores across the neighborhoods of 41 unique AMR gene models identified in 15 genomes of Dataset 2.1. The x-axis shows the number of gene models and y-axis indicates the average similarity scores between neighborhoods.

When the neighborhoods with lower average similarity scores were further investigated, the neighborhood of the AMR gene *mdtK* which lies on the maximum average similarity spectrum of histogram (Figure 2.6) showed minor variations. *mdtK* belongs to the AMR gene family of multidrug and toxic compound extrusion (MATE) transporters and exhibits the antibiotic efflux resistance mechanism against fluoroquinolone antibiotic. Figure 2.7 shows a heatmap of the similarity matrix of the neighborhood of 15 genomes generated based on the normalized bit-score of the homologous pairs of genes. The heatmap shows that most of the neighborhoods were very similar with average scores between 19 to 21. However, some exceptions were observed, most notably the genome with ID SA20160190 which had a similarity score of only 12 with most other neighborhoods. The distance matrix derived from the similarity scores is shown in Table 2.3. Most of the neighborhoods were very similar with maximum difference value of 0. The neighborhoods SA20160190 and SA20160323 showed the largest maximum differences of 0.5 between them indicating the gene mismatches in the neighborhood. The average difference value between the neighborhoods lies

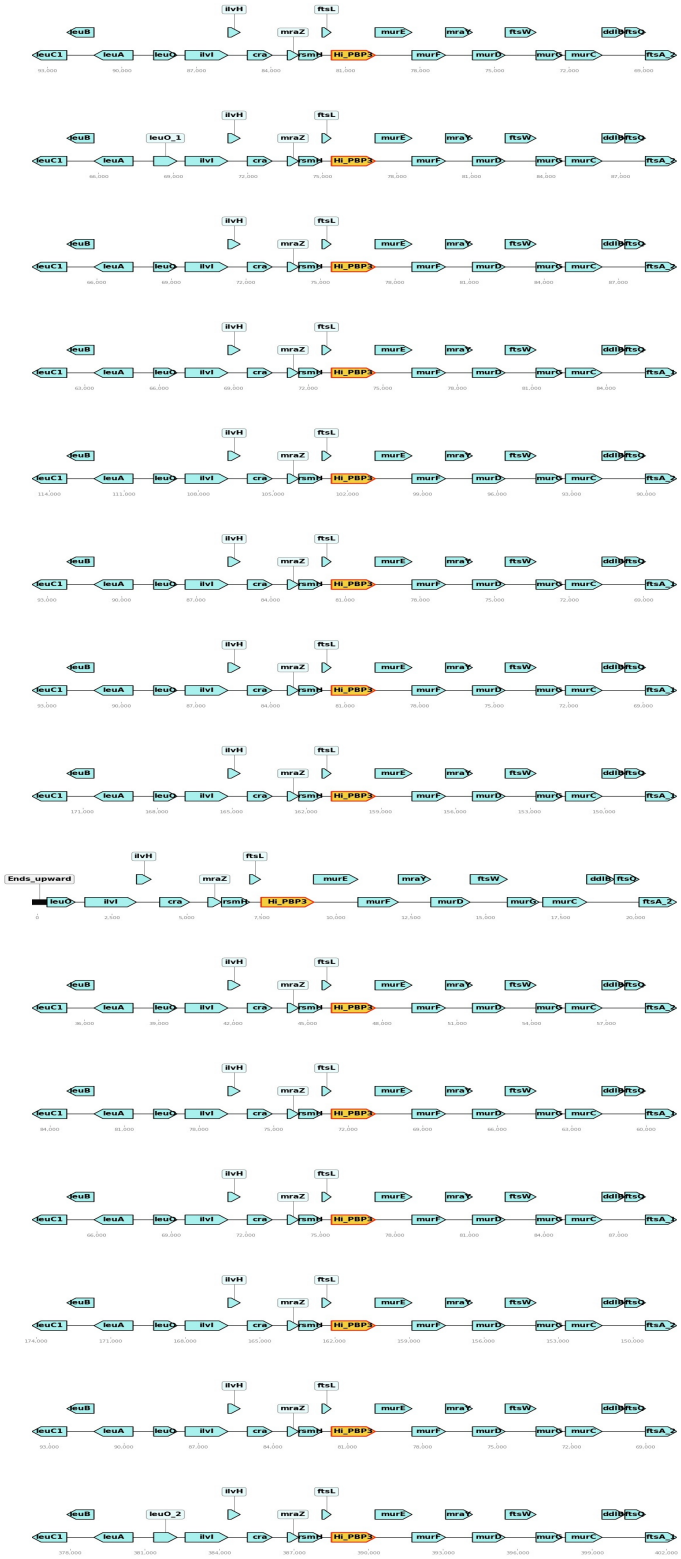

Figure 2.6: Neighborhoods of AMR gene model *HI_PBP3*: 15 neighborhoods, each originating from a different genome in the set, are shown, with the Strict hit AMR gene in yellow highlighted with a red border. The upstream and downstream genes of each neighborhood are represented using turquoise arrows with gene names.

between 0.03 and 0.15. The neighborhood analysis of this gene model revealed that insertion and deletion events can influence the neighborhood similarity.

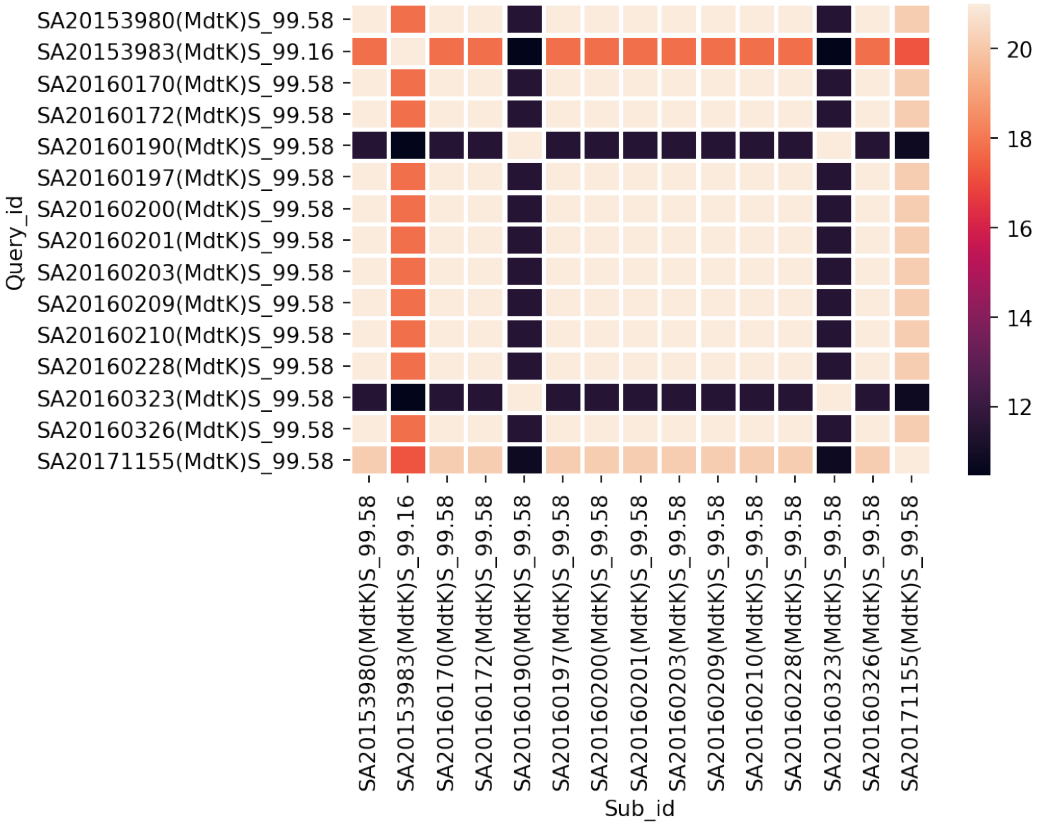

Figure 2.7: Variation of similarity scores across the neighborhood of *mdtK* gene model across 15 genomes represented in the form of an heatmap. The similarity scores range from 12 to 21. Query_id and Sub_id are the genome IDs of 15 genomes compared. The ID also indicates the name of the AMR gene and the percent identity match with the corresponding gene model.

Genome_ID	SA20153980	SA20153983	SA20160170	SA20160172	SA20160190	SA20160197	SA20160200	SA20160201	SA20160203	SA20160209	SA20160210	SA20160228	SA20160323	SA20160326	SA20171155
SA20153980	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20153983	0.151	0	0.151	0.151	0.502	0.151	0.151	0.151	0.151	0.151	0.151	0.151	0.502	0.151	0.179
SA20160170	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20160172	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20160190	0.452	0.502	0.452	0.452	0	0.452	0.452	0.452	0.452	0.452	0.452	0.452	0	0.452	0.485
SA20160197	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20160200	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20160201	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20160203	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20160209	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20160210	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20160228	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20160323	0.452	0.502	0.452	0.452	0	0.452	0.452	0.452	0.452	0.452	0.452	0.452	0	0.452	0.485
SA20160326	0	0.151	0	0	0.452	0	0	0	0	0	0	0	0.452	0	0.039
SA20171155	0.039	0.179	0.039	0.039	0.485	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.485	0.039	0

Table 2.3: Distance matrix for the neighborhood of AMR gene model *mdtK*. Rows and columns represent the 15 neighborhoods compared to obtain the distance matrix. Each value(i,j) in the symmetric matrix indicates the maximum difference obtained when the neighborhood on row i was compared with column j. The difference values range between 0.0-0.502.

UPGMA:

The distance matrix was provided as the input to UPGMA clustering method. Figure 2.8 shows the dendrogram of various clusters generated from the neighborhood distance matrix of *mdtK* gene model. There are many strategies that can be followed to cut a dendrogram at a position to obtain optimal clusters. We used the mean of the distance matrix as the measure to cut the dendrogram to obtain optimal clusters which can provide insights into neighborhood differences. Two clusters were obtained when this criterion was applied. Eleven genome neighborhoods in one cluster were identical and had a corresponding cluster height of zero; the other cluster consisted of two neighborhoods that were also identical.

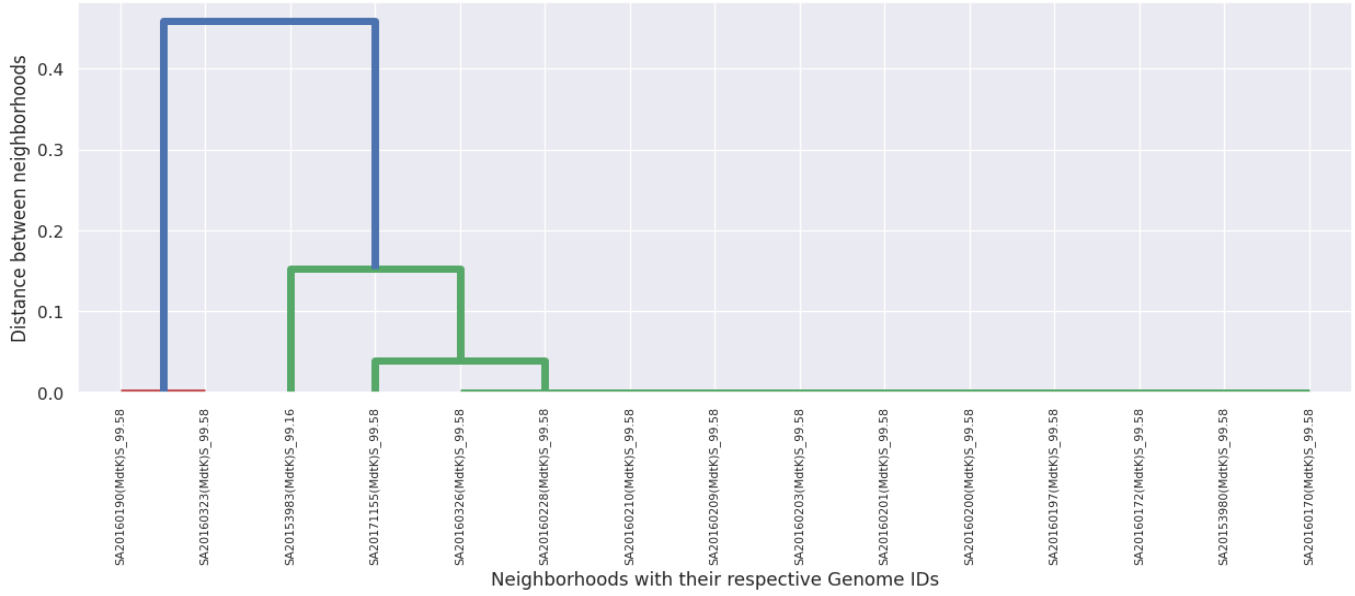

Figure 2.8: A dendrogram representing the clusters generated by UPGMA for the neighborhood of gene model *mdtK*. The x-axis shows the genome IDs of neighborhoods belonging to various clusters(red and green) as labels. The y-axis shows the distance between the neighborhoods at the time they were clustered.

MCL:

The similarity matrix was given as the input for the MCL algorithm as it uses similarity scores rather than distances. Each neighborhood was considered as a node connected by edges that are the similarity scores between the neighborhoods. 15 genomes were represented as nodes of indexes from 0 to 14 in the following order: 'SA20153980 - 0', 'SA20153983 -1', 'SA20160170 -2', 'SA20160172 -3', 'SA20160190 -4','SA20160197 -5', 'SA20160200 -6', 'SA20160201 -7', 'SA20160203 -8', 'SA20160209 -9', 'SA20160210 -10', 'SA20160228 -11', 'SA20160323 -12', 'SA20160326 -13', 'SA20171155 -14'. MCL exhibited different behaviors when the inflation parameter which controls the granularity of the output clusters was varied for a range of values. A good set of starting values are 1.4, 2, 4, and 6 according the MCL documentation (https://micans.org/mcl/man/mcl). When the inflation parameter was between 1.4 and 4, MCL assigned all the nodes into one big cluster indicating that there are no differences between neighborhoods (Figure 2.9). As the threshold was increased to 10 and above, MCL grouped the nodes 1, 4, 12, 13 and 14 together and all the remaining nodes were assigned to singleton clusters (Figure 2.10). MCL remained static and clusters were not changed even when the inflation parameter was varied between 1.4 and 10 (the optimal maximum value according to documentation). These findings proved that MCL clusters did not align completely with the hierarchical clustering as the differences between neighborhoods were not correctly accounted. Hence, MCL might not be a suitable clustering algorithm to be used while comparing the two neighborhoods as nodes in a graph in our case.

DBSCAN

The distance matrix was provided as the input to the DBSCAN algorithm that clustered the genomes based on the density parameter. The values of epsilon were varied over a range of values to evaluate the clusters in comparison with hierarchical clustering. The clusters generated by DBSCAN aligned with the clusters obtained from UPGMA method. DBSCAN provided 3 different clusters that showed major differences in the neighborhoods but did not consider the smaller difference values while clustering.

Figure 2.11 shows a detailed comparison of the clusters generated by UPGMA, MCL and DBSCAN for the AMR gene neighborhood of *mdtK*. Each neighborhood is

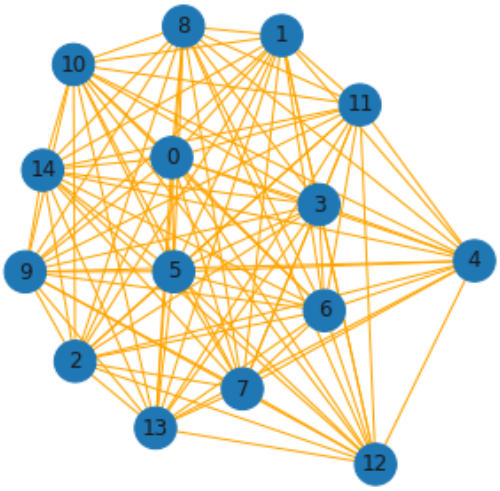

Figure 2.9: MCL clusters for the 15 neighborhoods of gene model *mdtK* when the inflation parameter is 10 and below. Each node (blue circles) represents a neighborhood containing the AMR gene *mdtK*.

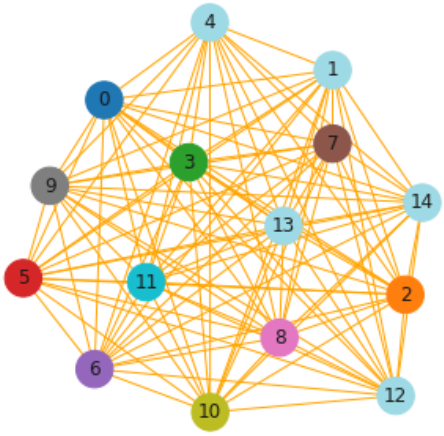

Figure 2.10: MCL clusters for the 15 neighborhoods of gene model *mdtK* when the inflation parameter is 10 and above. Each color indicates the cluster to which a particular node belongs. Nodes with same color indicated that the neighborhoods belong to the same cluster (blue nodes 1, 4, 12, 13 and 14)

assigned to a cluster and represented with a unique color for each cluster. UPGMA divides the 15 neighborhoods into 3 clusters (indicated using yellow, green and blue colors) with different cluster heights. All two (UPGMA and DBSCAN) techniques assign SA20153983 to a different cluster and this assignment is consistent and agrees with the similarity matrix shown in Figure 2.7. UPGMA and DBSCAN assigns most of the neighborhoods to same clusters (green and blue) which indicates that both the clustering techniques perform in a similar way when the difference between the neighborhoods is large as shown in the 2.3.

The clusters provided by MCL and DBSCAN did not completely agree with distance and similarity scores, as MCL mostly assigned many dissimilar neighborhoods to a single, large cluster, and DBSCAN did not consider minor similarity score differences which resulted in the formation of clusters that did not show notable key differences. Hence, in future analyses we applied only UPGMA clustering to cluster the neighborhoods of AMR gene models.

Although Dataset 2.1 was comprised of only 15 genomes, 41 unique models were identified indicating the necessity for a way to understand the overall variation of differences in the neighborhoods. These variations provide a better understanding of the level of conservation between neighborhoods and help us identify special cases that might show interesting differences used to gain information regarding the AMR genes. A histogram that showed the distribution of maximum differences between neighborhoods of the identified AMR gene models was generated (Figure 2.12) to provide the overall model statistics. The histogram shows that maximum differences varied between 0 (identical neighborhood) to the max value of 0.55. Maximum neighborhoods showed the difference between 0 and and 0.1 showing high levels of similarity. Only a single gene model showed a difference > 0.55 which supported our hypothesis of finding few neighborhoods with differences in this dataset of closely related genomes.

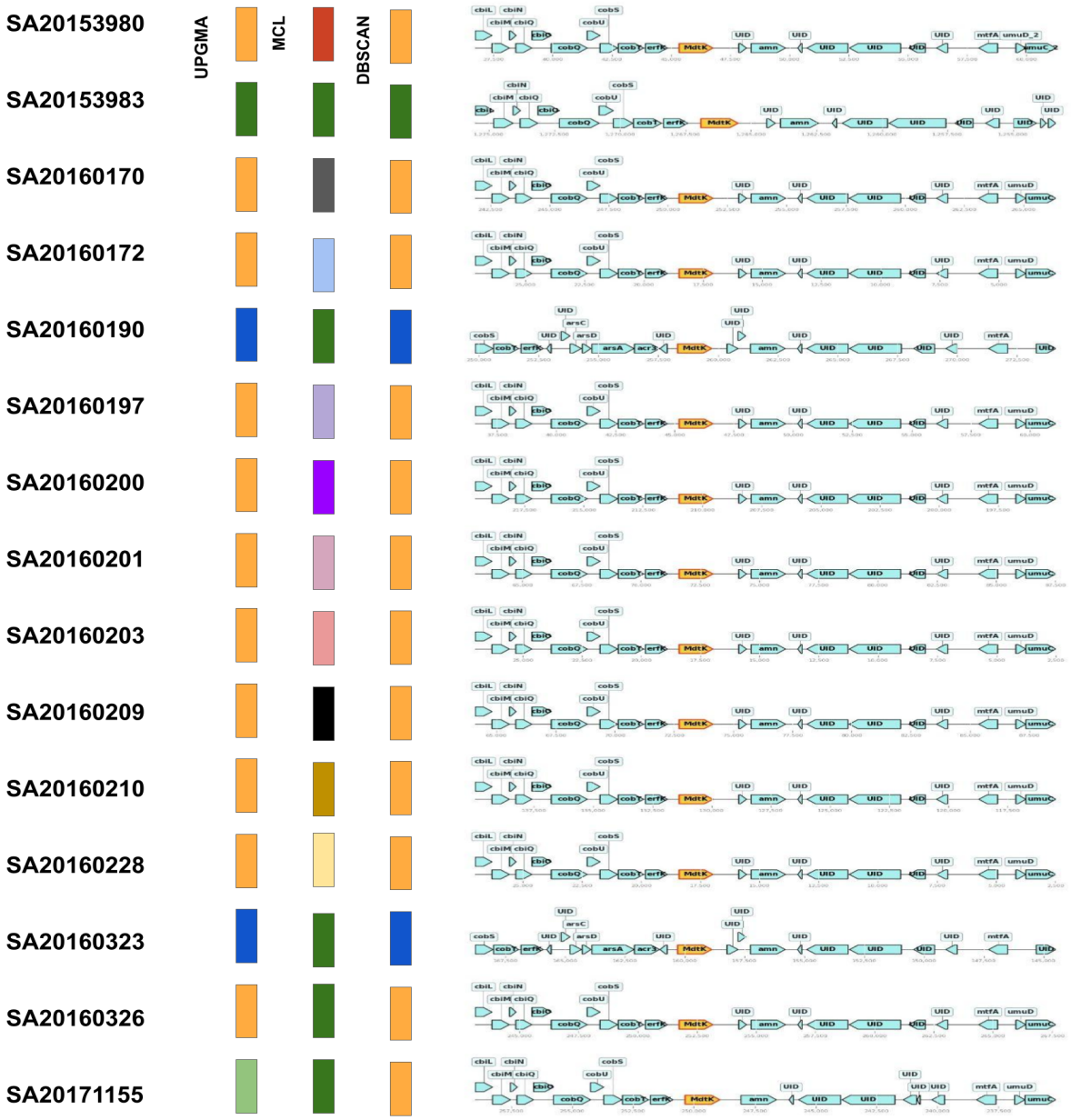

Figure 2.11: Comparison of various clusters generated by UPGMA, MCL and DBSCAN for the AMR gene model *mdtK*. Each color indicates the cluster to which the corresponding neighborhood with genome ID highlighted in bold belongs. Neighborhoods belonging to the same cluster are indicated with the same color. The gene orders of each corresponding neighborhood (*mdtK* with upstream and downstream genes) are shown towards the right end to visualize differences based on cluster information.

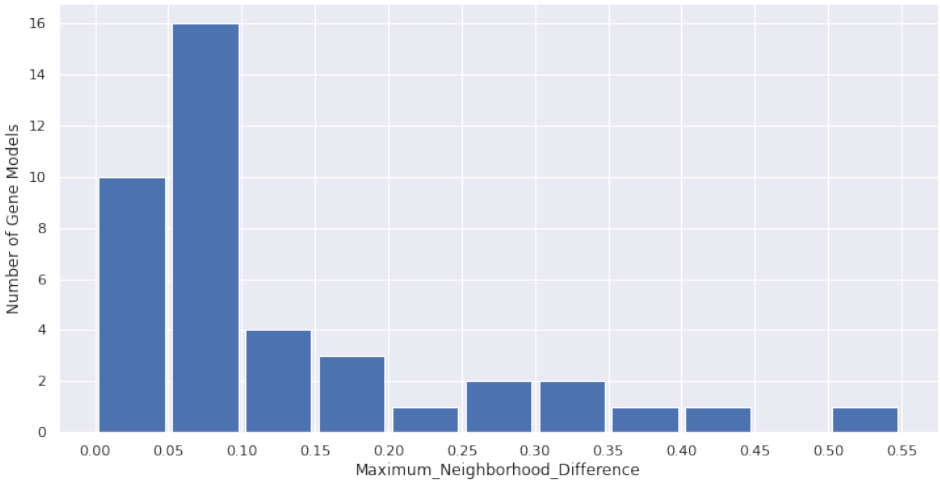

Figure 2.12: Histogram showing distribution of maximum differences across the neighborhoods of 41 unique AMR gene models identified across 15 genomes of Dataset 2.1. The x-axis shows the number of gene models and y-axis indicates the maximum difference between neighborhoods of each model.

2.4.2 Dataset 2.2: 100 S.Heidelberg genomes

As it is evident from the results obtained from Dataset 2.1, the level of conservation was high among the 15 genomes of S.Heidelberg, with a small number of exceptions. Hence, to evaluate and understand how the neighborhoods are conserved and the gene differences behave when the number of genomes are increased, we applied the methods to a larger set of 100 S.Heidelberg genomes. As we included more genomes of the same serovar, we expected conserved AMR gene neighborhoods with slightly greater variations compared to Dataset 2.1. A total of 100 genomes was analyzed, 31 unique gene models were obtained and UPGMA clustering technique was applied on the matrix derived from the neighborhoods. A specific AMR gene model was chosen for analysis only if it was present in more than 25% of the genomes of this dataset to avoid the analysis of those AMR gene models whose orthologs are not present in at least major part of genomes of the dataset.

Figure 2.13 shows the distribution of average similarity scores of neighborhoods of 31 gene models. As per the histogram, around 32% of the gene models had an average similarity score between 16 and 16.5 between neighborhoods indicating that the extent of similarity and gene order conservation was slightly less when compared

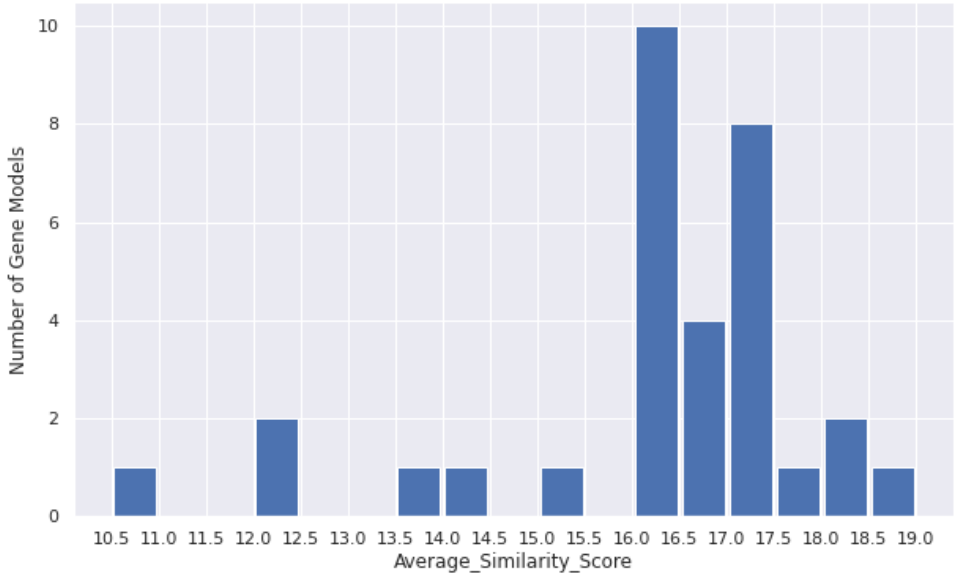

Figure 2.13: Histogram showing distribution of average similarity scores across the neighborhoods of 31 unique AMR gene models identified in 100 genomes of Dataset 2.2. The x-axis shows the number of gene models and y-axis indicates the average similarity scores between neighborhoods.

to Dataset 2.1. Around 10% of the total gene models had similarity scores from the minimum observed value of 10.5 to 14. The highest average similarity score among the 31 gene models was 19 out of a maximum of 21.

The neighborhoods of gene *mdtK* were analyzed with the intention of understanding the extent of conservation for a scaled dataset of 91 of the total 100 genomes. The neighborhoods of *mdtK* showed differences in the clusters with an average similarity scores of 15.725. Figure 2.14 shows the UPGMA clusters obtained for 91 S. Heidelberg neighborhoods. When the dendrogram was cut at the mean of distance matrix values, three major clusters were generated by UPGMA. A single neighborhood was sampled randomly from each of the two smaller clusters, while three were sampled from distinct lineages in the largest cluster.

Figure 2.15 shows each representative neighborhood from the three clusters obtained when the dendrogram was cut at a height of mean of the distance matrix. The neighborhoods 1, 2 and 3 belong to the large red cluster, 4 belongs to the blue cluster

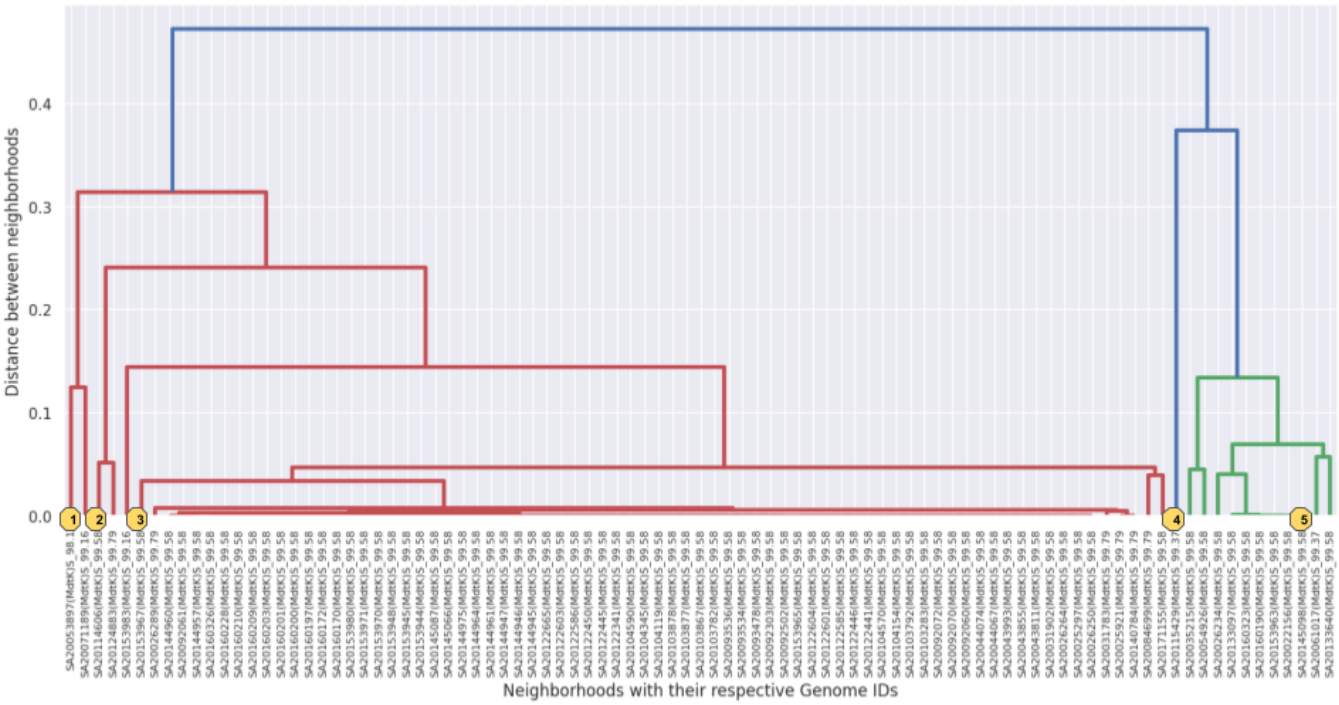

Figure 2.14: A dendrogram of clusters generated by UPGMA clustering denoting the key differences in the 91 neighborhoods of *mdtK* AMR gene model of Dataset 2.2. Each neighborhood from a major cluster obtained when the dendrogram was cut is numbered to visualize separately (yellow octagons with numbers); three representative neighborhoods were selected from the diverse red cluster. The x-axis shows the genome IDs of neighborhoods belonging to various clusters (red, blue and green) as labels. The y-axis shows the distance between the neighborhoods at the time they were clustered.

and the 5 was chosen from the green cluster as representative. The three neighborhoods 1,2 and 3 which all belong to the red cluster have all the 10 common upstream genes. The neighborhood 3 has only 4 upstream genes because it suffers from contig end and that is not penalized in our system. However, there are differences when we look at the downstream genes of these three neighborhoods in the red cluster. Neighborhood 1 has a unique gene *intS_3* which is absent in the other two neighborhoods and the neighborhoods 2 and 3 have only four downstream genes in common because there is an insertion of five unidentified genes between *amn* and *mtfA* in neighborhood 3. Neighborhood 4 is the only member of the blue cluster and does not share

major similarities with others in red and green. The downstream of neighborhood 4 is largely comprised of unidentified genes and the upstream region has many new genes - *acr3*, *arsA*, *arsD*, *arsC*. These two factors contribute to the low similarity scores and resultant is a blue cluster which is also a great way to choose dissimilar neighborhoods at first glance on dendrogram. Neighborhood 4 has ten genes in common (upstream and downstream) with the neighborhood 5 of green cluster whereas it shares only three genes in common with all the three neighborhoods of the red cluster.

Neighborhood 5 which belongs to the green cluster shares most of its upstream genes with neighborhood 4. The genes *amn* and the unidentified gene to the left of *amn* are conserved in almost all the neighborhoods except 1 (possible explanation: contig end).

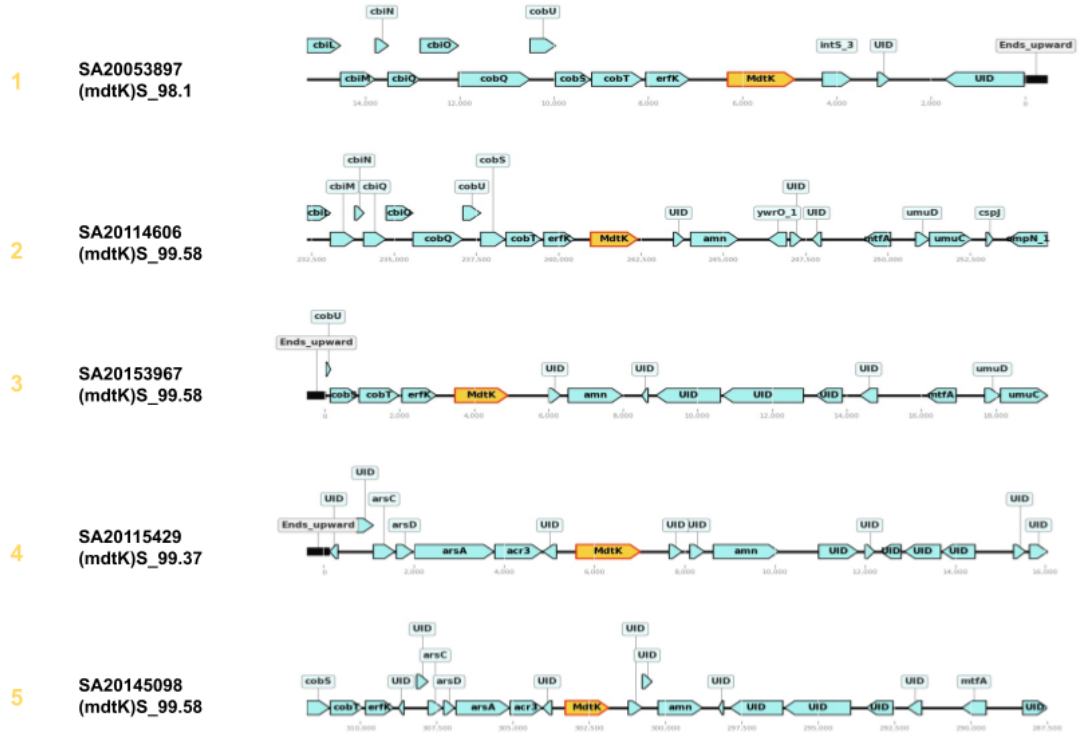

Figure 2.15: Visualization of each major cluster generated by UPGMA for the neighborhoods of *mdtK* AMR gene model in Dataset 2.2 denoting key differences. Each individual neighborhood is represented with the identifier of the corresponding lineage in the dendrogram followed by the genome ID of the neighborhood. The ID also indicates the name of the AMR gene and the percent identity match with the corresponding gene model.

A histogram was constructed to understand the overall statistics of the variation in

maximum differences between the neighborhoods of all the gene models of the Dataset 2.2 (Figure 2.16). According to the histogram, around 41% of the total 31 gene models had a maximum neighborhood difference between 0.35 and 0.45. Interestingly, 6% of the gene models were outliers with a maximum distance value between 0.66 to 0.7.

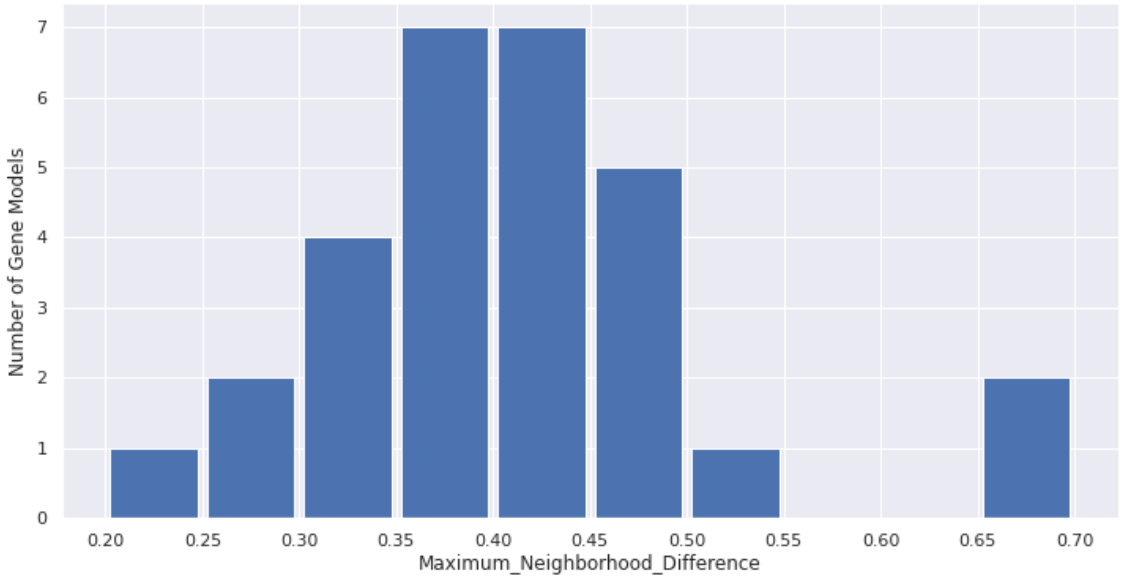

Figure 2.16: Histogram showing distribution of maximum differences across the neighborhoods of 31 unique AMR gene models identified across 100 genomes of Dataset 2.2. The x-axis shows the number of gene models and y-axis indicates the maximum difference between neighborhoods of each model.

When the gene models that lie on the spectrum of large maximum differences were further investigated, the neighborhood of an AMR gene *cpxA* which exhibits antibiotic efflux resistance mechanism against aminoglycosides and aminocoumarin antibiotics showed the maximum difference of 0.72 between neighborhoods. The *cpxA* model had orthologs in all 100 S. Heidelberg genomes.

Figure 2.17 shows the dendrogram generated by UPGMA for the 100 neighborhoods of the *cpxA* gene model. The dendrogram shows two major clusters (red and green) which are both divided into several smaller clusters. It is evident from the dendrogram that the neighborhoods showed major differences with maximum cluster height of 0.68. These neighborhoods in each cluster were closely analyzed for key differences in upstream and downstream genes by visualizing the gene order.

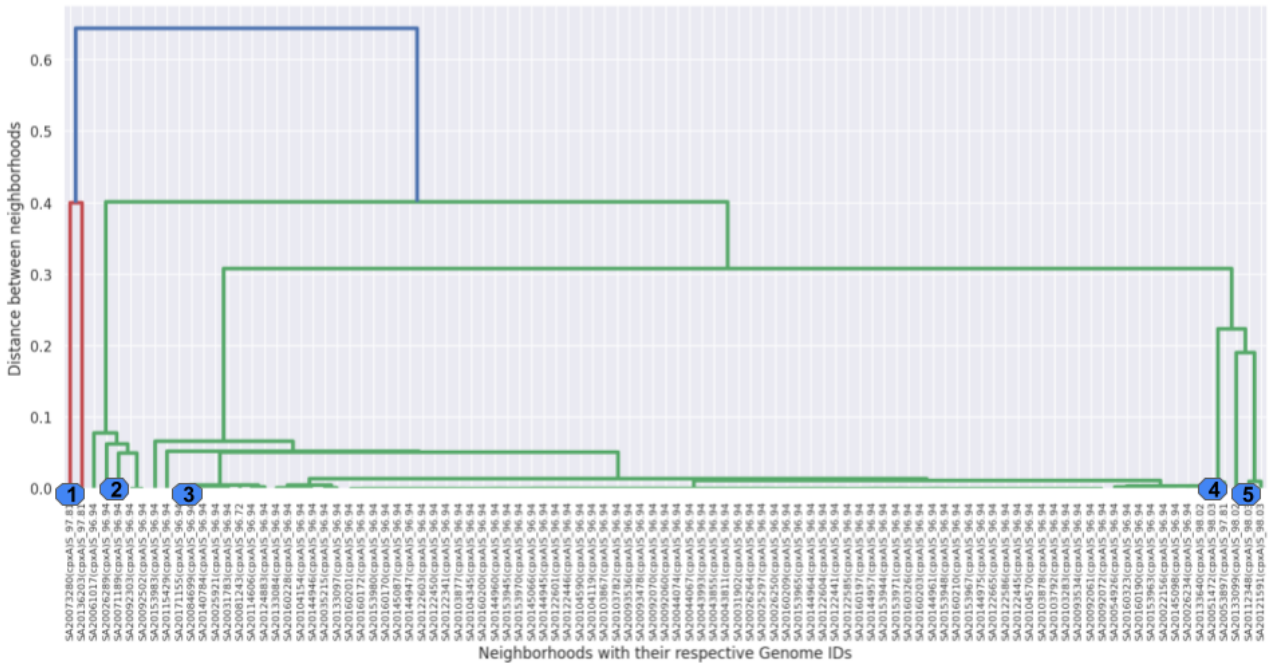

Figure 2.17: A dendrogram generated by UPGMA clustering denoting the key differences in the 100 neighborhoods of the *cpxA* AMR gene model of Dataset 2.2. One neighborhood from the red cluster and 4 representatives from the larger green cluster are chosen to visualize separately (blue circles with numbers). The x-axis shows the genome IDs of neighborhoods belonging to various clusters (red and green) as labels. The y-axis shows the distance between the neighborhoods at the time they were clustered.

Figure 2.18 shows each representative neighborhood from the two clusters obtained when the dendrogram was cut at a height of mean of the distance matrix. The neighborhood 1 which belongs to red cluster and 2, 3, 4 and 5 of the larger green cluster were chosen as representatives to visualize. Six genes in upstream [*cpxR*, *cpxP*, *fieF*, *pfkA*, *sbp*, *cdh*] of neighborhoods 3, 4 and 5 and four genes in downstream [*yjiM*, *sodA*, *rhaT*, *rhaR*] in the neighborhoods of 2, 3, 4 and 5 were conserved amongst members of the green cluster. Due to these large number of common genes that perform similar functions, the neighborhoods 2, 3, 4 and 5 are grouped into the same green cluster. The upstream of neighborhood 2 is quiet different from others in the cluster with 7 unique genes which includes many small unidentified open reading frames. The upstream of neighborhood 4 has an insertion of 3 genes - *tpiA*, *yfiS_1* and

UID in between *cdh* and an unidentified gene at the upstream end which are absent in the other three neighborhoods of the green cluster. The downstream regions of neighborhoods 2 and 3 are almost identical except for a short ORF in between *kdgM_2* and *rhaT* of neighborhood 2. The downstream of neighborhoods 4 and 5 are identical with all the genes conserved in the neighborhood. Due to these differences, the green cluster is further divided into several smaller clusters.

The neighborhood 1 of red cluster shares two genes upstream and four genes downstream in common with all the neighborhoods of the green cluster. Two unique genes *cvaA_1* and *lagD_1* with many short unidentified ORFs in upstream and *dapA_2* in the downstream leads to the decrease in average similarity scores when compared with other neighborhoods. Hence, the neighborhoods of red cluster share less similarity with the neighborhoods of green cluster.

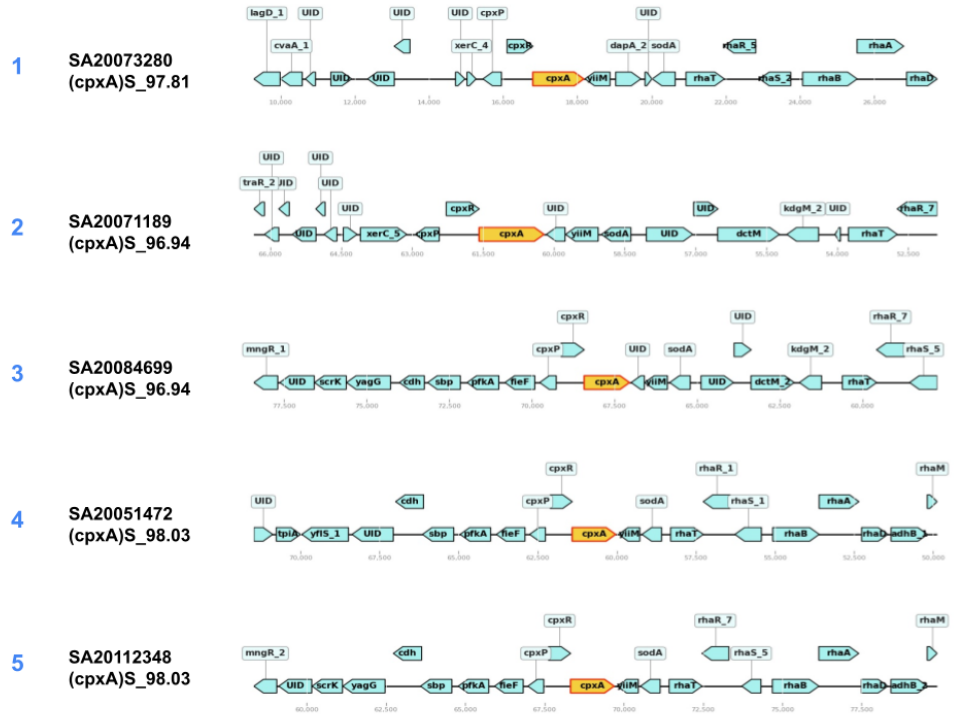

Figure 2.18: Visualization of each major cluster generated by UPGMA for the neighborhoods of the *cpxA* AMR gene model in Dataset 2.2 denoting key differences. Each individual neighborhood is represented with the identifier of the corresponding lineage in the dendrogram followed by the genome ID of the neighborhood. The ID also indicates the name of the AMR gene and the percent identity match with the corresponding gene model.

* The neighborhood analysis of this dataset provided a better understanding of how our methods performed on the large dataset of 100 closely related genomes. These interesting cases showed evidence of different evolutionary events such as insertion and deletion that influenced the AMR gene neighborhoods. According to this analysis, we can expect to find conserved neighborhoods in any dataset that has closely related genomes but the number of interesting cases that show insertions and deletions may vary depending on the properties of genomes of the dataset.

2.4.3 Dataset 2.3: Diverse *Salmonella* of 6 different serovars

When the methods were applied on the genomes belonging to same serovar (Heidelberg), the neighborhoods of many AMR gene models were similar and conserved. Hence, we decided to include genomes belonging to different serovars of *Salmonella enterica* species. Table 2.4 provides a detailed list of various genomes of the dataset with their respective serovars.

Genome ID	**Serovar**
SIDI01000010 SIEV01000010 SIHZ01000010	S.4,5,12:i
SIFG01000010 SIFH01000010 SIIR01000010 SIIV01000010	S.Hadar
SIIA01000010	S.Typhimurium
SIIH01000010	S.Kentucky
SIIK01000010 SIIO01000010 SIIP01000010	S.Heidelberg
SIIX01000010 SIIY01000010 SIWY01000010	S.Enteritidis

Table 2.4: Genomes with their respective serovars

Figure 2.19 shows the distribution of average similarity scores of the neighborhoods of 40 gene models. As per the histogram, almost 57.5% of the total 40 gene models had extremely similar neighborhoods with average similarity scores between 20.5 and 21 and almost 15% of the gene model neighborhoods had very similar neighborhoods with similarity scores between 20 to 20.5 and only 10% showed variations with average similarity score between 15 and 16.5. For this dataset of 15 genomes with 40 unique gene models, the highest average similarity score observed was 20.97, the lowest score was 15.5 and there were no neighborhoods with scores less than 15.5 indicating higher percentage of conserved neighborhood even though genomes from various serovars were included in the dataset.

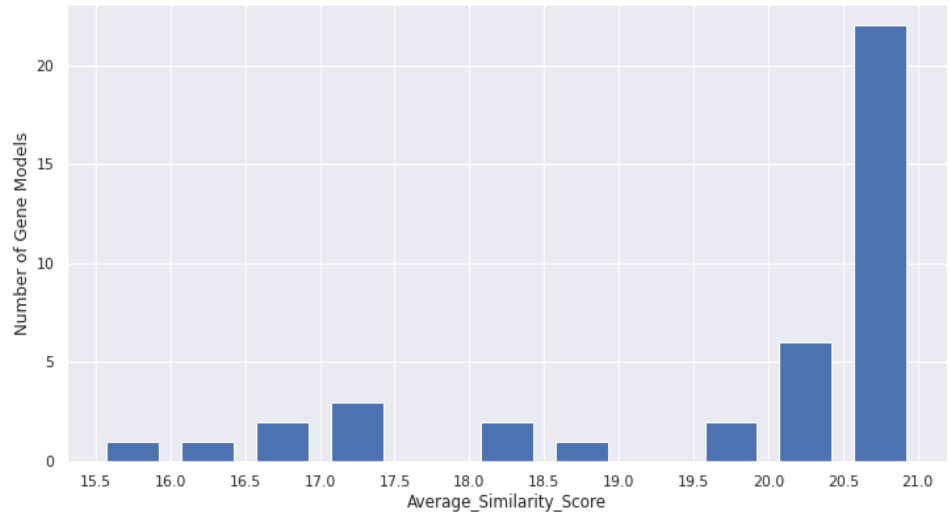

Figure 2.19: Histogram showing distribution of average similarity scores across the neighborhoods of 40 unique AMR gene models identified in 15 genomes of Dataset 2.3. The x-axis shows the number of gene models and y-axis indicates the average similarity scores between neighborhoods.

The gene neighborhood of the AMR gene model "*Escherichia coli UhpT* with mutation conferring resistance to fosfomycin (condensed as *Ec_UhpT*)" that exhibits antibiotic target alteration against fosfomycin antibiotics showed the average similarity score of 16.4 amongst 15 neighborhoods. *Ec_UhpT* showed the maximum differences between neighborhoods for this dataset. Figure 2.20 represents the dendrogram of clusters obtained from UPGMA clustering for the neighborhoods of *Ec_UhpT* AMR gene model where 15 neighborhoods were divided into three major clusters with a maximum cluster difference of 0.35 indicating that there are differences in the neighborhoods.

Figure 2.21 shows the visualization of one neighborhood from the red cluster and two from the blue cluster that are closely analyzed to find the key differences. All ten upstream genes are completely conserved between one neighborhood of the red cluster and two neighborhoods of the blue cluster indicating major similarities between the three. Neighborhood 1 of the red cluster has 5 unique downstream genes *nepl*, *gmuD*, *dgaR_2*, *levE* and *agac_2* and many unidentified short ORFs that are absent in the other two neighborhoods. These gene differences cause the division

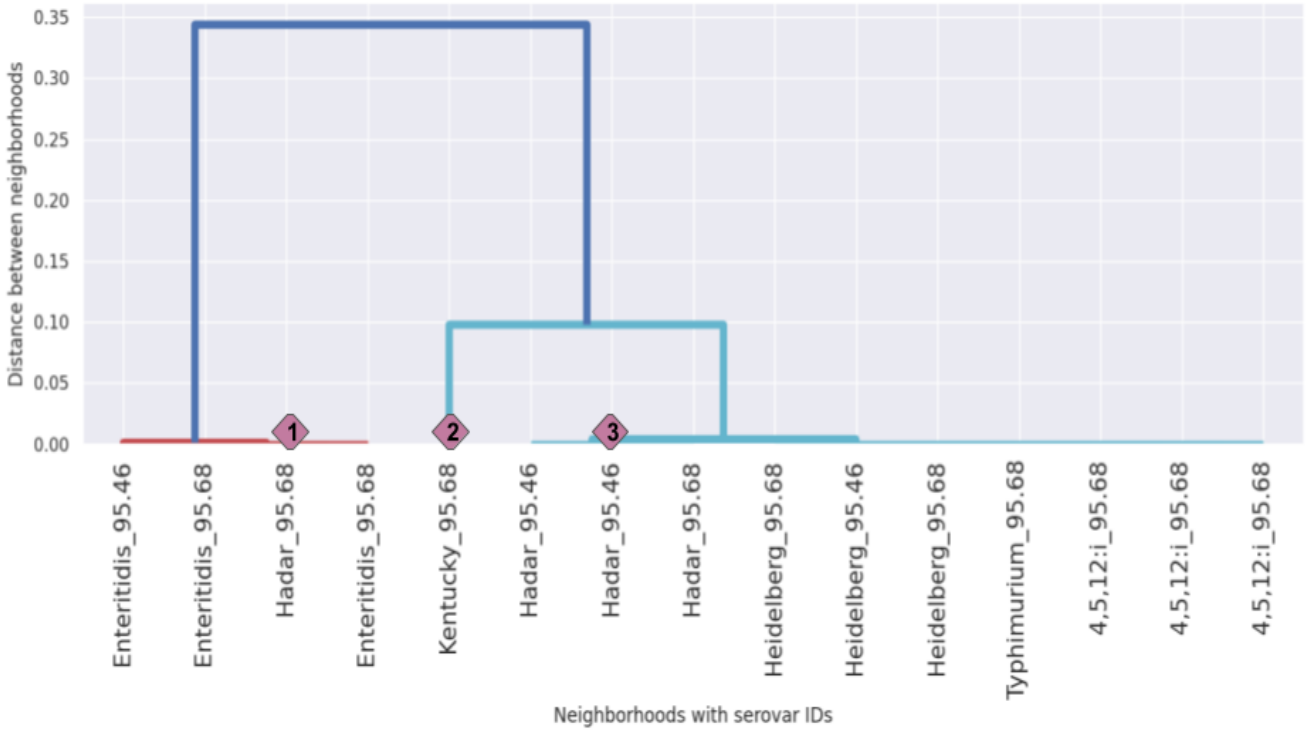

Figure 2.20: A dendrogram of clusters generated by UPGMA clustering method denoting the differences in the 15 neighborhoods of *Ec_UhpT* AMR gene model of Dataset 2.3. Three major neighborhoods from clusters obtained when the dendrogram was cut at mean value are numbered to visualize separately (pink diamonds with numbers).The x-axis shows the serovar IDs of neighborhoods belonging to various clusters (red and blue) as labels. The y-axis shows the distance between the neighborhoods at the time they were clustered.

between red and blue cluster at a cluster height of 0.35 in the dendrogram. The downstream of neighborhoods 2 and 3 of blue cluster are almost identical with genes of similar functions except an insertion of gene *ptsH_2* (8 places right from *Ec_UhpT*) indicating that there are only minor gene differences between the neighborhoods.

This AMR gene model was an interesting case to analyze. The neighborhood 1 - SIFG01000010 belonged to the serovar - Hadar (Table 2.4) but according to UPGMA this neighborhood belonged to the red cluster which is otherwise comprised only of the neighborhoods of S.Enteritidis. All the other genome neighborhoods of the

Hadar serovar were grouped as a part of the blue cluster. This difference suggests the possibility of lateral gene transfer among the genomes of the Hadar serovar.

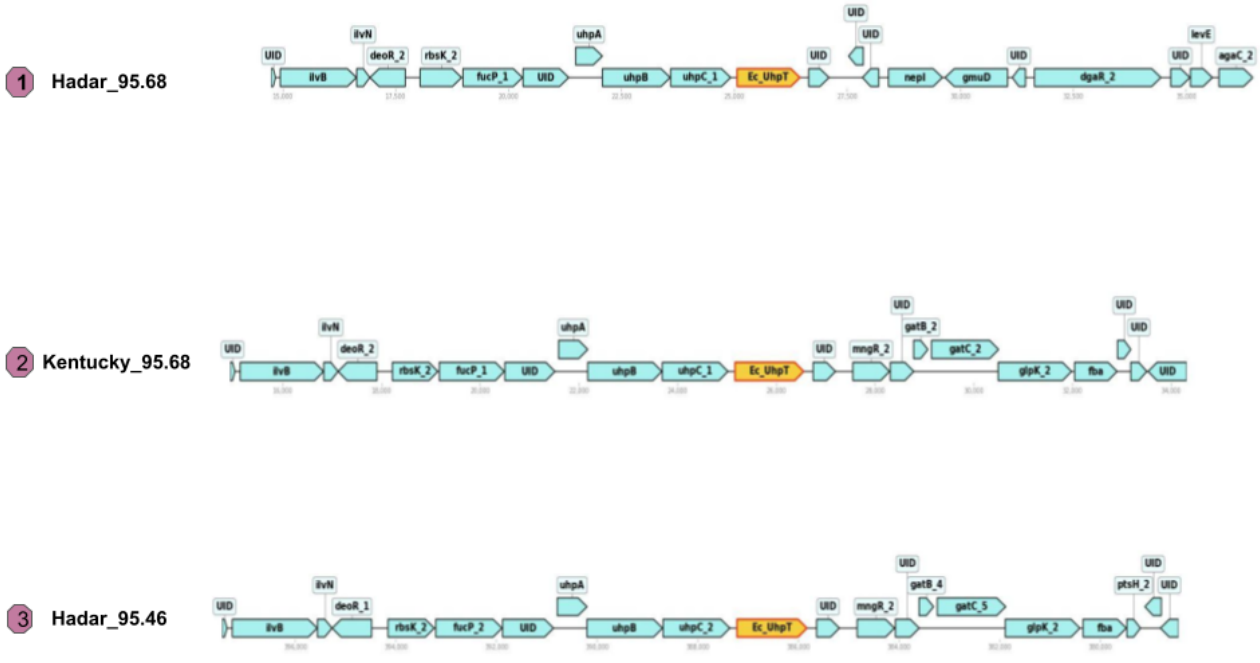

Figure 2.21: Visualization of each major cluster generated by UPGMA for the neighborhoods of *Ec_UhpT* AMR gene model in Dataset 2.3 denoting key differences. One neighborhood from each cluster is represented with pink diamond followed by the serovar ID of the corresponding neighborhood. The ID also indicates the name of the serovar and the percent identity match with the corresponding gene model.

As this dataset consisted 15 genomes of 6 different serovars, we further analyzed the clusters of gene models to find interesting cases. Figure 2.22 summarizes the overall statistics of the variation in maximum differences between the neighborhoods of all the gene models. According to the histogram, around 50% of the total 40 gene models had a maximum neighborhood difference between 0 to 0.05 indicating that neighborhoods are highly conserved even when different serovars were included. Interestingly, around 5% of the gene models had a maximum value between 0.4 to 0.45 indicating the existence of dissimilar neighborhoods even though the conservation was maintained.

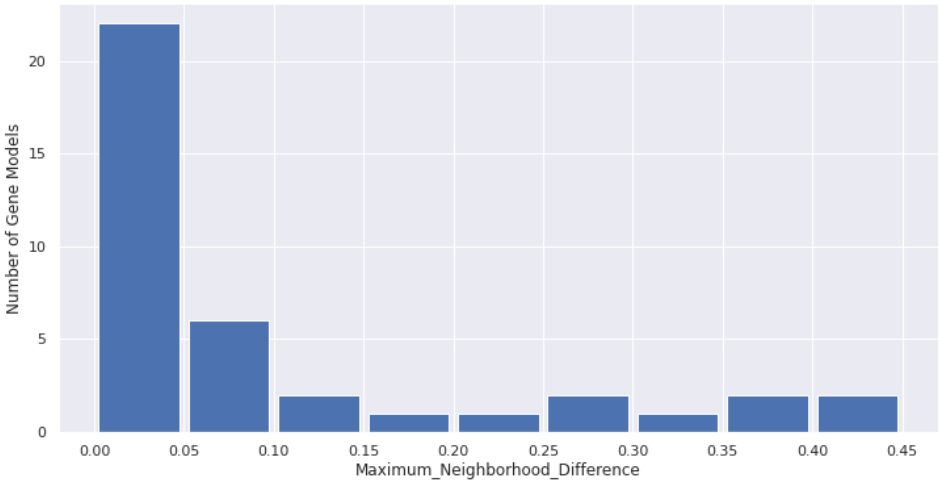

Figure 2.22: Histogram showing distribution of maximum differences across the neighborhoods of 40 unique AMR gene models identified across 15 genomes of Dataset 2.3. The x-axis shows the number of gene models and y-axis indicates the maximum difference between neighborhoods of each model.

We analyzed the very similar neighborhoods with maximum difference less than 0.05 to identify cases that showed perfect distinction between the neighborhoods based on their serovars. The neighborhood of the AMR gene *mdsB* that exhibits multidrug resistance against monobactam; carbapenem; cephalosporin; cephamycin; penam; phenicol and penem antibiotics was observed to have a highly conserved neighborhood of maximum difference of 0.007 and average similarity score of 20.7 which indicates the neighborhoods were identical with very small bit-score differences. Figure 2.23 shows the dendrogram generated for the 15 neighborhoods of AMR gene model *mdsB* which shows that even though all the 15 neighborhoods were extremely similar with a maximum difference of 0.007, the clusters showed notable distinction between the serovars except the neighborhoods of Hadar. The clusters in the dendrogram proved that UPGMA clustering captures even small differences in bit-scores which can be very useful in identifying variations in the dataset of genomes that has gene differences based on their serovars. We used the cluster information from the dendrogram (2.21. As we expect the neighborhoods of same serovars to posses similar genes, these cluster and gene order differences provides evidence that LGT between the genomes of same serovar has occurred.

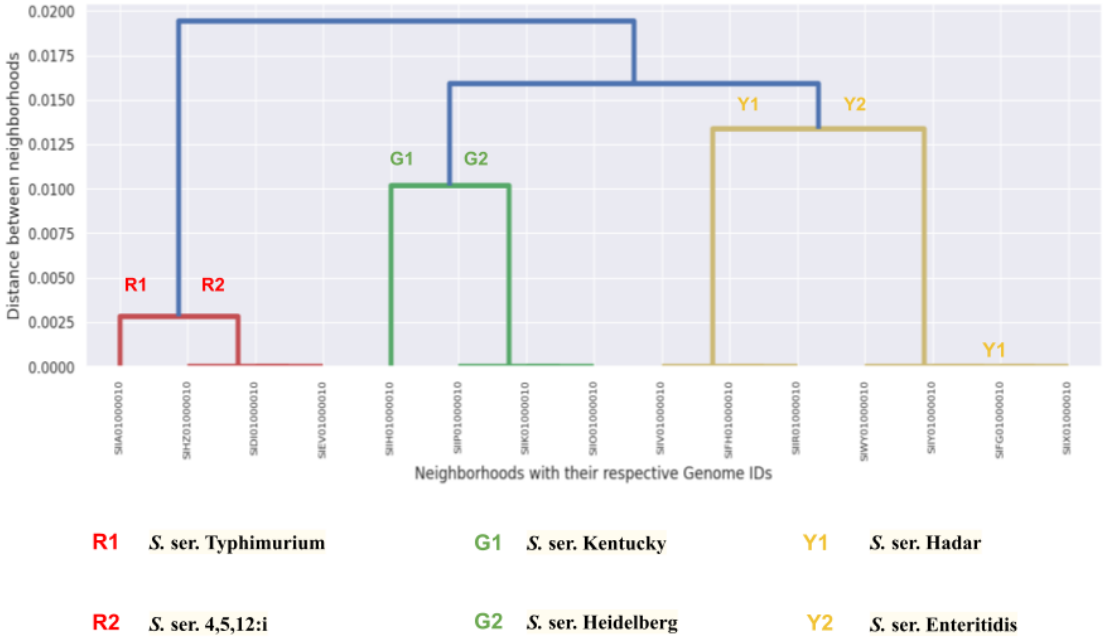

Figure 2.23: A dendrogram of clusters generated by UPGMA clustering denoting the differences in the neighborhood of *mdsB* AMR gene model of Dataset 2.3. Each major cluster visible (red, green and yellow) is highlighted and the corresponding serovar is mentioned below.

Chapter 3

INCLUSION AND ANALYSIS OF LOOSE AMR HITS

3.1 Introduction

The rise in global resistance has overshadowed the success of anti-bacterial drug discovery. The prevalence of AMR bacteria leads to the need for the development of advanced methods to identify AMR genes in bacterial pathogens. Databases such as CARD can be used to predict AMR genes in bacterial genomes. RGI which uses CARD as its central database, categorizes its predictions into three major types of hits. Even though Strict and Perfect hits are important, the analysis of Loose hits holds the potential to allow detection of homologs of AMR genes that correspond to previously uncharacterized genes and emerging threats [3]. In this section, we concentrate mainly on these Loose hits to understand whether a candidate Loose AMR gene really posseses resistant properties and also to observe if related neighborhoods can provide additional evidence for AMR functions of candidate CARD hits.

3.1.1 Uncertainty in prediction of AMR genes

Genes may be predicted to encode resistance by RGI or other methods, but there is a risk of false-positive predictions. Most of the conventional strategies to identify the genetically encoded AMR perform well while predicting previously known and conserved AMR genes, but tend to produce an inadmissible number of false positives if more-divergent sequences (sequences from distantly related taxa) are included [26].

There are at least 47 open-source bioinformatic tools developed for the purpose of detecting AMR determinants in DNA or amino acid sequence to date, including ResFinder, KmerResistance, MEGARes, CARD RGI, ANNOT, SRST2, Genefinder, ARIBA, and AMRFinder [52]. When the performance of public databases CARD and ResFinder was evaluated on 2587 isolates across five clinically relevant pathogens, the rate of false-positive results was higher in CARD (42.68%) than ResFinder (25.06%). The presence of AMR determinants such as efflux pumps caused erroneous predictions

since they are affected by gene regulation [79]. This study concluded that there is a requirement for further expansion of the AMR databases by improving the marker annotations such as (i) efflux-related AMR genes and mutations; (ii) aminoglycoside-modifying enzymes; and (iii) fluoroquinolone resistance-associated genes and mutations, per antibiotic rather than per antibiotic class (aminoglycosides, penams, tetracyclines, cephalosporins and other β- lactams, quinolones and aminoglycosides). Although numerous AMR predictive tools have been developed, most of these tools produce a certain number of false positive predictions. Hence, in-depth analysis of such predictions can provide additional insights into their functions and a better understanding to whether these false predictions contribute to AMR spread and transmission.

Recently, many studies have also started to use various machine-learning models for predicting AMR phenotypes. Most machine-learning based models utilize whole-genome sequences and extensively researched sets of AMR genes to predict AMR phenotypes. When the complete genome or the complete set of AMR genes from a genome is not available, the performance of the model deteriorates. Models that are built using all genes, rather than just those implicated in AMR, are more robust to incomplete data [88]. Bacteria contain putative genes that are likely to encode proteins but have no known function. These genes can share sequence similarities to already identified genes and thus can be inferred to share a similar function, but still the exact function of putative genes remains unknown [83]. Many machine-learning algorithms have also been developed to predict such putative AMR genes. These models use various characteristics of proteins such as regions associated with single-nucleotide polymorphisms (SNP) that are unique to AMR genes as features instead of using sequence similarity. These methods may fail when a feature-selection strategy that removes irrelevant and redundant features is not employed as these redundant features might compromise the accuracy of a machine-learning model [27] [6].

Several factors such as input data format, presence/absence of software for search within a database of AMR determinants, and also the search approach used for alignment or mapping have to be carefully considered while choosing a resource. Due to the limitations that arise while choosing a specific tool based on its sensitivity and specificity to predict AMR genes, only few of them have been emphasized in scientific

and research articles.

3.1.2 Phylogenetic and functional inference of orthologs

Although genome sequencing itself has become a routine task and identification of protein-coding genes has become more reliable, the methods that automatically assign functional roles to genes lack accuracy and sensitivity [101]. The performance of the function prediction methods can be enhanced by concentrating on the factors contributing to the existence of sequence similarity such as evolution [37]. In order to achieve more accurate and efficient results, integrating multiple lines of information such as phylogenetic relationships and sequence alignment is necessary. CAFA evaluates ensemble learning methods that show such integration of information [58]. Evolutionary relationships among various groups of species can be represented using phylogenetic trees.

The probability of a gene function changing after a duplication event is more than the probability of gene function changing after a speciation event because the sequence and function evolve in parallel. The study [39] states that this relation between sequence and function evolution is the underlying idea of many phylogenetic function annotation methods. The steps followed by most of the evolutionary based approaches for constructing a phylogenetic tree are: i) Homology assessment and multiple sequence alignment; ii) phylogenetic analysis using simple methods such as neighbor joining or maximum parsimony to obtain an initial distance based tree; iii) choosing the model based approaches best suitable for the data (a maximum likelihood tree; and iv) tree visualization [111]. Functional similarity can be inferred using phylogenetic trees which depict the pattern of evolution of a set of proteins [113].

3.2 Objective

In this chapter, we conducted a specific analysis to examine Loose RGI hits more carefully to see if related neighborhoods can provide additional evidence for AMR functions of candidate CARD hits. In particular, we consider whether Loose hits in one genome have homologs in other genomes that are Strict or Perfect matches to the same model, and whether the neighborhoods are similar. Phylogenetic analysis is used to further examine the relationship between Loose hits and their closest matches

in other genomes. We expect to observe interesting cases of AMR neighborhoods with varying degrees of conservation as we include datasets containing genomes of different serovars as well as genomes of distantly related taxa.

We examined two data sets to see if Loose hits could be supported by related neighborhoods. The **15_Salmonella-Diverse** dataset (Dataset 2.3) from chapter 2 was reused. We observed from that results of chapter 2 that gene order among neighborhoods of various AMR gene models was conserved within genomes of the same serovar, hence we decided to include genomes of distantly related species to analyze the similarity among the Loose hit AMR gene model neighborhoods. The second dataset comprised genomes from 5 different species: *Klebsiella pneumoniae*, *Citrobacter*, *Salmonella* Heidelberg, *Enterobacter* and *Escherichia coli*. A total of 100 genomes, 20 genomes of each species were combined together to form this dataset.

3.3 Process workflow

The methods of this chapter follow similar steps as in Chapter 2 with a few important modifications and additions. In the step of "Annotation and comparative analysis", one of the modifications was that the specifications for RGI were changed to obtain the Loose hits from the CARD database by including the option "include_loose" in the RGI command. After the Loose hits were identified and listed in the .txt files using the tabular format, the data was pre-processed by including a bit-score ratio filter to include only those genes that were above a specific threshold. The cut-off criteria were varied for each of the datasets depending upon the size. The bit-score ratio BR was calculated as follows:

$$BR = \frac{B_{CARD} * 100}{B_{PASS}}$$

where B_{CARD} is the bit-score value of match to top hit in CARD and B_{PASS} is the Strict detection bit-score cut-off applied by CARD to the given gene model.

The cut-off threshold BR proved to be essential as the RGI identifies gene hits with percent identity ranging from 20% to 100% which results in more than 4000 AMR gene predictions. By applying the threshold, the gene search for a specific gene model was narrowed and the number of false positives was reduced by almost 95%.

To gain more information regarding the predictions of AMR hits, three strategies

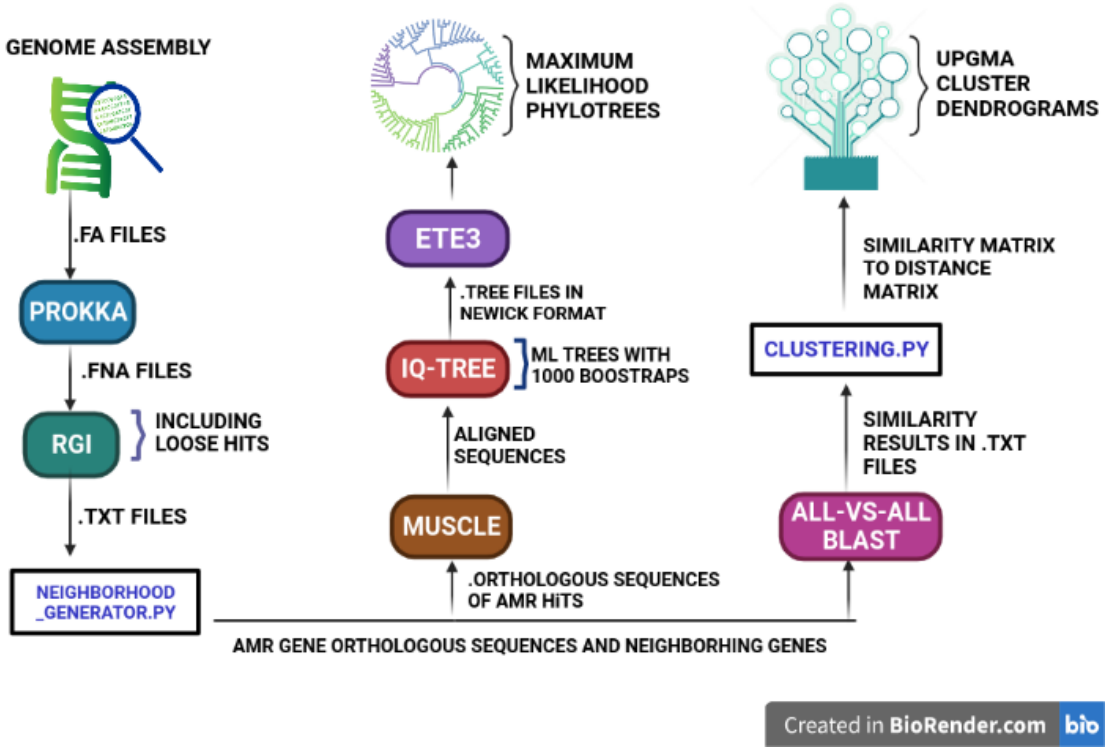

Figure 3.1: Process Workflow - 2: Various steps involved in the approach followed by highlighting bioinformatic tools used (curved rectangles of blue, green and purple), python scripts (.py extended names inside black rectangles) and the results obtained (pictorial representations inside a square). The output at every stage is shown using labelled emerging arrows.

were employed: hierarchical clustering of the similar AMR neighborhoods, phylogenetic analysis of the identified AMR orthologs, and predicting whether a Loose AMR hit confers resistance based on its neighborhood resistance score (NRS) accompanied with gene order visualization. The process of hierarchical clustering followed a similar approach as chapter 2 with the inclusion of Loose AMR CARD hits.

The process of detecting and highlighting the other AMR genes present in the neighborhood of a gene model was an addition to the methods in this chapter. The NRS was calculated based on the total AMR genes present in a neighborhood including the central AMR gene. Each AMR hit was assigned a score based on its cut off criteria: Loose - 1, Strict - 2 and Perfect - 3, with NRS equal to the sum of the individual scores of total AMR genes in the neighborhood. Figure 3.2 shows the visualization of the neighborhood of the *baeS* gene model that has more than

one AMR gene. The NRS of this neighborhood would be 8 as there are 3 Strict hits and 2 Loose hits. This score was helpful in deciding whether a candidate Loose hit conferred resistance.

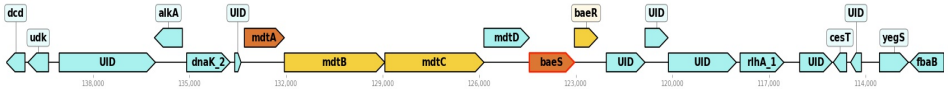

Figure 3.2: AMR genes in the neighborhood of a gene with a Loose match to the *baeS* AMR gene model. The central AMR gene *baeS* was highlighted using the red border, with other Strict and Loose hits highlighted in yellow and orange respectively.

3.3.1 Phylogenetic Analysis

Along with clustering, the phylogenetic tree approach was also implemented to gain more insights on the Loose hits. The initial steps followed while constructing the trees are :

1. The Locus ID and sequence were separated from the input of FASTA file which consists of the sequences of all the orthologs identified for a specific gene model.

2. The list of all taxa that share the same sequences was identified by comparing the sequence similarity between orthologous sequences. The percent identity between sequences was obtained by calculating the Levenshtein ratio.

3. Progressive sequence alignment was used to construct the multiple alignments of the sequences from the obtained list.

4. The clean sequences were aligned with the MUSCLE alignment tool, version v3.8.1551 [34] that uses the sum-of-pairs (SP) score - the sum over pairs of sequences of their alignment scores.

There are two categories of methods used to calculate phylogenetic trees: i) distance-matrix/clustering/algorithmic methods (e.g. UPGMA, neighbor-joining), ii) discrete data/tree searching methods (e.g. maximum likelihood, parsimony, Bayesian methods) [95]. The aligned sequences from MUSCLE were provided to IQ-TREE which is a fast and effective algorithm to infer phylogenetic trees that uses maximum likelihood. From multiple alignment sequences generated for each gene, maximum

likelihood trees were created with 1000 bootstrap replicates using IQ-TREE version 1.6.12. IQ-Tree provides many substitution model choices for the user to analyse proteins, DNA and codons. ModelFinder was used to choose the best fit model for our data and the next steps of the analysis were performed based on the chosen model. ModelFinder chooses the best model by comparing the log-likelihoods for various models (http://www.iqtree.org/doc/). The Figtree [102] software was used to visualize the tree files that were produced by IQ-Tree.

The bootstrap is a widely used procedure to assess the support for relationships shown in a phylogenetic tree [43]. Bootstrapping is used to verify whether a given relationship implied by the tree of the tree is robust to changes in the data. This is achieved by resampling the columns of an alignment with replacement in the data, building trees from each of the subsamples and calculating the frequency with which the various parts of the tree are reproduced in each of these random subsamples [36]. The lower bootstrap values indicate that the target node failed to be found in less than half of the bootstrap replicates.

3.4 Results and Discussion

3.4.1 Dataset 3.1: 15 *Salmonella* genomes with Loose hits

The methods were first applied on the smaller dataset to evaluate the results. More than 4000 genes per genome were identified by RGI including Loose hits of the CARD. These large number of hits were sorted and filtered based on the bit-score ratio. As the dataset contained only 15 genomes, all the hits with bit-score ratio above 50% were considered as the analysis of even the large number of Loose hit predictions for 15 genomes would be computationally infeasible. For this dataset, 24 unique AMR gene models had matches in more than 5 genomes of the dataset. Variation in the average similarity scores of the models is shown in Figure 3.3. Almost 21% of the total models had extremely similar neighborhoods with a average similarity score between 20 to 20.5. Around 9% of the total models had differences in their neighborhood with an average score between 16 to 16.5. For this dataset of 15 genomes and 24 unique AMR gene models, the highest average bit-score was 20.5 and the lowest was 16. There were no models with average score less than 16 indicating that most of the neighborhoods were highly conserved.

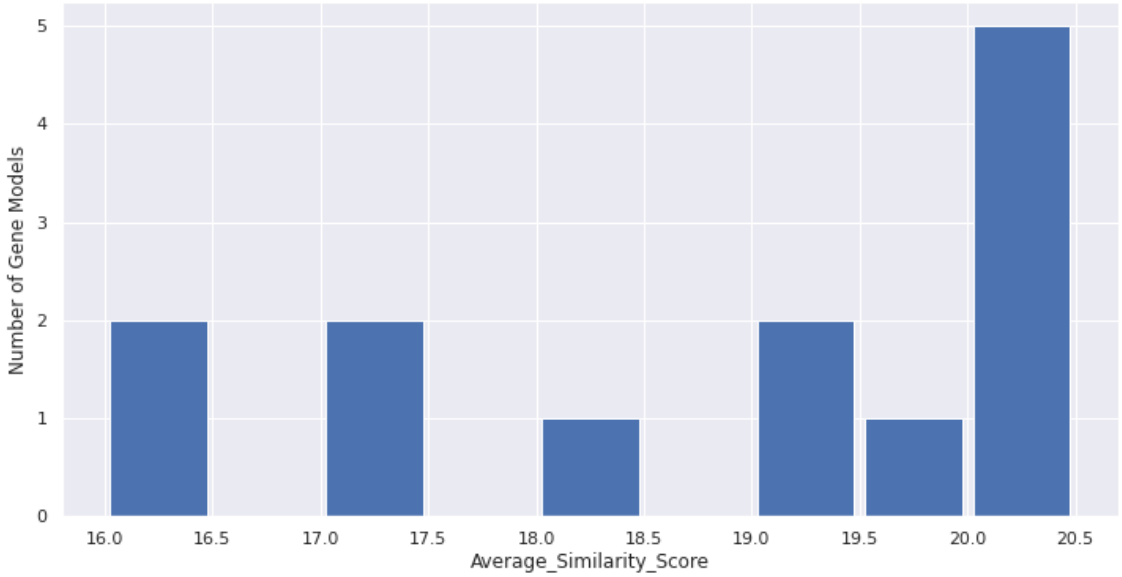

Figure 3.3: Histogram showing distribution of average similarity scores across the neighborhoods of 24 unique AMR gene models identified in 15 genomes of Dataset 3.1. The x-axis shows the number of gene models and y-axis indicates the average similarity scores between neighborhoods.

When the average scores were further analyzed in search of interesting cases and also the models with higher and lower average similarity scores were further investigated. The AMR gene model "*Escherichia coli emrE*" condensed as *Ec_emrE* which exhibits antibiotic efflux resistance mechanism against macrolide antibiotics had the lowest average similarity score of 16. The neighborhood of *Ec_emrE* was comprised of 15 Loose hits with a match of 57.8% percent identity and their neighboring genes. Figure 3.4 shows the dendrogram generated by UPGMA clustering technique for the neighborhoods of *Ec_emrE* gene model. The dendrogram shows that all the 15 genomes grouped into two major clusters (red and purple) with a cluster height of 0.4 that depicts maximum differences in the neighborhoods.

Figure 3.5 shows the 6 neighborhoods belonging to each different serovar included in the dataset that are visualized to understand the differences in their neighborhood when Loose hits of 57.8% identity were considered. The details of genomes with their corresponding are provided in Table 2.4. The red cluster in dendrogram consists of all the neighborhoods belonging to S.Enteritidis and one neighborhood of S.Hadar. The

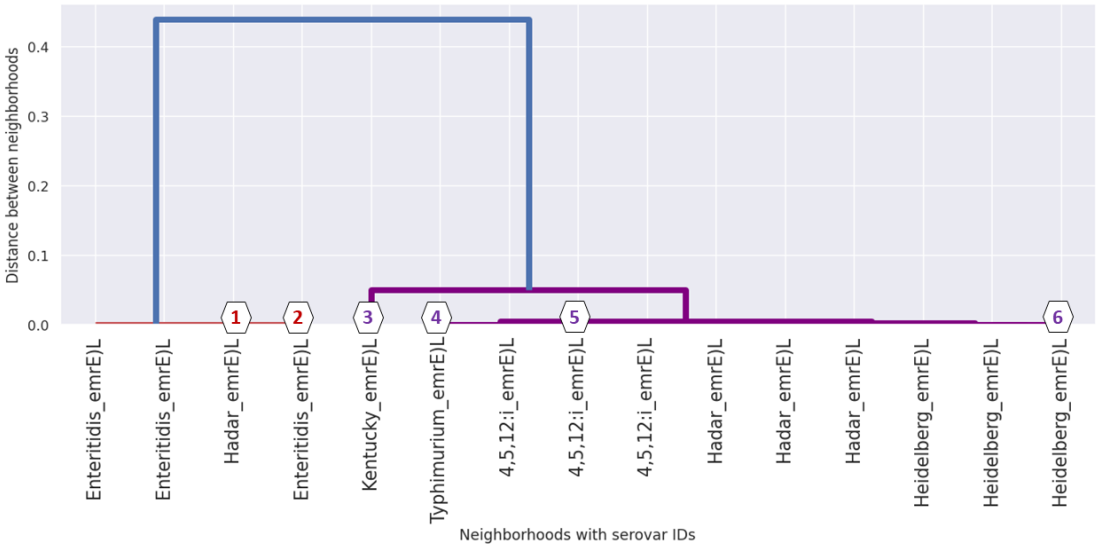

Figure 3.4: A dendrogram representing the clusters generated by UPGMA for the neighborhood of gene model *Ec_emrE*. The x-axis shows the serovar names along with percent identity match with the corresponding gene model belonging to two clusters (red and purple) as labels. The y-axis shows the distance between the neighborhoods at the time they were clustered.

neighborhood belonging to Hadar serovar in red cluster has a distinct set of genes when compared to the other members of the same serovar. As we expect closely related genomes to have similar neighborhoods, this cluster provides evidence of the occurrence of lateral gene transfer among the genomes of Hadar. The neighborhoods 3, 4, 5 and 6 of the purple cluster are very similar with same upstream and downstream genes even though each one belonged to different serovar. Neighborhood 3 has an insertion of an unidentified gene to the left of *Ec_emrE* which does not match with any genes of other neighborhoods, which explains this neighborhood's separation from the rest of the purple cluster. Neighborhoods 2 and 4 have a completely different set of upstream genes, many unidentified, that are responsible for the main split in the dendrogram.

The Loose hit gene models were further analyzed to investigate if there were any neighborhoods that were very similar across all serovars and isolates. The AMR gene model *arnA* that exhibits antibiotic target alteration of peptide antibiotics showed very similar neighborhoods with an average similarity score of 20.5. The neighborhood

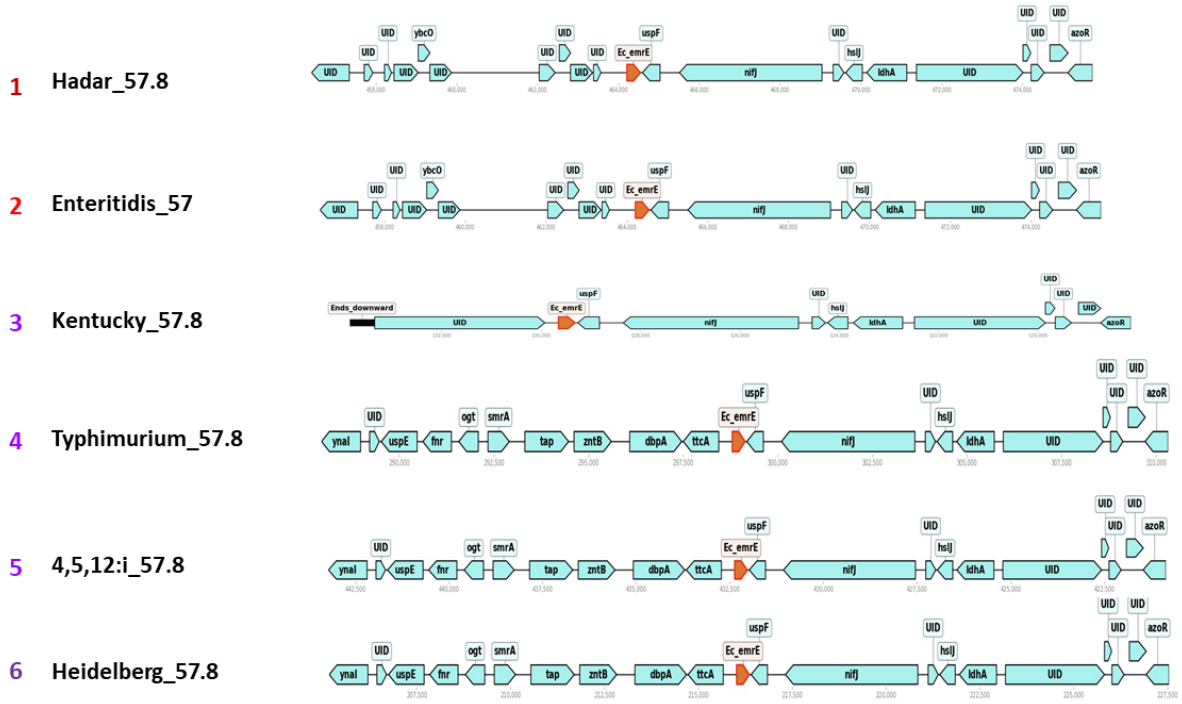

Figure 3.5: Visualization of one neighborhood from each serovar in the clusters generated by UPGMA for the neighborhoods of *Ec_emrE* AMR gene model in Dataset 3.1 denoting key differences. Each individual cluster is represented with a number in the color of the cluster to which it belongs followed by the genome ID of the neighborhood. The ID also indicates the name of the AMR gene and the percent identity match with the corresponding gene model.

of *arnA* was an interesting case for two reasons: first, it exhibits a high degree of similarity even though being a Loose hit of 68% identity range; second, a Strict hit *arnC* was found next to *arnA* (left) in all the neighborhoods. Figure 3.6 shows the dendrogram generated by UPGMA for the neighborhoods of *arnA*. It is clear from the dendrogram cluster heights that these neighborhoods were extremely similar with the maximum cluster height of 0.010. The dendrogram shows two clusters (blue and red) where the neighborhood of blue cluster belongs to S. Typhimurium whereas all other genomes of different serovars belonged to red cluster which is further divided into smaller cluster which shows each serovar distinction.

Figure 3.7 shows the 6 neighborhoods belonging to each different serovar included in the dataset that are visualized to understand the differences in their neighborhood when Loose hits of 68.62% identity range were considered. The members of both blue

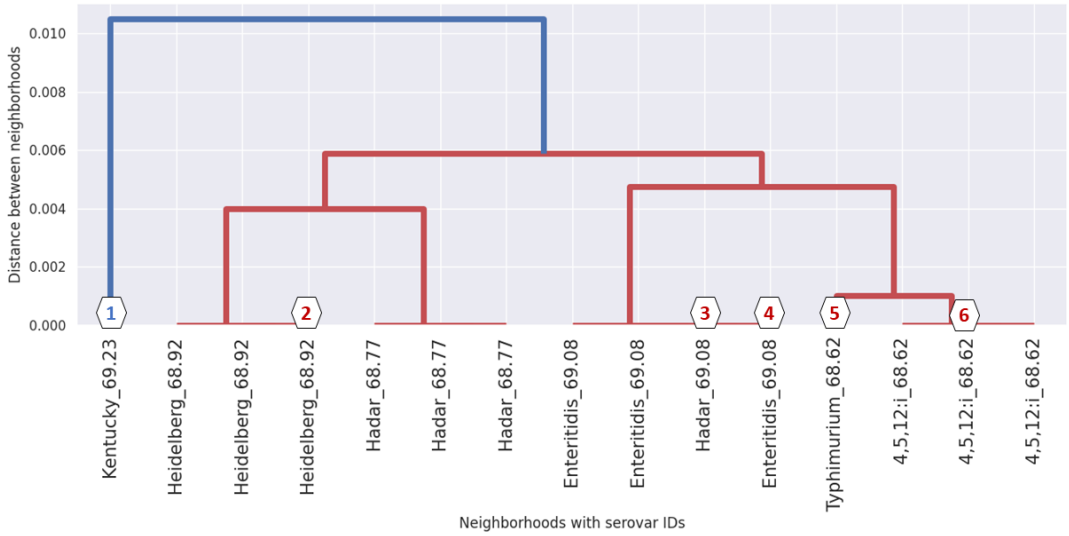

Figure 3.6: A dendrogram representing the clusters generated by UPGMA for the neighborhood of gene model *arnA*. The x-axis shows the serovar names along with percent identity match with the corresponding gene model belonging to two clusters (blue and red) as labels. The y-axis shows the distance between the neighborhoods at the time they were clustered.

and red clusters had identical neighborhoods with only smaller bit-score differences between serovars. The Strict hit *arnC* (represented in yellow) next to *arnA* is present in all the neighborhoods which suggests that neighborhoods are highly conserved even for Loose hits.

The neighborhoods of *Ec_emrE* and *arnA* were two distinct cases. The neighborhoods of *Ec_emrE* showed greater differences in gene content and order, but the lack of other RGI hits in the neighborhoods provides no evidence that suggests this Loose hit might confer to resistance. Conversely, the neighborhood of *arnA* showed very similar neighborhoods and also the consistent presence of Strict hit in the neighborhood provides suggests that this Loose hit may confer resistance.

There were also cases where one or more genomes had multiple Loose hits to the same gene model. One instance of such multiple copy hits was the AMR gene model *mdtG*, which confers resistance to fosfomycin antibiotics via efflux, showed variations in neighborhood content.

Two Loose hits of AMR gene *mdtG* with an average identity of 90% and 60%

Figure 3.7: Visualization of one neighborhood from each serovar in the clusters generated by UPGMA for the neighborhoods of *arnA* AMR gene model in Dataset 3.1 denoting key differences. Each individual cluster is represented with a number in the color of the cluster to which it belongs followed by the genome ID of the neighborhood. The ID also indicates the name of the AMR gene and the percent identity match with the corresponding gene model.

were identified in all 15 genomes of the *Salmonella* dataset. These neighborhoods were clustered by UPGMA and the dendrogram obtained is shown in 3.8. The dendrogram shows two major clusters (red and green) where cluster 1 consists of the 15 neighborhoods with 90% identity to the CARD model, and cluster 2 consists of all 15 neighborhoods with a 60% identity match. The neighborhood of cluster 1 showed high similarity with almost zero maximum difference in the neighborhoods and the neighborhoods of cluster 2 were further divided into smaller clusters with smaller cluster heights.

Each neighborhood from one serovar belonging to CLUSTER 1 is shown in Figure 3.9. All the neighborhoods of CLUSTER 1 were identical with the same upstream and downstream genes which is consistent with the dendrogram clusters in Figure 3.8. Similarly, each neighborhood from one serovar belonging to CLUSTER 2 are

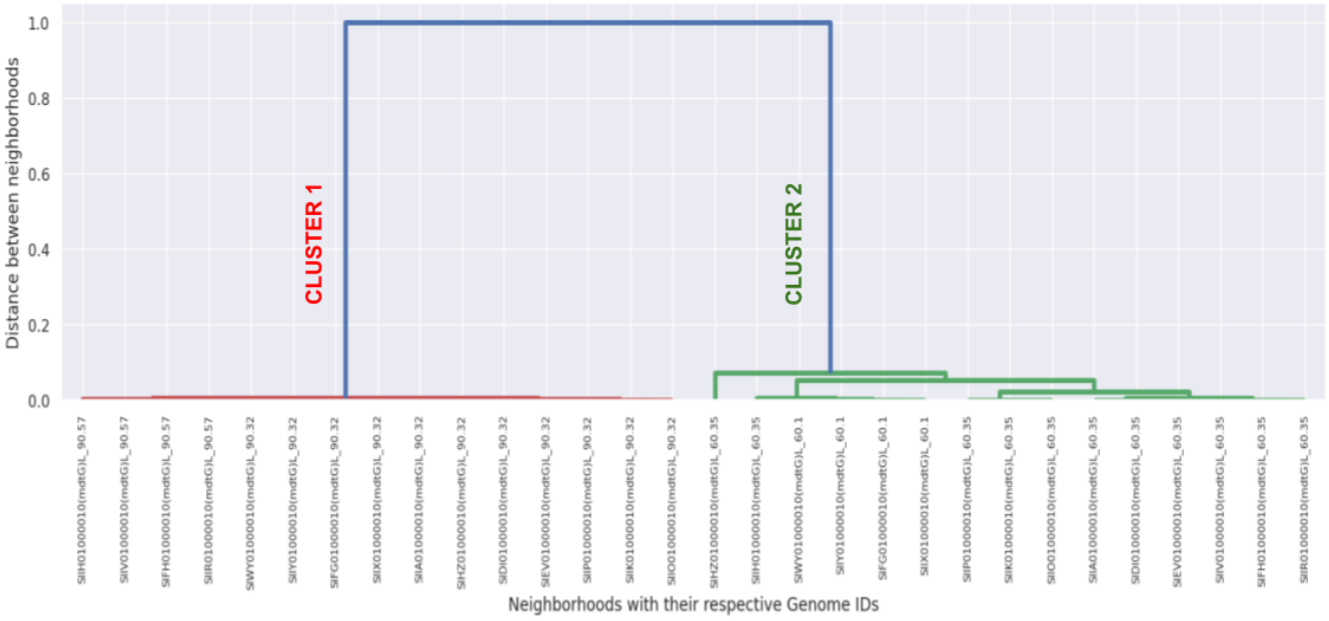

Figure 3.8: A dendrogram representing the clusters generated by UPGMA for the neighborhood of gene model *mdtG*. The x-axis shows the genome IDs of neighborhoods belonging to two clusters (red and green) as labels The y-axis shows the distance between the neighborhoods at the time they were clustered. The two major clusters are represented as CLUSTER 1 and CLUSTER 2 to visualize each cluster separately.

visualized in the Figure 3.10. The neighborhoods of these Loose hits of lower percent identity (60%) were highly conserved and showed greater similarity with same upstream and downstream genes. The major factor was the presence of a Strict hit *ampH* in the upstream (7 genes towards left of *ampH*) in all the neighborhoods at same position. This evidence of Strict hit suggests that this candidate Loose hit might confer resistance and not a mere false positive identified by CARD. Even though the neighborhoods of CLUSTER 1 were very similar and the percent identity of Loose hit was large, there is no significant evidence that the Loose hit of *mdtG* with 90% identity has AMR gene properties.

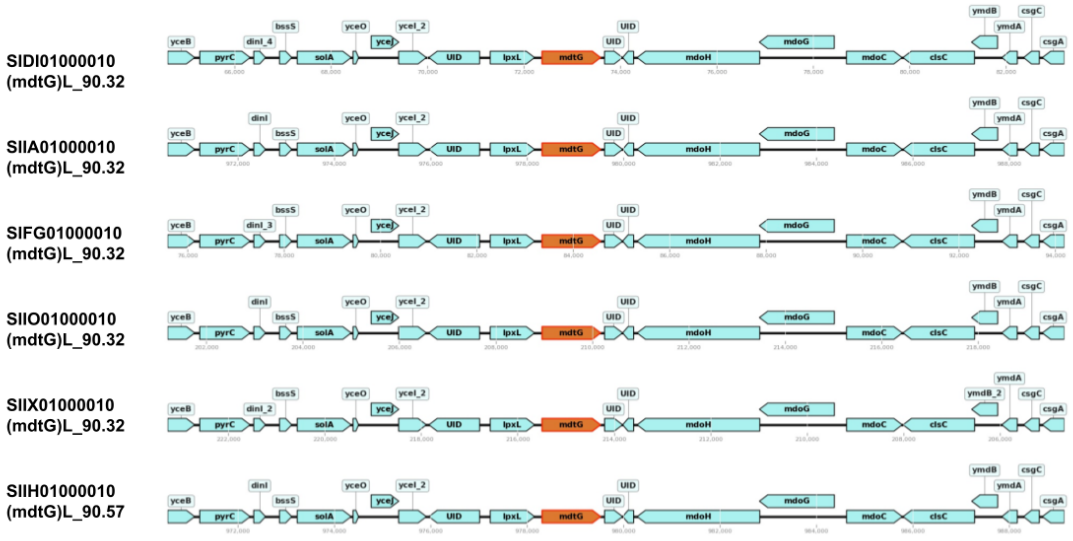

Figure 3.9: Visualization of one neighborhood of Loose hit of 90% identity range from each serovar in the CLUSTER 1 generated by UPGMA for the neighborhoods of *arnA* AMR gene model in Dataset 3.1 denoting key differences. Each individual cluster is represented by the genome ID of the neighborhood. The ID also indicates the name of the AMR gene and the percent identity match with the corresponding gene model.

3.4.2 Dataset 3.2 : 100 genomes of 5 species

After evaluating the results on the smaller dataset by applying methods on closely related species, we observed that gene order and neighborhoods were greatly conserved even when Loose hits of CARD were included. Hence, we decided to test the methods on scaled genomes of distantly related species. This dataset consists of 100 genomes of 5 different genera from the order Enterobacterales. Thirty unique AMR Loose hit models above the threshold were identified. As 20 genomes of each different species were included in this dataset, to ensure that AMR hits with orthologs in at least more than one species were included, we choose specific models that are present in more than 25% of the total genomes of the dataset. As this dataset had 100 genomes, the total percentage of Loose predictions were very high and included many false positives. Hence to narrow down our research to more interesting and probable AMR hits we included only those hits with bit-score ratios greater than 70%.

Figure 3.11 shows the distribution of average neighborhood similarity scores of 30 identified AMR gene models. As per the histogram, almost 60% of the total 30 gene

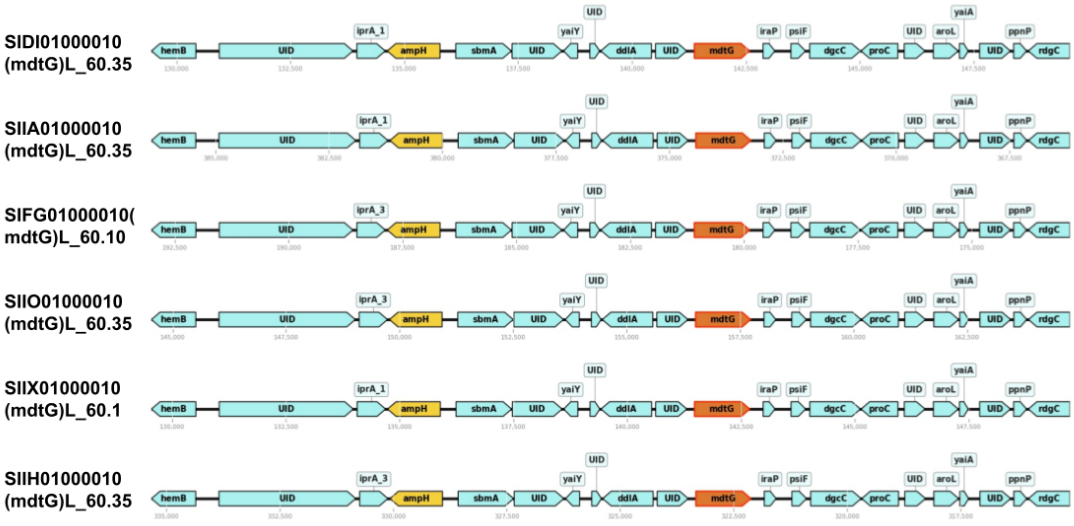

Figure 3.10: Visualization of one neighborhood of Loose hit of 60% identity range from each serovar in the CLUSTER 2 generated by UPGMA for the neighborhoods of *arnA* AMR gene model in Dataset 3.1 denoting key differences. Each individual cluster is represented by the genome ID of the neighborhood. The ID also indicates the name of the AMR gene and the percent identity match with the corresponding gene model.

models showed average similarity scores between 12 and 13 which denotes that the level of similarity and conservation was low among most of the gene models. Also, there were no models with average similarity greater than 14.5 in the dataset. Almost 13% of the gene models had the lowest observed similarity scores between 10.5 to 11 which indicates that these models show greater variations in the neighborhood where more than half of the neighborhood is dissimilar with mismatches. For this dataset of 100 genomes with 30 unique gene models, the highest average similarity score observed was 17.5 and the lowest score was 10.5.

Each model was closely analyzed on the basis of low average similarity scores and high NRS with a goal of finding more substantial evidence to consider that a Loose hit is indeed a true AMR gene. The AMR gene model *baeS* that exhibits antibiotic efflux resistance mechanism against aminocoumarin antibiotics showed very dissimilar neighborhoods with an average similarity score of 12.5 but the neighborhood resistance score was very high.

The gene model *baeS* was found in all 100 genomes belonging to five different species of the dataset. The percent identity of the orthologs identified varied between

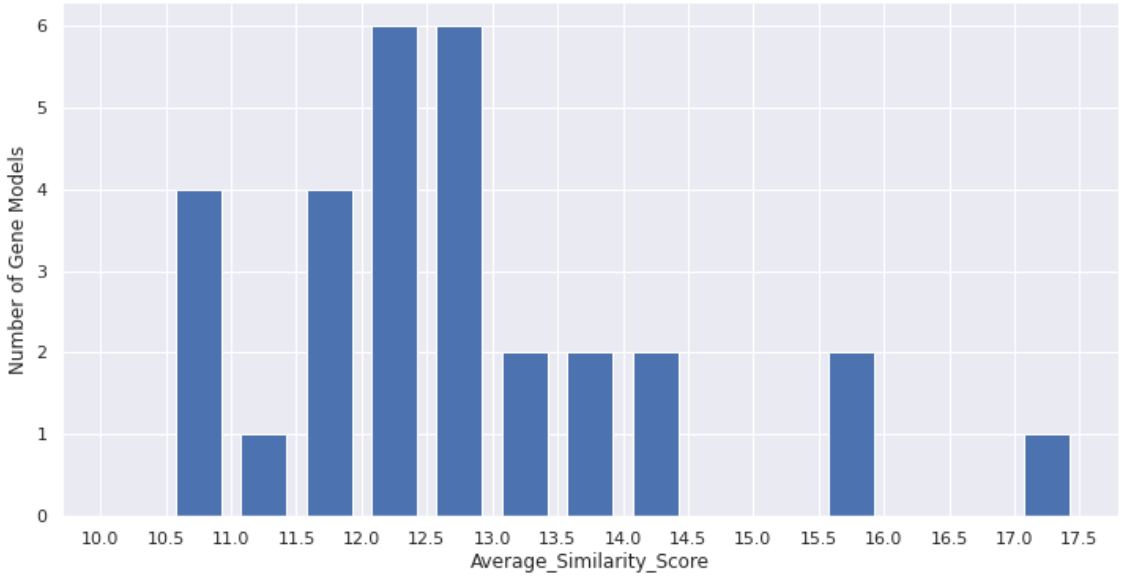

Figure 3.11: Histogram showing distribution of average similarity scores across the neighborhoods of 30 unique AMR gene models identified in 100 genomes of Dataset 3.2. The x-axis shows the number of gene models and y-axis indicates the average similarity scores between neighborhoods.

74.3 to 100% amongst all the genomes. The orthologs of this gene model were found to be a combination of Loose, Strict and Perfect hits. When all the 100 neighborhoods were clustered using UPGMA, the dendrogram obtained is shown in Figure 3.12. The cluster heights in the dendrogram are very high with a value of 0.8 indicating that there are major differences in the neighborhoods. UPGMA divided the 100 neighborhoods into several cluster groups, with some clusters comprising multiple species. For instance, Cluster 9 (labelled pink in the dendrogram) contains a subset of all neighborhoods of *E.coli*, *Enterobacter* and *Citrobacter* grouped together. The similarity of these neighborhoods across subsets of genomes from multiple genera was surprising and suggestive of LGT.

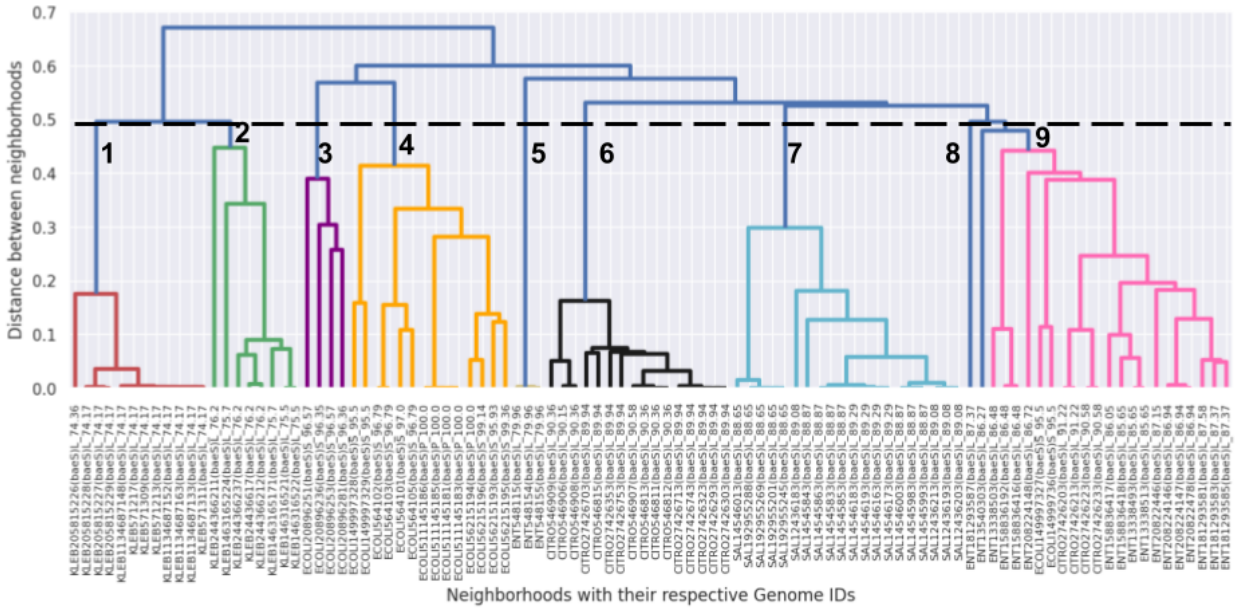

Figure 3.12: A dendrogram representing the clusters generated by UPGMA for the neighborhood of gene model *baeS*. The x-axis shows the genome IDs of neighborhoods belonging to several clusters (red, green, purple, orange, yellow, black, and pink) as labels. The y-axis shows the distance between the neighborhoods at the time they were clustered. The dendrogram is cut at a distance of 0.5 denoted by black dashed line which divides the dendrogram into 9 major clusters represented by bold digits numbered from 1 to 10. The genome ID's start with uppercase letters that indicate the species to which the corresponding genome neighborhood belongs to along with the gene model and percent identity of the match [KLEB-*Klebsiella pneumoniae*, CITRO-*Citrobacter*, SAL-*Salmonella* Heidelberg, ENT-*Enterobacter* and ECOLI-*Escherichia coli*.

The UPGMA dendrogram was cut at a distance of 0.5 to examine and visualize the sample of each cluster. One neighborhood from each of the 9 clusters obtained is visualized in Figure 3.13. An important factor to notice at the first glance of the figure is the presence of the AMR genes *mdtA, mdtB, mdtC, and baeR* in all nine neighborhoods irrespective of their percent identity and their cut-off criterion (Loose, Strict or Perfect). The NRS for these neighborhoods was very high with a maximum score of 15 (5 Perfect AMR genes) which contributes to support for the AMR property of *baeS* gene. The upstream and downstream regions of each neighborhood differ from one another with many insertions, deletions and unidentified

genes. There are many Loose *baeS* matches which were dissimilar from the reference gene model with varying percent identity. The conserved gene order and the evidence of transfer based on clustering relationships shown by UPGMA dendrogram (Figure 3.12) strongly suggests that these Loose hits are indeed real and confer to resistance. Even though the difference between each cluster is very high (dendrogram), high NRS provides substantial evidence that this candidate *baeS* AMR gene has the property of antimicrobial resistance and that it is not merely a false positive identified by CARD.

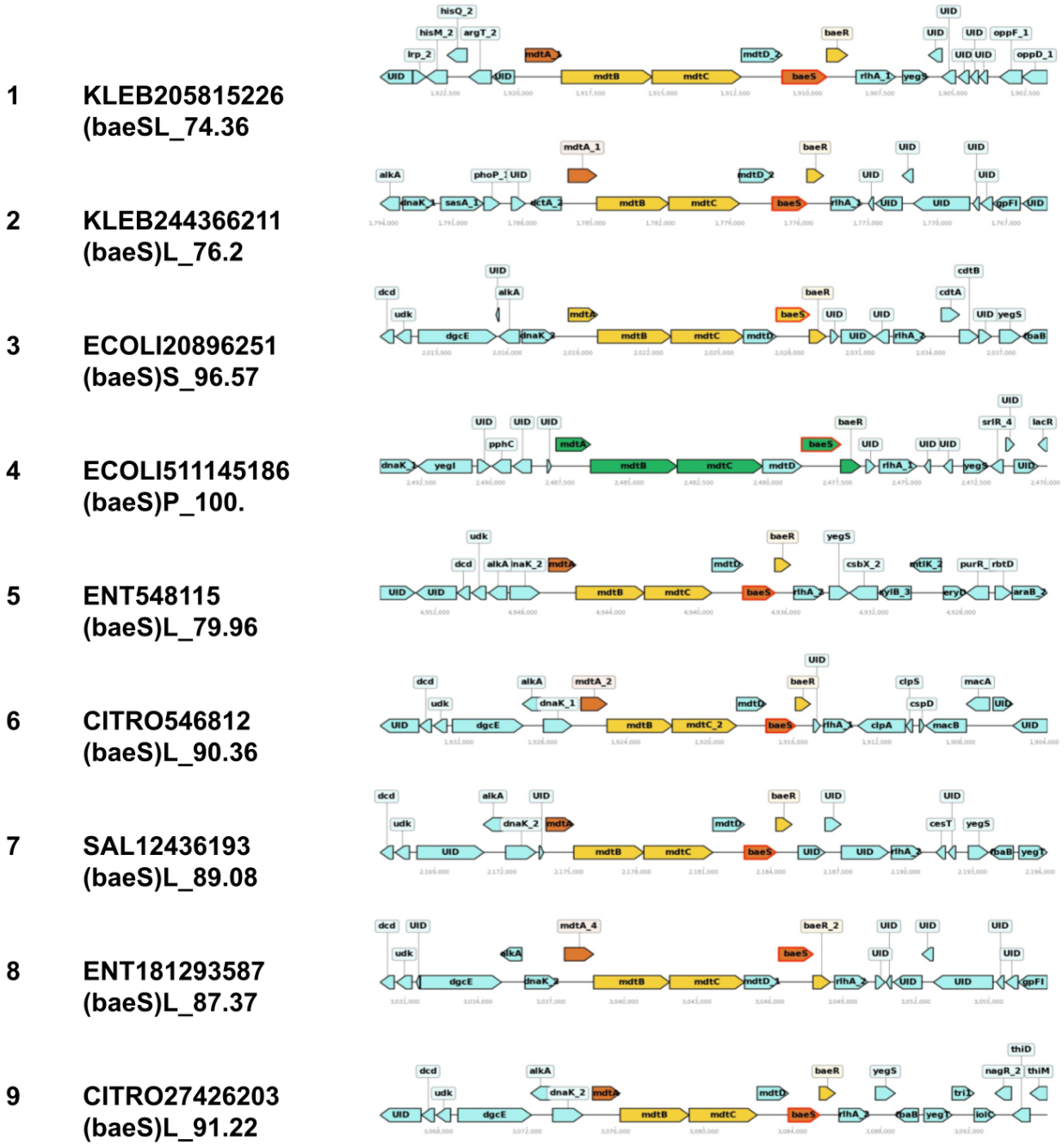

Figure 3.13: Visualization of each representative from labelled cluster obtained for the *baeS* AMR gene model in Dataset 3.2 when the dendrogram was cut at a distance of 0.5 denoting key differences. Each individual neighborhoods are represented with back bold digits of corresponding clusters of the dendrogram followed by the genome ID of the neighborhood. The ID also indicates the name of the AMR gene and the percent identity match with the corresponding gene model.

The phylogenetic tree constructed for the orthologous sequences of *baeS* matches is shown in Figure 3.14. The relationships shown in the phylogenetic tree are largely consistent with those seen in the dendrogram for the *baeS* model: for example, cluster 5 (yellow) which consists of three *Enterobacter* Loose matches with 79.96% identity is consistent with the tree. The large clusters of *Salmonella*, *Citrobacter* and *Escherichia coli* show similarities with the branches of the tree. The nodes of sequences belonging to *Klebsiella* are divided into two separate branches which is consistent with with the red and green clusters of the dendrogram shown in Figure 3.12. While many patterns were conserved, there were also cases where the clusters did not align with the nodes of the tree. For instance, cluster 9 (pink in dendrogram) which consists of neighborhoods from three different species (15 from *Enterobacter*, 4 from *Citrobacter*, and 2 from *E.coli*) was not observed in the branches of the tree, as the *Enterobacter* genomes comprised a group that was distinct from the other genomes. Further analysis of such unique instances can provide more insights into the AMR properties of the *baeS* Loose hit matches in that cluster.

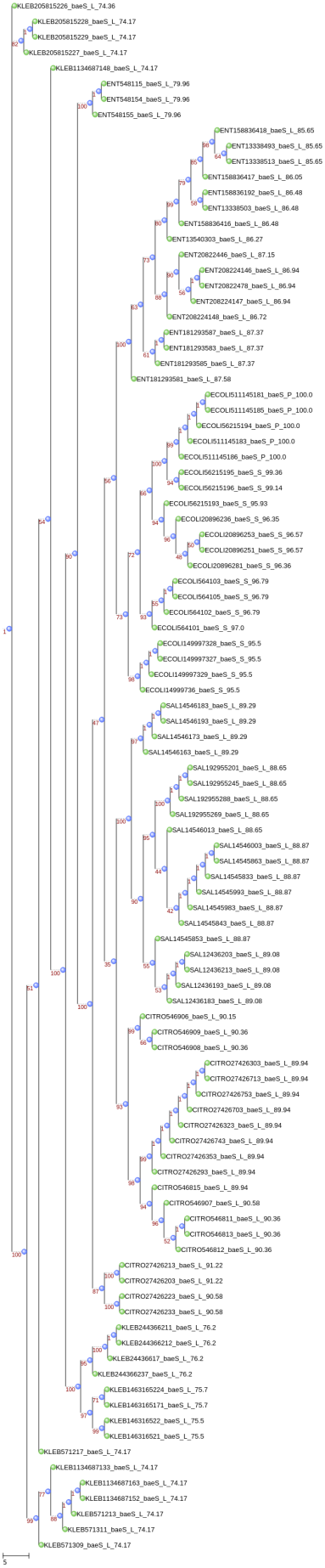

Figure 3.14: The phylogenetic tree constructed from the sequences of 100 orthologous sequences of *baeS* Loose, Strict and Perfect matches. The genome ID of the corresponding sequence are represented as tree labels. The bootstrap values ranging between 1 to 100 for each branch are shown on the tree in red.

Similar to Dataset 3.1, Dataset 3.2 also had AMR models with multiple copies per genome. Analysis of multiple copy gene models was also very informative as it provided better understanding of distribution of neighborhoods of different identities and also to understand which copy might confer resistance based on NRS and upstream and downstream genes. When the multiple copy gene models were analyzed, the AMR gene model of *mdtN* that exhibits antibiotic efflux resistance mechanism against nucleoside antibiotic and acridine dye drug classes showed interesting variations.

The gene model *mdtN* was found in the three genomes of *Citrobacter*, 20 genomes of *E.coli*, 3 genomes of *Enterobacter* and 13 genomes of *Klebsiella*. The neighborhoods were a combination of Loose, Strict and Perfect hits with percent identities varying between 60 to 100%. Most of the neighborhoods had the presence of two or more other AMR hits which increased the NRS and indicating that the corresponding Loose hit central gene will also confer resistance. Figure 3.15 shows the dendrogram generated by UPGMA clustering technique for the total 39 neighborhoods of *mdtN* gene model. UPGMA divides the neighborhoods into two major clusters (red and green) that are further divided into several smaller smaller clusters. The large cluster height of 1.0 indicates that there is very less similarity between neighborhoods. To visualize a sample of the clusters, the dendrogram was cut at a distance of 0.4 to obtain 10 major clusters.

Figure 3.16 shows one neighborhood from each of the 10 clusters obtained by cutting the dendrogram. The visualization is a combination of neighborhoods from 4 different species, which exhibited varied percent identities and CARD confidence levels (Loose, Strict and Perfect). Neighborhood 10 comprises Perfect hits to three consecutive AMR models -*mdtN, mdtO and cusC_1* and this order is conserved in almost all the neighborhoods except neighborhood 7. When these neighborhoods were closely analyzed, the upstream and downstream genes differ in almost all the neighborhoods. Neighborhood 4 has a Loose hit *mdtO* and a Strict hit *cusC_2* next to *mdtN*. The gene *yjcS* which is found in the upstream (two places left of *mdtN*) of almost all the neighborhoods that contain the three AMR matches is a part of multidrug efflux pump "yjcRQP" [114] that is also related with antimicrobial resistance providing more evidence that *mdtN* confers resistance. These AMR genes are

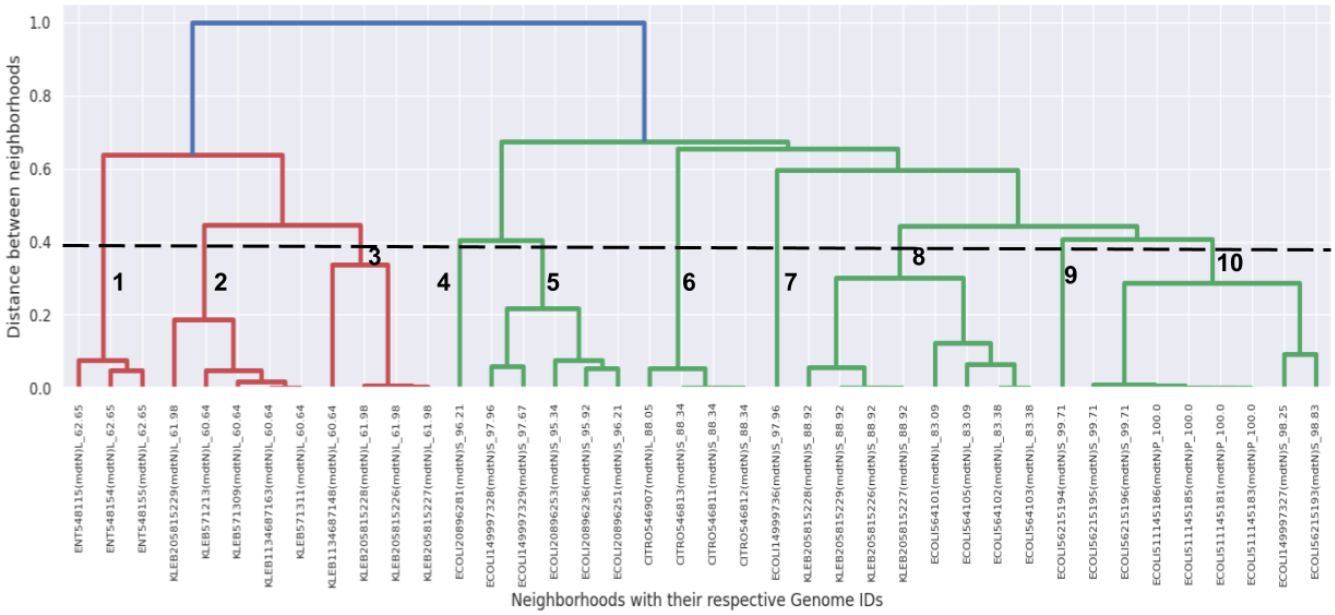

Figure 3.15: A dendrogram representing the clusters generated by UPGMA for the neighborhood of gene model *mdtN*. The x-axis shows the genome IDs of neighborhoods belonging to several clusters (red and green) as labels. The y-axis shows the distance between the neighborhoods at the time they were clustered. The dendrogram is cut at a distance of 0.4 denoted by black dashed line which divides the dendrogram into 10 major clusters represented by bold digits numbered from 1 to 10.

consistently maintained in most of the Strict and Perfect neighborhoods. As the two hits *mdtO* and *cusC* are always present in the neighborhoods of Strict and Perfect *mdtN* genes, their presence in the neighborhoods of a Loose hit of *mdtN* provides evidence that the Loose hit also has the AMR property and exhibits resistance against nucleoside antibiotics.

The phylogenetic tree constructed for the orthologous sequences of *mdtN* is shown in Figure 3.14. Most of the bootstrap values for the branches are greater than 60% which indicates moderate to strong support for the corresponding branches. The red cluster of the dendrogram that contains the neighborhoods of Loose hit matches in *Enterobacter* and *Klebsiella* aligns with the tree where the two nodes of *Enterobacter* Loose hits are split accordingly. Cluster 6, which consists of four Strict hits of *Citrobacter*, is consistent with a branch that has a bootstrap value of 100 for the nodes of orthologous sequences of *Citrobacter*match hits for the *mdtN* model. The concordance of unusual patterns of AMR distribution from both the UPGMA neighborhood clustering and the gene tree lends additional support to the hypotheses that some of the Loose hits are indeed AMR genes, and that gene transfer between species has occurred.

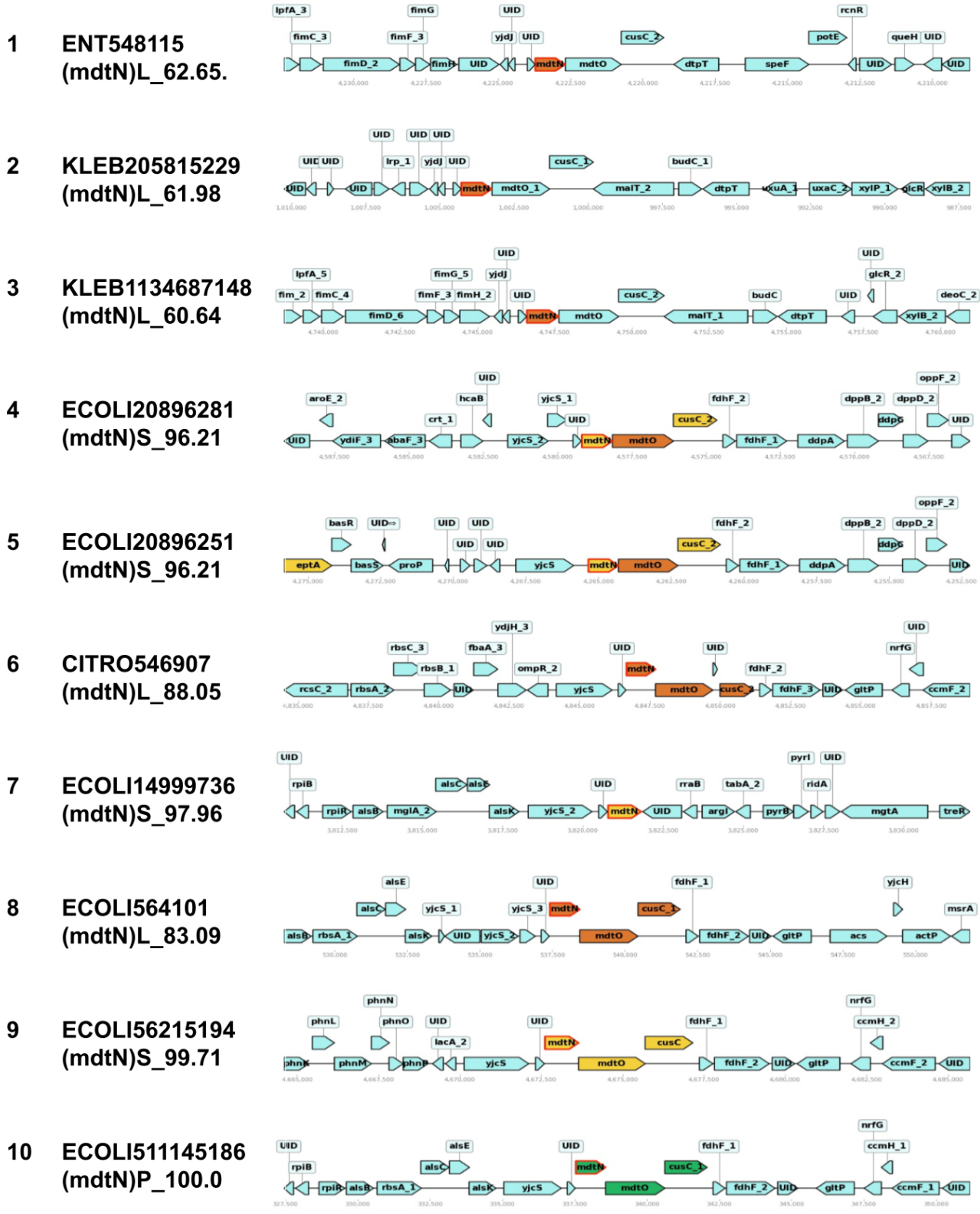

Figure 3.16: Visualization of each representative from labelled cluster obtained for the *mdtN* AMR gene model in Dataset 3.2 when the dendrogram was cut at a distance of 0.4 denoting key differences. Each individual neighborhoods are represented with back bold digits of corresponding clusters of the dendrogram followed by the genome ID of the neighborhood. The ID also indicates the name of the AMR gene and the percent identity match with the corresponding gene model.

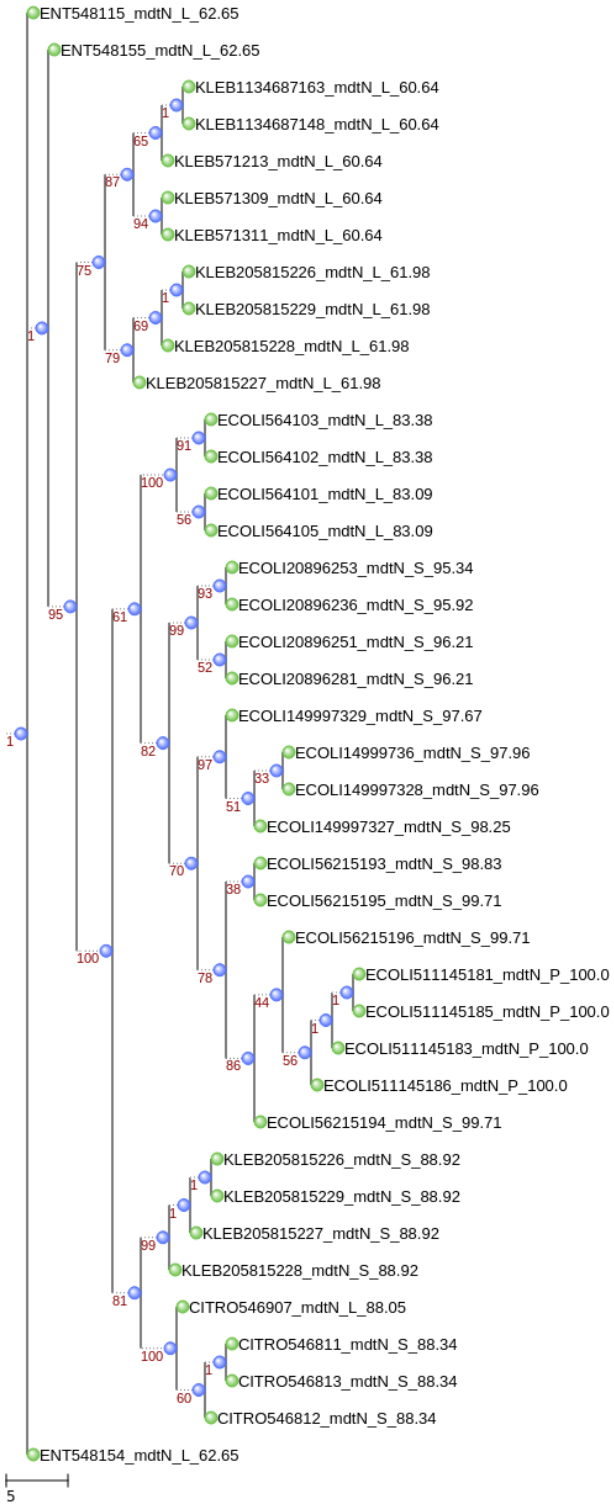

Figure 3.17: The phylogenetic tree constructed from the sequences of 39 orthologous sequences of *mdtN* Loose, Strict and Perfect matches. The genome ID of the corresponding sequence are represented as tree labels. The bootstrap values for each branch are shown on the tree in red.

Chapter 4

CONCLUSION AND FUTURE WORK

Genes that perform similar functions are frequently present close together in bacterial genomes. Analysing gene order among prokaryotic genomes can reveal interesting gene conservation patterns, clusters of similar neighborhoods and allow functional prediction of uncharacterized genes. Although there are many caveats of using gene order information, analyzing the local gene neighborhood of genes provides a deeper understanding of the evolutionary relationships between genomes and detection of potentially interacting proteins [129]. Hence, we designed methods that use gene order among the neighborhoods and were successful in obtaining convincing results that provided important insights on AMR.

In chapter 2, we analyzed more than 130 genomes which belonged to same or different serovars of *Salmonella* and analyzed neighborhoods of more than 100 unique Strict and Perfect gene models. We found that neighborhoods were highly conserved when the neighborhoods of genomes that belonged to the same serovar were considered. This conservation was also generally maintained even when the genomes of different serovars were included. However, there were also several interesting and unique cases where the neighborhoods were dissimilar and very few genes in upstream and downstream were conserved. The key differences in the neighborhoods were due to insertions, deletions and lateral gene transfers which reduced the average similarity between neighborhoods; for example, the neighborhoods of AMR gene model *Ec_UhpT* which showed the evidence of lateral gene transfer among the two genomes of Hadar serovar. Amongst the three clustering techniques that we applied to cluster similar neighborhoods, UPGMA proved to be the best-performing algorithm that generated meaningful clusters which provided clear indication of identical and different neighborhoods in the form of dendrograms. Hierarchical clustering with UPGMA was an effective approach given that neighborhoods such as those of *mdtK* and *cpxA* gene models showed differing degrees of similarity, rather than falling into natural

discrete clusters. Visualizing the gene order between each representative neighborhood allowed us to spot individual gene differences and helped us to interpret the clustering results.

To understand the extent of conservation among neighborhoods when the CARD Loose hits were included, we applied methods on the datasets of genomes with various serovars of *Salmonella* and genomes of five different species. We observed that the neighborhoods of Loose hit gene models were highly conserved within Salmonella, whereas the average similarity drastically decreased when the genomes of different species were analysed. The analysis of Loose hit neighborhoods revealed interesting cases where the candidate CARD Loose hit with lower percent similarity showed stronger evidence of association with AMR. Using a phylogenetic approach to build trees of orthologous sequences provided additional evidence regarding the gene-order conservation between genome neighborhoods of different species. We observed very distinct neighborhoods with low similarity scores when we analyzed the 100 genomes from 5 different species. However, there were interesting cases where the other conserved AMR genes in the immediate neighborhood influenced the resistance factor and the chances of the Loose hit conferring resistance (neighborhoods of *baeS*). We also observed cases where the dendrogram clusters were a combination of neighborhoods from multiple species and the similarity between these neighborhoods was surprising as it showed evidences of LGT (neighborhoods of *mdtN*). The neighborhood resistance score (NRS) accompanied with gene-order visualization assisted in deciding whether a Loose AMR hit is associated with other probable resistance genes or if it is more likely to be a false positive hit by CARD.

This analysis could be used to detect interesting cases of Loose hit neighborhoods with the help of histograms, dendrograms generated by UPGMA and gene-order visualizations. These various results can provide additional evidence where the total resistance factor of neighborhoods could influence the AMR properties of the candidate Loose hit [88]. Databases such as CARD catalogue many homologous sequences and partial hits that may or may not contribute to AMR [3]. Our method analyzes the neighborhoods of such Loose hits by CARD and provides additional insight into these hits.

4.1 Future Work

Putative AMR genes sometimes localized to very short contigs with few or no neighboring genes around the target AMR gene model, and these were not considered for analysis. These short contigs with fewer neighboring genes may lead to problems when the target AMR gene is located on such contigs. We intend to handle this limitation by including a threshold for contig length during the time of annotation of genomes so that each contig contains at least a minimum number of coding sequences. The initial steps of the analysis include running the annotation tool - Prokka and RGI on the genomes. As the GenBank data is very large, the time taken by these tools was very high.

In future, we would like to modify our methods to handle the above-mentioned limitations and improve the efficiency. One of the ways would be to consider code optimization using generator objects and decorator caching that stores the intermediate computation results to improve runtime while computing the similarity matrix from the All-vs-All results to accommodate large datasets [80]. The scope of the project can be expanded by testing our methods on larger datasets of diverse genomes and increasing the number of upstream and downstream genes considered for neighborhood analysis so that our method can be applied to perform well on various types of data and constraints. We also intend to connect the similarity between neighborhoods and conserved operons with mobile genetic elements such as integrons as it is evident from parts of the introduction that MGEs play a vital role in AMR acquisition and transmission.

The major goal of the project to analyse the neighborhoods of AMR genes of interest in a reference genome was successfully achieved by the results obtained by applying our methods on various datasets. The scoring schemes we developed were easily scaled up to 100 genomes and can be applied to larger datasets with approximate runtime less than ten minutes if Prokka and RGI annotations are readily available. Although it is impractical to visualize hundreds of gene neighborhoods at once, the hierarchical clustering approach allows the selection of representative neighborhoods of size dependant on the dataset. This is achieved by assigning neighborhoods to right clusters based on distance matrix. AMR distribution from the gene tree lends additional support to the hypotheses that some of the Loose hits indeed confer resistance,

and provides evidence of gene transfer between species. We observed interesting cases where some clusters of the dendrogram aligned with tree branches while other showed differences (neighborhoods of *baeS*). Deeper analysis of such unique clusters and tree branches provide additional insights into the AMR functions of Loose hits in that cluster. Hence, The UPGMA clustering, gene order visualization of neighborhoods and phylogenetic analysis provided important insights into the probable functions of AMR genes, and future work will include integrating these methods into prediction pipelines to improve the identification of AMR genes by analysing the neighborhoods of various AMR genes in many diverse genomes.

The major contribution of this thesis work is the tool that can be used by fellow researchers to analyze the various AMR gene neighborhoods within their datasets. When compared to other tools that compare shared orthologs between two genomes and identified conserved patterns such as [61], our approach provides greater flexibility in choosing the number of neighboring genes to be analyzed thus detecting long range patterns. Beginning from the annotation of raw genome assemblies to providing suggestions on various interesting results to look at, the entire process is automated. We limited the neighborhood size to 10 and observed interesting long conserved patterns in the AMR gene neighborhoods, but one could vary this size depending on the requirement of the project and also based on the dataset. UPGMA proved to be the best performing algorithm for our approach, but MCL and DBSCAN could be revisited to modify the distance matrix and similarity scores. Our approach can also be applied to genes other than AMR genes and also there is a flexibility to include different databases other than CARD.

Bibliography

[1] WA Adedeji. The treasure called antibiotics. *Annals of Ibadan postgraduate medicine*, 14(2):56, 2016.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Introduction to pathogens. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[3] Brian P Alcock, Amogelang R Raphenya, Tammy TY Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, et al. Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic acids research*, 48(D1):D517–D525, 2020.

[4] Gregory CA Amos, Semina Ploumakis, Lihong Zhang, Peter M Hawkey, William H Gaze, and Elizabeth MH Wellington. The widespread dissemination of integrons throughout bacterial communities in a riverine system. *The ISME journal*, 12(3):681–691, 2018.

[5] A Andino and I Hanning. Salmonella enterica: survival, colonization, and virulence differences among serovars. *The Scientific World Journal*, 2015, 2015.

[6] Dionysios A Antonopoulos, Rida Assaf, Ramy Karam Aziz, Thomas Brettin, Christopher Bun, Neal Conrad, James J Davis, Emily M Dietrich, Terry Disz, Svetlana Gerdes, et al. Patric as a unique resource for studying antimicrobial resistance. *Briefings in bioinformatics*, 20(4):1094–1102, 2019.

[7] L Aravind. Guilt by association: contextual information in genome analysis. *Genome Research*, 10(8):1074–1077, 2000.

[8] Vicente Arnau, Sergio Mars, and Ignacio Marín. Iterative cluster analysis of protein interaction data. *Bioinformatics*, 21(3):364–378, 2005.

[9] MD Avsaroglu, R Helmuth, E Junker, S Hertwig, A Schroeter, M Akcelik, F Bozoglu, and B Guerra. Plasmid-mediated quinolone resistance conferred by qnrs1 in salmonella enterica serovar virchow isolated from turkish food of avian origin. *Journal of antimicrobial chemotherapy*, 60(5):1146–1150, 2007.

[10] Ariful Azad, Georgios A Pavlopoulos, Christos A Ouzounis, Nikos C Kyrpides, and Aydin Buluç. Hipmcl: a high-performance parallel implementation of the markov clustering algorithm for large-scale networks. *Nucleic acids research*, 46(6):e33–e33, 2018.

[11] Francois Balloux and Lucy van Dorp. Q&a: What are pathogens, and what have they done to and for us? *BMC biology*, 15(1):1–6, 2017.

[12] Xavier Bellanger, Sophie Payot, Nathalie Leblond-Bourget, and Gérard Guédon. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiology Reviews*, 38(4):720–760, 2014.

[13] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, Barbara A Rapp, and David L Wheeler. Genbank. *Nucleic acids research*, 28(1):15–18, 2000.

[14] Patrick Boerlin and Richard J Reid-Smith. Antimicrobial resistance: its emergence and transmission. *Animal Health Research Reviews*, 9(2):115, 2008.

[15] Peer Bork, Thomas Dandekar, Yolande Diaz-Lazcoz, Frank Eisenhaber, Martijn Huynen, and Yanping Yuan. Predicting function: from genes to genomes and back. *Journal of molecular biology*, 283(4):707–725, 1998.

[16] E Fidelma Boyd, Salvador Almagro-Moreno, and Michelle A Parent. Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends in microbiology*, 17(2):47–53, 2009.

[17] Marit S Bratlie, Jostein Johansen, Brad T Sherman, Richard A Lempicki, Finn Drabløs, et al. Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC genomics*, 11(1):1–17, 2010.

[18] Lauren Brinkac, Alexander Voorhies, Andres Gomez, and Karen E Nelson. The threat of antimicrobial resistance on the human microbiome. *Microbial ecology*, 74(4):1001–1008, 2017.

[19] Maryury Brown-Jaque, William Calero-Cáceres, and Maite Muniesa. Transfer of antibiotic-resistance genes via phage-related mobile elements. *Plasmid*, 79:1–7, 2015.

[20] Karen Bush. Alarming β-lactamase-mediated resistance in multidrug-resistant enterobacteriaceae. *Current opinion in microbiology*, 13(5):558–564, 2010.

[21] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):1–9, 2009.

[22] Guillaume Cambray, Anne-Marie Guerout, and Didier Mazel. Integrons. *Annual review of genetics*, 44:141–166, 2010.

[23] Srikanth Celamkoti, Sashidhara Kundeti, Anjan Purkayastha, Raja Mazumder, Charles Buck, and Donald Seto. Geneorder3. 0: software for comparing the order of genes in pairs of small bacterial genomes. *BMC bioinformatics*, 5(1):1–5, 2004.

[24] Lihong Chen, Jian Yang, Jun Yu, Zhijian Yao, Lilian Sun, Yan Shen, and Qi Jin. Vfdb: a reference database for bacterial virulence factors. *Nucleic acids research*, 33(suppl_1):D325–D328, 2005.

[25] Gursharan S Chhatwal. Anchorless adhesins and invasins of gram-positive bacteria: a new class of virulence factors. *Trends in microbiology*, 10(5):205–208, 2002.

[26] Abu Sayed Chowdhury, Douglas R Call, and Shira L Broschat. Antimicrobial resistance prediction for gram-negative bacteria via game theory-based feature evaluation. *Scientific reports*, 9(1):1–9, 2019.

[27] Abu Sayed Chowdhury, Douglas R Call, and Shira L Broschat. Pargt: A software tool for predicting antimicrobial resistance in bacteria. *Scientific reports*, 10(1):1–7, 2020.

[28] ST Cole, K Eiglmeier, J Parkhill, KD James, NR Thomson, PR Wheeler, N Honore, T Garnier, C Churcher, D Harris, et al. Massive gene decay in the leprosy bacillus. *Nature*, 409(6823):1007–1011, 2001.

[29] Fernando De La Cruz, Laura S Frost, Richard J Meyer, and Ellen L Zechner. Conjugative dna metabolism in gram-negative bacteria. *FEMS microbiology reviews*, 34(1):18–40, 2010.

[30] Yang Deng, Xuerui Bao, Lili Ji, Lei Chen, Junyan Liu, Jian Miao, Dingqiang Chen, Huawei Bian, Yanmei Li, and Guangchao Yu. Resistance integrons: class 1, 2 and 3 integrons. *Annals of clinical microbiology and antimicrobials*, 14(1):1–11, 2015.

[31] Damien Devos and Alfonso Valencia. Intrinsic errors in genome annotation. *TRENDS in Genetics*, 17(8):429–431, 2001.

[32] Xavier Didelot, Daniel Lawson, Aaron Darling, and Daniel Falush. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics*, 186(4):1435–1449, 2010.

[33] Xavier Didelot and Martin CJ Maiden. Impact of recombination on bacterial evolution. *Trends in microbiology*, 18(7):315–322, 2010.

[34] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[35] Damodar Reddy Edla, Prasanta K Jana, and IEEE Senior Member. A prototype-based modified dbscan for gene clustering. *Procedia Technology*, 6:485–492, 2012.

[36] Bradley Efron, Elizabeth Halloran, and Susan Holmes. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 93(14):7085–7090, 1996.

[37] Jonathan A Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome research*, 8(3):163–167, 1998.

[38] François Enault, Karsten Suhre, and Jean-Michel Claverie. Phydbac" gene function predictor": a gene annotation tool based on genomic context analysis. *BMC bioinformatics*, 6(1):1–10, 2005.

[39] Barbara E Engelhardt, Michael I Jordan, Susanna T Repo, and Steven E Brenner. Phylogenetic molecular function annotation. In *Journal of Physics: Conference Series*, volume 180, page 012024. IOP Publishing, 2009.

[40] Anton James Enright. *Computational analysis of protein function within complete genomes*. PhD thesis, University of Cambridge, 2002.

[41] José Antonio Escudero*, Céline Loot*, Aleksandra Nivina, and Didier Mazel. The integron: adaptation on demand. *Microbiology spectrum*, 3(2):3–2, 2015.

[42] Gang Fang, Eduardo PC Rocha, and Antoine Danchin. Persistence drives gene clustering in bacterial genomes. *BMC genomics*, 9(1):1–14, 2008.

[43] Joseph Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *evolution*, 39(4):783–791, 1985.

[44] Tom C Freeman, Leon Goldovsky, Markus Brosch, Stijn Van Dongen, Pierre Mazière, Russell J Grocock, Shiri Freilich, Janet Thornton, and Anton J Enright. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS computational biology*, 3(10):e206, 2007.

[45] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[46] Ralph A. Giannella. *Salmonella - Medical Microbiology - NCBI Bookshelf*. 1996. (Accessed on 05/24/2021).

[47] Howard S Gold and Robert C Moellering Jr. Antimicrobial-drug resistance. *New England journal of medicine*, 335(19):1445–1453, 1996.

[48] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A Chapman, Jillian Rowe, Christopher H Tomkins-Tinch, Renan Valieris, and Johannes Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, 15(7):475–476, 2018.

[49] Jörg Hacker. *Pathogenicity islands and the evolution of pathogenic microbes*. Springer, 2012.

[50] Ruth M Hall. Salmonella genomic islands and antibiotic resistance in salmonella enterica. *Future microbiology*, 5(10):1525–1538, 2010.

[51] Ruth M Hall and Christina M Collis. Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Molecular microbiology*, 15(4):593–600, 1995.

[52] Rene S Hendriksen, Valeria Bortolaia, Heather Tate, Gregory H Tyson, Frank M Aarestrup, and Patrick F McDermott. Using genomics to track global antimicrobial resistance. *Frontiers in public health*, 7:242, 2019.

[53] William WL Hsiao, Korine Ung, Dana Aeschliman, Jenny Bryan, B Brett Finlay, and Fiona SL Brinkman. Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet*, 1(5):e62, 2005.

[54] Curtis Huttenhower, Matt Hibbs, Chad Myers, and Olga G Troyanskaya. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 22(23):2890–2897, 2006.

[55] Martijn Huynen, Berend Snel, Warren Lathe, and Peer Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome research*, 10(8):1204–1210, 2000.

[56] Martijn A Huynen and Peer Bork. Measuring genome evolution. *Proceedings of the National Academy of Sciences*, 95(11):5849–5856, 1998.

[57] Ryota Ito, Mustapha M Mustapha, Adam D Tomich, Jake D Callaghan, Christi L McElheny, Roberta T Mettus, Robert MQ Shanks, Nicolas Sluis-Cremer, and Yohei Doi. Widespread fosfomycin resistance in gram-negative bacteria attributable to the chromosomal fosa gene. *MBio*, 8(4):e00749–17, 2017.

[58] Yuxiang Jiang, Tal Ronnen Oron, Wyatt T Clark, Asma R Bankapur, Daniel D'Andrea, Rosalba Lepore, Christopher S Funk, Indika Kahanda, Karin M Verspoor, Asa Ben-Hur, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1):1–19, 2016.

[59] Christopher M Johnson and Alan D Grossman. Integrative and conjugative elements (ices): what they do and how they work. *Annual review of genetics*, 49:577–601, 2015.

[60] I King Jordan, Kira S Makarova, John L Spouge, Yuri I Wolf, and Eugene V Koonin. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Research*, 11(4):555–565, 2001.

[61] Grigory Kolesov, H-W Mewes, and D Frishman. Snapping up functionally related genes based on context information: a colinearity-free approach. In *Bioinformatics and Genome Analysis*, pages 29–63. Springer, 2002.

[62] Eugene V Koonin and Michael Y Galperin. Principles and methods of sequence analysis. In *Sequence—Evolution—Function*, pages 111–192. Springer, 2003.

[63] Liisa B Koski and G Brian Golding. The closest blast hit is often not the nearest neighbor. *Journal of molecular evolution*, 52(6):540–542, 2001.

[64] Akiko Kubomura, Tsuyoshi Sekizuka, Daisuke Onozuka, Koichi Murakami, Hirokazu Kimura, Masahiro Sakaguchi, Kazunori Oishi, Shinichiro Hirai, Makoto Kuroda, and Nobuhiko Okabe. Truncated class 1 integron gene cassette arrays contribute to antimicrobial resistance of diarrheagenic escherichia coli. *BioMed research international*, 2020, 2020.

[65] Maxat Kulmanov and Robert Hoehndorf. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020.

[66] Morgan GI Langille, William WL Hsiao, and Fiona SL Brinkman. Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology*, 8(5):373–382, 2010.

[67] Warren C Lathe III, Berend Snel, and Peer Bork. Gene context conservation of a higher order than operons. *Trends in biochemical sciences*, 25(10):474–479, 2000.

[68] David Lebeaux, Jean-Marc Ghigo, and Christophe Beloin. Biofilm-related infections: bridging the gap between clinical management and fundamental aspects of recalcitrance toward antibiotics. *Microbiology and molecular biology reviews: MMBR*, 78(3):510, 2014.

[69] I Lebrun, R Marques-Porto, AS Pereira, A Pereira, and EA Perpetuo. Bacterial toxins: an overview on bacterial proteases and their action as virulence factors. *Mini reviews in medicinal chemistry*, 9(7):820–828, 2009.

[70] Jongin Lee, Woon-young Hong, Minah Cho, Mikang Sim, Daehwan Lee, Younhee Ko, and Jaebum Kim. Synteny portal: a web-based application portal for synteny block analysis. *Nucleic acids research*, 44(W1):W35–W40, 2016.

[71] Frédéric Lemoine, Olivier Lespinet, and Bernard Labedan. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC evolutionary biology*, 7(1):1–18, 2007.

[72] Chih-Yu Liang, Chih-Hui Yang, Chung-Hsu Lai, Yi-Han Huang, and Jiun-Nong Lin. Comparative genomics of 86 whole-genome sequences in the six species of the elizabethkingia genus reveals intraspecific and interspecific divergence. *Scientific reports*, 9(1):1–11, 2019.

[73] libretext. Bacterial genomes, Jan 2021.

[74] Dang Liu, Martin Hunt, and Isheng J Tsai. Inferring synteny between genome assemblies: a systematic evaluation. *BMC bioinformatics*, 19(1):1–13, 2018.

[75] Carl Llor and Lars Bjerrum. Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem. *Therapeutic advances in drug safety*, 5(6):229–241, 2014.

[76] Yaniv Loewenstein, Elon Portugaly, Menachem Fromer, and Michal Linial. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics*, 24(13):i41–i49, 2008.

[77] Christian Lopez, Scott Tucker, Tarik Salameh, and Conrad Tucker. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. *Journal of biomedical informatics*, 85:30–39, 2018.

[78] Padmanabhan Mahadevan, John F King, and Donald Seto. Data mining pathogen genomes using geneorder and coregenes and cgug: gene order, synteny and in silico proteomes. *International Journal of Computational Biology and Drug Design*, 2(1):100–114, 2009.

[79] Norhan Mahfouz, Inês Ferreira, Stephan Beisken, Arndt von Haeseler, and Andreas E Posch. Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. *Journal of Antimicrobial Chemotherapy*, 75(11):3099–3108, 2020.

[80] Ami Marowka. Python accelerators for high-performance computing. *The Journal of Supercomputing*, 74(4):1449–1460, 2018.

[81] David MA Martin, Matthew Berriman, and Geoffrey J Barton. Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC bioinformatics*, 5(1):1–17, 2004.

[82] Marcus Miethke and Mohamed A Marahiel. Siderophore-based iron acquisition and pathogen control. *Microbiology and molecular biology reviews: MMBR*, 71(3):413, 2007.

[83] Kentaro Mishima, Tomonori Hirao, Miyoko Tsubomura, Miho Tamura, Manabu Kurita, Mine Nose, So Hanaoka, Makoto Takahashi, and Atsushi Watanabe. Identification of novel putative causative genes and genetic marker for male sterility in japanese cedar (cryptomeria japonica d. don). *Bmc Genomics*, 19(1):1–16, 2018.

[84] G Moreno-Hagelsieb and G Santoyo. Predicting functional interactions among genes in prokaryotes by genomic context. *Prokaryotic Systems Biology*, pages 97–106, 2015.

[85] Arcady R Mushegian and Eugene V Koonin. Gene order is not conserved in bacterial evolution. *Trends in genetics: TIG*, 12(8):289–290, 1996.

[86] Nathan L Nehrt, Wyatt T Clark, Predrag Radivojac, and Matthew W Hahn. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol*, 7(6):e1002073, 2011.

[87] Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471, 2012.

[88] Marcus Nguyen, Robert Olson, Maulik Shukla, Margo VanOeffelen, and James J Davis. Predicting antimicrobial resistance using conserved genes. *PLoS computational biology*, 16(10):e1008319, 2020.

[89] NIH. Basic research on antimicrobial (drug) resistance, February 11 2020. available at=https://www.niaid.nih.gov/research/antimicrobial-resistance-basic-research-support.

[90] Chao Niu, Dong Yu, Yuelan Wang, Hongguang Ren, Yuan Jin, Wei Zhou, Beiping Li, Yiyong Cheng, Junjie Yue, Zhixian Gao, et al. Common and pathogen-specific virulence factors are different in function and structure. *Virulence*, 4(6):473–482, 2013.

[91] Mehwish Noureen, Ipputa Tada, Takeshi Kawashima, and Masanori Arita. Rearrangement analysis of multiple bacterial genomes. *BMC bioinformatics*, 20(23):1–10, 2019.

[92] Anne E Osbourn and Ben Field. Operons. *Cellular and Molecular Life Sciences*, 66(23):3755–3775, 2009.

[93] Ross Overbeek, Michael Fonstein, Mark D'souza, Gordon D Pusch, and Natalia Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, 96(6):2896–2901, 1999.

[94] Jelili Oyelade, Itunuoluwa Isewon, Funke Oladipupo, Olufemi Aromolaran, Efosa Uwoghiren, Faridah Ameh, Moses Achas, and Ezekiel Adebiyi. Clustering algorithms: their application to gene expression data. *Bioinformatics and Biology insights*, 10:BBI–S38316, 2016.

[95] Roderick DM Page and Edward C Holmes. *Molecular evolution: a phylogenetic approach*. John Wiley & Sons, 2009.

[96] Sally R Partridge, Stephen M Kwong, Neville Firth, and Slade O Jensen. Mobile genetic elements associated with antimicrobial resistance. *Clinical microbiology reviews*, 31(4), 2018.

[97] William R Pearson. An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, 42(1):3–1, 2013.

[98] Vinita Periwal and Vinod Scaria. Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics*, 31(1):1–9, 2015.

[99] Morgan N Price, Adam P Arkin, and Eric J Alm. The life-cycle of operons. *PLoS Genet*, 2(6):e96, 2006.

[100] Nicola Principi, Ettore Silvestri, and Susanna Esposito. Advantages and limitations of bacteriophages for the treatment of bacterial infections. *Frontiers in pharmacology*, 10:513, 2019.

[101] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227, 2013.

[102] Andrew Rambaut. Figtree v1. 3.1. *http://tree. bio. ed. ac. uk/software/figtree/*, 2009.

[103] Sangeeta Rao, Lyndsey Linke, Enrique Doster, Doreene Hyatt, Brandy A Burgess, Roberta Magnuson, Kristy L Pabilonia, and Paul S Morley. Genomic diversity of class i integrons from antimicrobial resistant strains of salmonella typhimurium isolated from livestock, poultry and humans. *Plos one*, 15(12):e0243477, 2020.

[104] David A Rasko and Vanessa Sperandio. Anti-virulence strategies to combat bacteria-mediated disease. *Nature reviews Drug discovery*, 9(2):117–128, 2010.

[105] Wanda C Reygaert. An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS microbiology*, 4(3):482, 2018.

[106] Louis B Rice and RA Bonomo. Mechanisms of resistance to antibacterial agents. 2007.

[107] Igor B Rogozin, Kira S Makarova, Janos Murvai, Eva Czabarka, Yuri I Wolf, Roman L Tatusov, Laszlo A Szekely, and Eugene V Koonin. Connected gene neighborhoods in prokaryotic genomes. *Nucleic acids research*, 30(10):2212–2223, 2002.

[108] Burkhard Rost, Jinfeng Liu, Rajesh Nair, Kazimierz O Wrzeszczynski, and Yanay Ofran. Automatic prediction of protein function. *Cellular and Molecular Life Sciences CMLS*, 60(12):2637–2650, 2003.

[109] Gabino Sanchez-Perez, Alex Mira, Gábor Nyirő, Lejla Pašić, and Francisco Rodriguez-Valera. Adapting to environmental changes using specialized paralogs. *Trends in Genetics*, 24(4):154–158, 2008.

[110] Samantha Sayers, Li Li, Edison Ong, Shunzhou Deng, Guanghua Fu, Yu Lin, Brian Yang, Shelley Zhang, Zhenzong Fa, Bin Zhao, et al. Victors: a web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic acids research*, 47(D1):D693–D700, 2019.

[111] Imke Schmitt and F Keith Barker. Phylogenetic methods in natural product research. *Natural product reports*, 26(12):1585–1602, 2009.

[112] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[113] Amarda Shehu, Daniel Barbará, and Kevin Molloy. A survey of computational methods for protein function prediction. In *Big data analytics in genomics*, pages 225–298. Springer, 2016.

[114] Tomohiro Shimada, Kaneyoshi Yamamoto, and Akira Ishihama. Involvement of the leucine response transcription factor leuo in regulation of the genes for sulfa drug efflux. *Journal of bacteriology*, 191(14):4562–4571, 2009.

[115] SR Shinde, MV Bhailume, and V SHamde. Transformation of rhizobium with plasmid from pseudomonas spp. a 1113 to degrade dimethoate.

[116] Marie Skovgaard, Lars Juhl Jensen, Søren Brunak, David Ussery, and Anders Krogh. On the total number of genes and their length distribution in complete microbial genomes. *TRENDS in Genetics*, 17(8):425–428, 2001.

[117] Moses Stamboulian, Rafael F Guerrero, Matthew W Hahn, and Predrag Radivojac. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. *Bioinformatics*, 36(Supplement_1):i219–i226, 2020.

[118] Mikita Suyama and Peer Bork. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends in Genetics*, 17(1):10–13, 2001.

[119] Javier Tamames. Evolution of gene order conservation in prokaryotes. *Genome biology*, 2(6):1–11, 2001.

[120] Roman L Tatusov, Michael Y Galperin, Darren A Natale, and Eugene V Koonin. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, 28(1):33–36, 2000.

[121] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6):697–700, 2003.

[122] James Vlasblom and Shoshana J Wodak. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC bioinformatics*, 10(1):1–14, 2009.

[123] Andrea Von Groll, Anandi Martin, Pontus Jureen, Sven Hoffner, Peter Vandamme, Françoise Portaels, Juan Carlos Palomino, and Pedro Almeida da Silva. Fluoroquinolone resistance in mycobacterium tuberculosis and mutations in gyra and gyrb. *Antimicrobial agents and chemotherapy*, 53(10):4498–4500, 2009.

[124] Corneliu Ovidiu Vrancianu, Laura Ioana Popa, Coralia Bleotu, and Mariana Carmen Chifiriuc. Targeting plasmids to limit acquisition and transmission of antimicrobial resistance. *Frontiers in Microbiology*, 11, 2020.

[125] MATTHEW K Waldor, HELMUT Tschäpe, and John J Mekalanos. A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in vibrio cholerae o139. *Journal of bacteriology*, 178(14):4157–4165, 1996.

[126] Hidemi Watanabe, Hirotada Mori, Takeshi Itoh, and Takashi Gojobori. Genome plasticity as a paradigm of eubacteria evolution. *Journal of molecular evolution*, 44(1):S57–S64, 1997.

[127] James D Watson, Roman A Laskowski, and Janet M Thornton. Predicting protein function from sequence and structural data. *Current opinion in structural biology*, 15(3):275–284, 2005.

[128] Benjamin P Westover, Jeremy D Buhler, Justin L Sonnenburg, and Jeffrey I Gordon. Operon prediction without a training set. *Bioinformatics*, 21(7):880–888, 2005.

[129] Yuri I Wolf, Igor B Rogozin, Alexey S Kondrashov, and Eugene V Koonin. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome research*, 11(3):356–372, 2001.

[130] Andrew Wong and Hagit Shatkay. Protein function prediction using text-based features extracted from the biomedical literature: the cafa challenge. In *BMC bioinformatics*, volume 14, pages 1–14. Springer, 2013.

[131] Rachel AF Wozniak and Matthew K Waldor. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews Microbiology*, 8(8):552–563, 2010.

[132] Xinghuo Zeng, Matthew J Nesbitt, Jian Pei, Ke Wang, Ismael A Vergara, and Nansheng Chen. Orthocluster: a new tool for mining synteny blocks and applications in comparative genomics. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pages 656–667, 2008.

[133] Naihui Zhou, Yuxiang Jiang, Timothy R Bergquist, Alexandra J Lee, Balint Z Kacsoh, Alex W Crocker, Kimberley A Lewis, George Georghiou, Huy N Nguyen, Md Nafiz Hamid, et al. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20(1):1–23, 2019.

[134] Zhiyong Zong, Sally R Partridge, and Jonathan R Iredell. Isecp1-mediated transposition and homologous recombination can explain the context of blactx-m-62 linked to qnrb2. *Antimicrobial agents chemotherapy*, 54(7):3039, 2010.

[135] Valentin Zulkower and Susan Rosser. Dna features viewer, a sequence annotations formatting and plotting library for python. *BioRxiv*, 2020.