Identifying The Structure Of Genomic Islands In Prokaryotes

Reem Aldaihani

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Application

Lenwood S. Heath, Chair
Frank O. Aylward
Na Meng
Rym M'Hallah
Liqing Zhang

14 July 2022
Blacksburg, Virginia

Keywords: Prokaryotes, Horizontal Gene Transfer, Genomic Islands, Patterns,
Bacteriophages, Connections

Identifying The Structure Of Genomic Islands In Prokaryotes

Reem Aldaihani

(ABSTRACT)

Prokaryotic genomes evolve via horizontal gene transfer (HGT), mutations, and rearrangements. HGT is a mechanism that plays a significant role in prokaryotic evolution and leads to biodiversity in nature. One of the important components of HGT is the genomic island (GI) which is a subsequence of the genome created by HGT. This research aims to identify the structures of the prokaryotic GIs that have a fundamental role in the adoption of prokaryotes and the impact of the species on the environment. Previous computational biology research has focused on developing tools that detect GIs in prokaryotic genomes, while there is little research investigating GI structure. This research introduces a novel idea that has not yet been addressed intensively, which is identifying additional structures of the GIs in prokaryotes. There are two main directions in this research used to study the prokaryotic GIs structure from each different perspective. In the first direction, the aim is to investigate GI patterns and the existence of biological connections across bacterial phyla in terms of GIs on a large scale. This direction mainly aims to pursue the novel idea of connecting GIs across prokaryotic and phage genomes via patterns of protein families across many species. A pattern is a sequence of protein families that is found to frequently occur in the genomes of a number of species. Here the large data set available from the IslandViewer4 database and protein families from the Pfam database have been combined. Furthermore, implementing a comprehensive strategy to identify patterns that makes use of HMMER, BLAST, and MUSCLE; also implement Python programs that link the analysis into a single pipeline. Research results demonstrate that related GIs often exist in multiple species that are not

evolutionarily related and indeed may be from multiple bacterial phyla. Analysis of the discovered patterns led to the identification of biological connections among prokaryotes and phages through their GIs. A connection is an HGT relation represented as a pattern that exists in a phage and a number of prokaryotic species. These discovered connections suggest quite broad HGT connections across the bacterial kingdom and its associated phages. In addition, these connections provide the basis for additional analysis of the breadth of HGT and the identification of individual HGT events that span bacterial phyla. Moreover, these patterns can suggest the basis for discovering the specific patterns in pathogenic GIs that could play a crucial role in antibiotic resistance. The second direction aims to identify the structure of the GIs in terms of their location within the genome. Prokaryotic GIs have been analyzed according to the genome structure that they are located in, whether it be a circular or a linear genome. The analysis is performed to study the GIs' location in relation to the oriC, investigating the nature of the distances between the GIs, and determining the distribution of GIs in the genome. The analysis has been performed on all of the GIs in the data set. Moreover, the GIs in one genome from each species and the GIs of the most frequent species are in the data set, in order to avoid bias. Overall, the results showed that there are preferable sites for the GIs in the genome. In the linear genomes, they are usually located in the origin of replication area and terminus, and in the circular genomes they are located in the terminus.

Identifying The Structure Of Genomic Islands In Prokaryotes

Reem Aldaihani

(GENERAL AUDIENCE ABSTRACT)

Prokaryotes are one of the most abundant species on earth that play an essential role in naturally shaping the planet and its life. This research aims to identify the structure of a component in these species that has a fundamental role in the adoption of prokaryotes and the impact of the species on the environment. This component is a part of the genome named the genomic island (GI). This dissertation aims to identify the structure of the GIs in two different ways that have not yet been addressed extensively. The first direction aims to discover patterns in the GIs and then use them to bring to light biological connections between prokaryotic and bacteriophages. In this direction, a comprehensive strategy has been utilized to identify patterns and connections. This strategy uses several tools such as BLAST, HMMER, and MUSCLE. Furthermore, Python programs that link the analysis into a single pipeline have been implemented. In the second direction, an investigation has been performed to understand the nature of the GIs' locations within the genome. This direction addresses three different analysis techniques to achieve its target. The three analyses are studying the GIs' location in relation to the origin of replication, investigating the nature of the distances between the GIs, and discovering the location distribution of GIs in the genome. The analysis is performed on linear genomes and circular genomes separately. In each group of GIs, the data set has been utilized to see the results from different perspectives. The overall analysis in both directions relived several findings. In the first direction, the discovered patterns merit deep investigation based on the possibility that they are related to diseases. In addition, in prokaryotic genomes, there are specific sites where the GIs can be

frequently seen that need further search to understand the relation between the GIs' location and the content of the GI in terms of proteins.

Acknowledgments

I would like to thank my supervisor Professor Lenwood Heath for his invaluable advice, support with full encouragement and enthusiasm, noble guidance in everything, constructive feedback, and assistance at every stage of this research. I am deeply grateful and proud to have had the opportunity to work with him. It's an experience that has impacted and inspired me a great deal both in my career and personally. I have benefited greatly from his wealth of knowledge and meticulous editing. Words cannot express my gratitude to my family without whom I could not have undertaken this journey, on account of their generosity by providing me with support, motivation, and guidance. Their belief in me has kept my spirits and motivation high during this journey. I would like to express my deepest appreciation to my committee members, Professor Frank O. Aylward, Professor Na Meng, Professor Rym M'Hallah, and Professor Liqing Zhang, for their time, support, efforts, guidance, and encouragement. Their encouraging words and thoughtful, constructive feedback have been of great value and have assisted me in broadening my horizons. I would like to express my gratitude to Professor Aylward for providing valuable suggestions and answering my questions from a biological perspective. I would like to extend my sincere thanks to the program director Professor Shafer for his guidance and advice. My sincerest gratitude must be expressed towards all of my colleagues in Professor Heath's lab, for their time, support, and feedback sessions. I would like to extend my sincere thanks to Mr. Robert Hunter, system administrator, for his very prompt technical assistance. My research work has been financially supported by Kuwait University, for which I am truly grateful. I would like to thank Virginia Tech University, especially the Computer science department, for the amazing facilities and excellent research environment that they offer all the students all over the world. They made our stay in this professional community satisfying and memorable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview and Motivation

On planet Earth, organisms experience a gradual process of change over the course of many years. This process has been demonstrated by science and named evolution. Charles Darwin introduced the concept of evolution by natural selection in 1859 [34]. The study of evolutionary processes, known as evolutionary biology, can happen through mechanisms such as natural selection, genetic drift, and sexual selection, which are the reasons for the vast diversity of life seen on Earth. The theory that Darwin proposed is by no means the entirety of what is possible within the field of evolutionary biology. Based on the standards of today, Darwin's concept of evolution is fairly limited, and, since his time, there have been many discoveries by evolutionary biologists, who have studied the history of life on our planet in a variety of ways. Consequently, evolution has surpassed the original ideas of Darwin, and there are still many things that remain to be discovered in this field.

Recent research into the field of evolutionary biology investigates numerous issues in which ideas from differing areas, such as microbial genetics evolution, are included. Microbial genetics is the study of microorganisms. It has numerous objectives, like studying the mechanisms of the information that are able to be inherited in microorganisms, including prokaryotes. The genetically driven changes that take place in microorganisms and are not eventually discarded are known as microbial evolution. It is considered likely that some genetic changes that happen in microbial evolution occur due to the presence of genetic

material that moves around within a genome. These genetic materials are called Mobile Genetic Elements (MGEs), and they can also be carried over from one species to another. MGEs have a significant role in evolution, since they exist in all organisms, more specifically, microorganisms. Types of MGEs are DNA transposable elements, bacteriophages, and plasmids. The transfer of genes between bacteria does not just happen between species that are related, but also among those that are unrelated. This gene transfer, which is horizontal, is seen as a crucial part of microbial evolution that takes place in nature.

Horizontal Gene Transfer (HGT), is the process of transferring sequences of genomic DNA between organisms horizontally, which is unlike the regular transfer that is from parent to their offspring in the reproduction process [53]. In short, a genome that acquires DNA fragments from other organisms horizontally is considered as a genome that is the result of different evolutionary histories. Consequently, the resulting genome from the HGT process leads to more difficulties and complexities in the study of lineages and species and how they are related evolutionarily. Frederick Griffith was the first person to observe HGT in 1928 in one of his experiments [49]. In his experiment on *Streptococcus pneumoniae*, it could be seen that virulence had the ability to pass from a virulent strain to a non-virulent strain. He showed that, through a mechanism called transformation, genetic information has the capacity to be horizontally transferred between bacteria. Similarly, conjugation and transduction were observed in the 1940s and 1950s, which provided proof that there were additional mechanisms of HGT [126] [148] [67]. Prokaryotic genomes may contain a sequence of genes that were obtained from other microorganisms by horizontal DNA transfer; this sequence of genes is called a genomic island (GI). GIs are composed of mobile DNAs that have a central role in prokaryotic evolution. As a result, the interest in identifying GIs in prokaryotic genomes has recently increased. Even though numerous computational methods have been devised for the detection of the different kinds of GIs, not one method to this present day has been capable of detecting all GIs on its own. For this reason, it seems to be advantageous to know the structure of these GIs to discover why there is difficulty in

producing a method to detect all GIs .

1.2 Problem Statement

This research introduces a novel idea that has not yet been addressed intensively, which is understanding the structure of the GIs in prokaryotes. There are two main directions in this research with various research questions for each direction. The first direction is the structure of the GIs in terms of their function, which was studied by analyzing the protein families of the proteins in the GIs to generate patterns and connections. A pattern is a sequence of protein families that is found to frequently occur in the genomes of a number of prokaryotic species. In nature, the HGT events are seen as a crucial part of microbial evolution. In these HGT events, there could be an interaction between numerous microorganisms that are related or distantly related. Therefore, this research introduces the term connection, which is defined as an HGT relation resulting from an interaction between a phage and a number of prokaryotic species. In this direction of the research, the structure of the GIs was studied in terms of their biological function, which was performed by analyzing the protein families of the proteins in the GIs. Therefore, hmmsearch and Pfam-33 were used to transfer the proteins to protein families. A priori, a data mining algorithm, was used with newly implemented methods to generate patterns. The resulting patterns went through numerous filters to obtain patterns of potential interest. The in-depth analysis of the patterns was performed using BLAST, MUSCLE, and newly implemented algorithms built using Python. It was observed that the patterns in the GIs are mostly made of phage protein families. Phage protein families are protein families of protein sequences that usually exist in phages, which are viruses that infect prokaryotes. These phage protein families mean that there is a biological connection between prokaryotic and phages in terms of GIs. Moreover, the analysis led to discovering significant connections between phages and bacteria, and they were filtered to get the promising ones. The resulting connections bring to light shared prokaryotic GIs

between specific species. Regarding the discovered connections, they indicate an HGT event between a phage and a number of prokaryotic species. Therefore, their main advantage is an assist in understanding the prokaryotic evolution resulting from the HGT mechanism.

The second direction aims to identify the structure of the GIs in terms of their location within the genome. The GIs in this research direction have been preprocessed to overcome the overlapped GIs in the data set. Furthermore, the GIs have been divided into two groups; circular group and linear group, according to the genome structure that the GI located in. This direction has been analyzed using different techniques. This was performed by studying the GIs location in relation to the origin of replication, investigating the nature of the distances between the GIs within the genome, and analyzing the location distribution of GIs in the genome. Each of the three analyses were performed using three cases according to the data input utilized: first, all of the GIs in the data set; second, the GIs in one genome from each species; and third, the GIs of the most frequent species in the data set.

There are several GI databases that store the resulting GI sequences from computational methods, one of which is the IslandViewer website that has been used in this research. The IslandViewer4 database (i.e., genomes, GIs, and proteins), which is the latest update and the most comprehensive database, was used as an input for this research to achieve the desired results.

1.3 Dissertation Organization

This dissertation is organized as follows; Chapter 2 contains the biological background that explains the prokaryotic genomics, bacteriophages, protein families, horizontal gene transfer, and genomic islands. Later, in Chapter 3, the literature review can be found, which presents the HGT and GIs in prokaryotes, the related work to my research, and the GI databases. After that, Chapter 4 presents the data set used in the research. Chapter 5 then focuses on

the first research question, which is connection GIs across prokaryotic and phage genomes via protein families. This chapter demonstrates methods used to produce the novel patterns and connections. Later, in Chapter 6, the second part of the research is presented, which is investigating the nature of GIs' locations within a genome. Finally, Chapter 7 contains the conclusions and the future research directions.

Chapter 2

Biological Background

A domain, in the field of biological taxonomy, is known as the highest taxonomic rank of organisms that is found in the current three-domain system of taxonomy, which was proposed in 1990 by Carl Woese et al [136]. Within this system, every living organism, apart from viruses, can be categorized into three domains: Eukarya, Bacteria, and Archaea. The latter two are prokaryotic microorganisms, single-celled organisms containing no nucleus within their cells. Biological scientists have discovered that prokaryotes,which are the subject of this research, are the earliest forms of land life on the planet Earth from more than three billion years ago [59] and are also among the most diverse organisms. Furthermore, they are able to live and proliferate in numerous diverse environments and exchanging their genetic material.

The biological background section is divided into five subsections. The research focused on prokaryotes; therefore, an explanation of prokaryotic genomics is provided at first. Second, biological information about the bacteriophages is presented. Third, understanding the components of the genome is accomplished by delving deeper into them in terms of knowledge of prokaryotic protein families. Fourth, illustrating the basic idea and methods of HGT will be carried out. Fifth, the underlying biological concept of GIs is illustrated.

2.1 Prokaryotic Genomics

Genomics is a field in molecular biology relate to the study of genomes from numerous prospective such as studying or analyzing the structure of the genomes. Biological sequence analysis plays a vital part in bioinformatics by means of comparing, analyzing, aligning, and indexing biological sequences. The benefits of sequence analysis are used in this research to analyze the genomes that contain GIs, as well as analyze their GIs and the associated proteins. The aim of this analysis is to discover the structure of GIs that make them transferable as a cluster of genes from one genome to another. The genetic material in prokaryotes typically made of a chromosome that consists of a DNA molecule that exists in a circular form, named a genophore. Along with the genophore, prokaryotic cells can contain more circular DNA molecules known as plasmids. In a mechanism called bacterial conjugation, bacterial cells use their sex pili to exchange plasmids, which facilitates the evolution of bacteria in order to create features for the next generation. The conjugation mechanism is one of the HGT mechanisms used by prokaryotes to transfer genetic materials such as GIs. These mechanisms are illustrated next in the HGT section (Section 2.4).

In general, there are sequences in DNA that code for proteins, but there are other sequences that do not (i.e. noncoding DNA). Prokaryotic genomes have less noncoding DNA than eukaryotic DNA. On average, twelve percent of the genomic sequences in prokaryotes are noncoding, whereas in eukaryotes there are around ninety-eight percent [3] [83]. This could be due to the small size of prokaryotic genomes.

2.2 Bacteriophages

Bacteriophages (Phages) are one of the most abundant living entities. Phages are viruses that predominantly infect bacteria and usually kill them. Bacteriophages discovered in 1915 and 1917 by Frederick William Twort in England and then in 1917 by Felix d'Herelle, their

role within the ecosystem, in terms of bacterial evolution and microbial balancing, is a crucial one. Moreover, when compared to bacteria, the number of phages that exist in the environment is tenfold making it probable that they are the most abundant organism on the planet [21]. They play a critical role in regulating bacterial populations and nutrient cycling. In general, phages are ubiquitous since they are isolated from numerous environments. They exist in areas such as the food [93], humans, animals, soil, and water [77].

Bacteriophages are highly diverse in terms of their shape, size, capsid symmetry and structure. They can present variations in DNA and RNA, such as double-stranded (ds) or single-stranded (ss). [25] [1]. Bacteriophages are classified by a number of policies that were created by the International Committee for Taxonomy of Viruses (ICTV) under a multitude of properties [1], of which the most important are their nucleic acid type (i.e. DNA or RNA), their morphological characteristics, the location where they are predominantly detected, bacteria target (i.e. the bacterial species which they are able to infect and kill), and the biological cycle (e.g. lytic) [104].

Classification is crucial for phylogenetic studies since it permits comparisons, which facilitate the understanding of the relationships among phage groups. Moreover, classification is required to identify novel phages. According to estimates, more than 96% of phages across the globe are part of the Caudovirales order characterised by tailed helical particles, provided with adhesion structures (spikes, fibres and baseplate) and dsDNA genome, which demonstrate a range of sizes between 18kb and 500kb [1] [95]. The order is composed of the following three families: Myoviridae (contractile tail), Podoviridae (short non-contractile tail) and Siphoviridae (long non-contractile tail). It should be noted that there is an increasing number of sequenced bacteriophages and the reclassification or creation of new orders along with families is constantly ongoing, and, for this reason, this list is not entirely complete.

Bacteriophages play a major role in the microbial ecosystem by regulating its actions, which

is done by way of gene transfer, metabolic reprogramming, and killing [43]; [65]. Many human diseases can be a result of altering the virulence of bacteria, which is carried out by phages using the HGT mechanism between bacteria. Numerous studies have been performed in order to use bacteriophages for biological applications such as antimicrobial drug discovery [65], disease diagnostics [11] [133] and phage therapy [84] [81]. Bacteria and bacteriophages contain a sophisticated network of interactions where continued coexistence and survival of bacteriophages and their host bacteria is permitted. Bacteria and bacteriophages constantly battle in a manner where bacteria build up a resistance to phages, while phages develop methods to tackle that resistance [22]. Evidence of such battles dates back to three billion years ago [57], which has given rise to a vast amount of diversity in phage and bacterial genomes by way of horizontal gene transfer and novel mutations [57] [122]. A high prevalence of horizontal gene transfer in oceanic and coastal environments was documented by McDaniel and coworkers 2010, confirming that 47% of the cultivable natural microbial community were gene recipients. Proving that at least a partial amount of this HGT was executed by viral-like particles released by the bacteria, they proposed that marine bacteria could adapt more easily to varying environmental conditions. It is possible that HGT is the principal factor for spreading antibiotic resistance within bacteria. Genome variety expands due to HGT, as it introduces new genes that have novel functions facilitating the improvement of adaptation to specific environments.

2.3 Protein Families

A protein family is a set of proteins that are related evolutionarily. The proteins that belong to the same family would share the same ancestor. Moreover, they commonly have notable sequence similarity and related functions. By contrast, proteins that do not descend from the same ancestor will not show a sufficient sequence similarity between them. Part of this research is to study the protein families that exist in GIs. Answering this question

requires the identification of the protein families of the proteins in the GIs. There are several databases that identify protein families using sequence alignment and clustering methods, such as the Protein families (Pfam) database. The Pfam database is the database that is used in this research, which is a set of curated protein families, each represented by Hidden Markov models (HMMs) and Multiple Sequence Alignments (MSA).

2.4 Horizontal Gene Transfer

Horizontal Gene Transfer (HGT) is a process of transfer of genetic information between organisms where MGEs and GIs are seen as the agents of HGT. There are two processes that work in collaboration to form the genome in the evolution of prokaryotic organisms: HGT and direct gene inheritance (i.e., vertical gene transfer). Previously, it was believed that HGT rarely occurs and that the main process in evolution was inheritance. With advancements in DNA sequencing, it has become obvious that HGT is also important. There are three major mechanisms for HGT in which the genetic material transfers between organisms: transduction, transformation, and conjugation [124].

The transduction mechanism is the transfer of a DNA fragment from a bacterium to another by a bacteriophage (i.e. virus), which means the bacteriophages work as vehicles of foreign DNA transfer between organisms. Transformation involves taking exogenous DNA from the environment. The conjugation mechanism involves cell to cell contact using plasmids to transfer the genetic material. The ability to obtain foreign DNA horizontally from other organisms is one reason for prokaryotic adaptability and genetic diversity, which results in genes being rearranged, acquired or removed [96]. The evidence from genome analyses, which continues to grow, clearly shows that HGT is a crucial force for driving prokaryotic evolution [31] [48] [96]. This can be seen in prokaryotic genomes they possess a large number of foreign genes due to the HGT process [76].

2.5 Genomic Islands

Genomic islands (GIs) are composed of sets of genes and have a size that usually ranges from 10 kb to 200 kb [55]. The genomic islet is a GI with a distance not higher than 10 kb. The definition of GIs is broad and enters into the territories of other MGEs. For this reason, the term GI was created to assist in generally describing a genomic sequence consisting of a set of horizontally acquired genes.

GIs can be identified computationally with some success from genomic sequences by analyzing a number of different features of GIs. They contain several features that assist in identifying them from the remainder of a bacterial genome based on their sequence and structure [79]. One of the features of GIs that stands out the most is that their phyletic patterns are different to the pattern of the host genome patterns because of the sporadic distribution of GIs. Moreover, there are also specific kinds of protein-coding genes that are linked with GIs.

To detect GIs computationally, two main bioinformatics methods are employed [79]: methods based on comparative genomics and methods based on sequence composition. Comparative genomics is a biological research field whose aim is to compare the genomic features of different organisms sequences, such as using probabilistic approaches [113]. By contrast, the aim of sequence composition is to study the frequency of DNA bases, such as GC-content, that are present within a genome. Composition-based methods are advantageous due to their need for only a single genome or sequence to be analyzed, while methods that use comparative genomics require more and compare numerous genome sequences to identify GIs. Genome sequences are used by numerous computational methods for GI prediction to recognize the features and signature of the GIs. Identifying the prokaryotic GIs is highly important in studying the prokaryotic genomes, especially infectious disease pathogens.

Chapter 3

Review of Literature

3.1 Horizontal Gene Transfer and Genomic Islands in Prokaryotes

HGT, which is an significant part in the evolution of a prokaryotes, is the distribution of genetic materials by any means other than the vertical transmission between species [53] [16]. By presenting new genes assists in improving adaptation while also containing novel functions, HGT expands the variety of genomes. This process advances the fast spread of genetic information across lineages, which usually appear as genomic sequences known as GIs. These GIs exist in various types, which are categorized based on their content. A number of computational approaches have been built to recognize various kinds of GIs, yet there is no method that can recognize every GI [10]. Therefore, identifying the structure of GIs is another way of detecting GIs that is more general and comprehensive that has been targeted in this research.

Although it is clear that GIs are able to carry various kinds of genes, there are some specific kinds of genes that have been demonstrated to occur disproportionately in GIs than in the remainder of the chromosome. For example, research conducted in the past by Dr. Fiona Brinkman in her laboratory has proven that GIs are one of the main sources of novel genes in bacteria [62]. This was determined because of the large number of proteins of unknown functions and without any known homologs being found to be linked with GIs. These are said to originate from the *phage gene pool* which is much more extensive gene pool than the

bacterial gene pool [123] and represent a source of greater diversity for bacteria. Various different studies have demonstrated that, generally, phage-related genes are linked with GIs [130] [132]. Phages belonging to the Caudovirales order and other phages represent the largest number of entities on earth. Bacterial cells are believed to be infected by these bacteriophages each second [72]. Moreover, numerous antibiotic resistance (ABR) genes typically spread via GIs and plasmids. The spread of ABR genes in several instances is owing to such HGT via phage transduction [20] or bacterial conjugation [131].

In short, GIs have demonstrated associations with a range of genes and a wide array of novel or uncharacterized genes, of which all significantly play roles in regards to the evolution of bacterial species. Additionally, GIs have also demonstrated the ability to carry alternative metabolism genes, antibiotic resistance genes, and other genes for adaptation that contribute in the adjustment to environmental changes for microbes. It is likely that HGT is an essential factor in the spreading of antibiotic resistance in prokaryotes [71].

GIs also have the ability to disrupt or change the host gene expressions. From the perspective of an infectious disease, GIs could contribute to rapid alteration of microbial phenotypes. This can be done by either acquiring a set of virulence genes that can suddenly turn into a pathogenic strain from a harmless strain, or by permitting the survival of microbes by means of sanitation procedures by obtaining genes for resistance against sanitation. An all round interest in identifying GIs exists due to all of the previously mentioned reasons.

3.2 Related Work

In recent times, it has become obvious that GIs have had a large impact on prokaryotic evolution, as the amount of genomes has continually increased. The use of bioinformatics approaches in order to detect such regions become a crucial part of the study of microbial function, structure, and evolution. Therefore, in this research, a large number of GIs have

been studied and analyzed to measure features linked to islands so that the commencement of the process of improving the identification and structure of them can be carried out. GIs assist in the part of HGT that contributes to the adaptation and diversification of bacteria, but the structure of these GIs are not understood well. Nonetheless, researchers have discovered several of the main features of GIs and that they share some common structural and sequence features that can distinguish them from the remainder of the genome. Firstly, genomic islands are irregularly distributed in strains of the same species or in species that are closely related. Secondly, it is possible for GIs to be recognized by means of nucleotide statistics, such as GC content, that are usually unlike the other parts of the genome. Thirdly, GIs often insert adjacent to tRNA genes. Fourth, most GIs that have been found until now are somewhat large in terms of their segments of DNA, which have a range of 10kb to 200 kb [54]. Fifth, GIs frequently hold mobility genes such as transposases and integrases. Sixth, GIs can be regularly flanked by IS elements or direct repeats. Seventh, GIs typically hold genes that give host bacteria selective benefits in islands such as resistance and pathogenicity islands [38] [111]. Every feature does not necessarily exists in a region for that region to be classified as a GI, since not all the features exist in all of the reported GIs. However, the existence of a group of these features is good evidence of the presence of GIs.

There have been various computational tools developed for the prediction of islands within sequenced genomes [62] [106] [128] [129] [132] which, for the most part, make use of the sequence biases existing in prokaryotes that occur naturally. This is performed in order to detect regions that may have a foreign sequence composition [70] [130].

In 2014, FB Guo et al conducted a statistical analysis on five conserved features of GIs of prokaryotes [52]. In this analysis, GIs were found by using comparative methods based on 104 recognized GIs. Four of these features included abnormal GC content, sequence size, embedded mobility gene, and flanking tRNA gene, and G+C homogeneity. As mentioned previously, GIs generally have a length of 10-200 kb. Regarding the GIs that were not inside of the typical length range (i.e. 10 to 200 kb), there were only 12 GIs. As a result,

more short GIs than long GIs appeared based on the characteristics of the 104 GIs. As a result, based on the characteristics of the 104 GIs, there are few long GIs than short GIs. Regarding GC content, two classes were made for the 104 GIs; GC-richer and AT-richer. As a result it appears that the absolute value of GC deviations of GC-richer island is similar to the absolute value of GC deviations of the AT-richer. Furthermore, the homogeneity feature was investigated, which was used by Zhang and colleagues to be one of the standards approaches used for the prediction of novel GIs [144] [145] [146] [135] [51]. In relation to the G+C homogeneity feature, FB Guo found that an h homogeneity index that is less than 0.1 was found in 85% of GIs, demonstrating that the G+C content of GI is orderly for the most part. In regards to the mobility genes (i.e. transposases and integrases) they could assist in the integration of GIs to the host genome [130] [118]. Moreover, tRNA genes frequently flank GI regions [130] [118].

Out of the 88 genomic islands, 87.5% had three of the five features. There was just one GI that had a single conserved feature, while there were no genomic islands without any of the features.

A machine learning method for the identification of genomic islands in prokaryotic based on eight criteria was introduced by GS Vernikos et al. These criteria included presence/absence of integrase, the interpolated variable order motif (IVOM) score, size, RNA, presence/absence of phage-related protein domains, repeats, insertion points, and density [130]. Their research show that genomic islands can be interpreted as a superfamily of mobile elements and were performed in three distinct prokaryotic genera: Salmonella, Staphylococcus and Streptococcus. The three genus-specific GI models demonstrate core and variable structural features containing recognizable genus-specific signatures. Moreover, they have indicated that for GI prediction the features with the most importance are the presence of phage-related proteins, the sequence composition bias, the size of the region, and the existence of integrase. Conversely, tRNA genes, the presence of flanking direct repeats, and gene density, carry less importance despite being informative.

With respect to the genes within GIs, bioinformatics studies have demonstrated that novel genes appear more in GIs [62]. The encoding of proteins that perform the basic functions for sustaining cellular life is thought to be carried out by essential genes. An analysis into the essential genes in 28 prokaryotes was conducted by X Zhang et al. This was done using a statistical method, and it concluded that significant genes in genomic islands occur in significantly lower numbers than the ones that are outside of GIs [147].

In 2007, Hsiao analyzed prokaryotic genomic islands (most of which were pathogenicity islands) and investigated their computational characterization [61]. A set of prior reported genomic islands of 60 prokaryotic genomes were analyzed to quantify features related to islands for the first stages of improving their structure and identification. Some of the features associated with islands were integrated into it, of which the main ones were dinucleotide bias of gene clusters, GC content, tRNA genes, and occurrence of mobility genes. Hsiao demonstrated that, in a wide range of phyla, the gene placement in the islands against the distribution in the remainder of a given genome appears not to be random. Specific protein functional categories, such as novel hypothetical proteins and virulence factors, are more common in genomic islands.

Bush et al. introduced a new strategy in understanding GIs [23] by implementing a tool named, xenoGI. xenoGI is a powerful tool that allows users to analyze the history of genomic island insertions, which is represented as a clade of microbes. The information it produces is important as it assists in comprehending the adaptive path which has developed living species.

3.3 Genomic Island Databases

There are numerous GI databases that are available and contain a large number of prokaryotic GIs. However, there are some newest GIs databases that containing a certain type of

bacterial species such as VCGIDB which is a database for the GIs from Vibrio cholerae [64]. In this section, a number of the most significant GI databases that contain a numerious kind of prokaryotic species are reviewed.

The Pathogenicity Island Database (PAIDB) [141] [140] is a web-based tool that has a variety of verified pathogenicity islands (PAIs), which are a category of GIs that are influential in microbial virulence. To browse PAIs, there are three possibilities: search with BLAST, carry out a text search, or browse by species.

IslandPath [60] is a database that has a visual interface that assists in the detection of GIs, tRNA genes, mobility genes, and genes with a significant deviation from the dinucleotide bias or average GC content. The output of IslandPath is a graphical view of the whole genome, which draws attention to features associated with GIs and assists in the identification of putative GIs manually.

Islander [88] [63], a database of GIs with the site of their tRNA integration in genomes, was first presented in 2004 by Mantri et al. It was subsequently developed and then redeployed in 2015 by Hudson et al. GI predictions were made using tmRNA and tRNA genes predicted by BruCE and trNAscan-SE in a BLAST search. In Islander, it is possible to browse GIs by GI name, integration site or organism name.

Predicted genomic islands (Pre_GI) [101] is a free comprehensive database that contains a major number of prokaryotic GIs that have been identified in chromosomes and plasmids. The purpose of Pre_GIs is to be a web-resource for two kinds of analysis; ontological paths analysis between islands, cartographic analysis.

IslandViewer [12] is one of the most commonly used GI prediction and interactive visualization web servers for prokaryotic genomes, which has also been updated recently. The database of IslandViewer is composed of four different island prediction methods that are used to detect HGT fragments in bacterial genomes but without recognizing the ontological links between them. IslandViewer is the first web server to integrate four GI prediction tools:

SIGI-HMM [132] that is based on a HMM approach with codon usage bias; IslandPick [78], which is a tool that is built on a comparative genomics approach, Islander [63], a database of precisely mapped GIs in tRNA and tmRNA genes; and IslandPath-DIMOB [60], which is a tool that is dependent on nucleotide bias and the existence of MGEs.

In this research, IslandViewer is used since it integrates the four approved island prediction methods, as mentioned previously. Also, it contains a large number of GIs that have been updated recently compared to other GI databases. It was demonstrated that IslandViewer4 and GIHunter gave the most precise and accurate results in a very extensive study whose aim was to examine and compare a large number of GI prediction tools for their methodology, user friendly, precision, and accuracy. [13]

Chapter 4

Data Sets

The data set of GIs used in this research was downloaded from the IslandViewer4 website in April, 2020. The GIs are predicted by different tools. The database of IslandViewer4 is composed of four different island prediction methods that are used to detect HGT fragments in prokaryotic genomes but without recognizing the ontological links between them. In IslandViewer4, there are numerous prokaryotic genomes of which most contain a set of GIs along with their respective genes and proteins. These GIs and their genes and proteins are located on the genome page that the GIs belong to, and they can all be downloaded in separate FASTA files.

The data set in this study contained 14,332 prokaryotic genomes of which 420 genomes did not contain GIs. Therefore, as the study mainly depended on prokaryotic GIs, the genomes that did not have GIs were eliminated from the study, meaning that the total number of genomes used was 13,912. In these genomes, there were a total of 384,236 GIs of which 74 could not be downloaded due to their large size causing server issues with IslandViewer4. This meant that the total number of GIs used in this research is 384,162 and the total number of protein sequences in these GIs is 4,725,173, which is based on every single protein sequence in the GIs, even if there is an overlap between some of the GIs. To elaborate further, some GIs are overlapped, which means part of their sequence shares the same area in the host genome, which is due to these GIs being discovered from different GI prediction tools. Therefore, the shared area between the overlapped GIs contains proteins, and they are counted separately according to the number of the GIs.

In IslandViewer4, the protein sequences of each GI are stored in FASTA files, as mentioned previously. Each GI file contains the protein sequences that are in a GI where each sequence comes with a description line, which contains the main information of each of these protein sequences. The loci of protein sequences in the GI is also included (i.e. coordinates), as well as the protein sequence accession number. These are the two main pieces of information needed from the heading line, however, the GI number and genome accession number, which are also important to this research, are not represented by default. For this reason, in this study, the heading line was updated to include the GI number, which is the position number of the GI on its genome page on the IslandViewer4 website, as well as the genome accession number, which represents the species that a GI belongs to. All of these updates were performed to facilitate the analysis process.

The phage data set used in this research to detect the HGT connections is downloaded from the NCBI website. Furthermore, the Pfam database version 33 was used along with HMMER to generate the protein families of protein sequences in the GIs and in the phage genomes. The Pfam database comprises a massive number of protein families represented as Multiple Sequence Alignment (MSA) and Hidden Markov Model (HMM) [44].

Chapter 5

Connecting Genomic Islands across Prokaryotic and Phage Genomes via Protein Families

5.1 Research Problem

I formulate the research problem mathematically. The challenge in computational biology is that there are several GI prediction tools that each detect GIs according to specific criteria. However, at present, there is still not enough research on the structure of the prokaryotic GIs and the possibility of their presence in organisms from other kingdoms. The main aim of this research is to connect GIs across prokaryotic and phage genomes via protein families. This is performed by identifying the structure of the GIs in terms of their function by analyzing their protein families and discovering patterns within them before performing intensive analysis on the resulting patterns to obtain connections between prokaryotes and phages. Starting with the data set, the IslandViewer4 website is a comprehensive database that contains a number, N, of prokaryotic genomes ($\mathcal{G}$) from numerous prokaryotic species:

$$\mathcal{G} = \{G_1, G_2, G_3, \ldots, G_N\} \tag{5.1}$$

Each genome G_i is located on the IslandViewer4 website on a separate webpage that contains

information about the G_i genome and the GIs that have been discovered in the that genome. Each genome is from a known species; some species have many genomes. A GI is a subsequence of the genome, and, in each genome G_i, there are a number of GIs in IslandViewer4, each predicted by some GI prediction tool.

The webpage for a genome G_i contains the genome sequence, the GI sequences in the genome, the protein sequences in each GI, and other information related to the genome and the GIs. As the main idea of the research is based on protein families, the research input is the proteins within each GI. Each GI contains a sequence of genes, in order, which give rise to a sequence of proteins, say,

$$P_1, P_2, P_3, \ldots, P_K$$

where K is the number of the proteins in the GI.

In this research, I analyzed the entirety (or universe) of the GIs through a number of steps illustrated in the pipeline in Figure 5.1. The pseudocode for all the pipeline algorithms can be found in the appendix.

The first stage of the pipeline is to retrieve the GI proteins from the IslandViewer4 website. Biologically, each protein, according to its biological function, is classified under a specific protein family. Therefore, the aim is to obtain the protein family for each protein found within GIs. This means that each protein P_i in a GI is mapped to its corresponding protein family $f(P_i)$ using hmmsearch and Pfam. The GI can then be thought of more coarsely as the sequence of protein families:

$$f(P_1), f(P_2), f(P_3), \ldots, f(P_K).$$

Thinking of each such sequence as a set of protein families. The other objective of this stage is to filter the GIs. Section 5.2.1.2 presents more detail about filtering the GIs.

In stage two, the first step is to use the Apriori algorithm to identify the most frequent sets

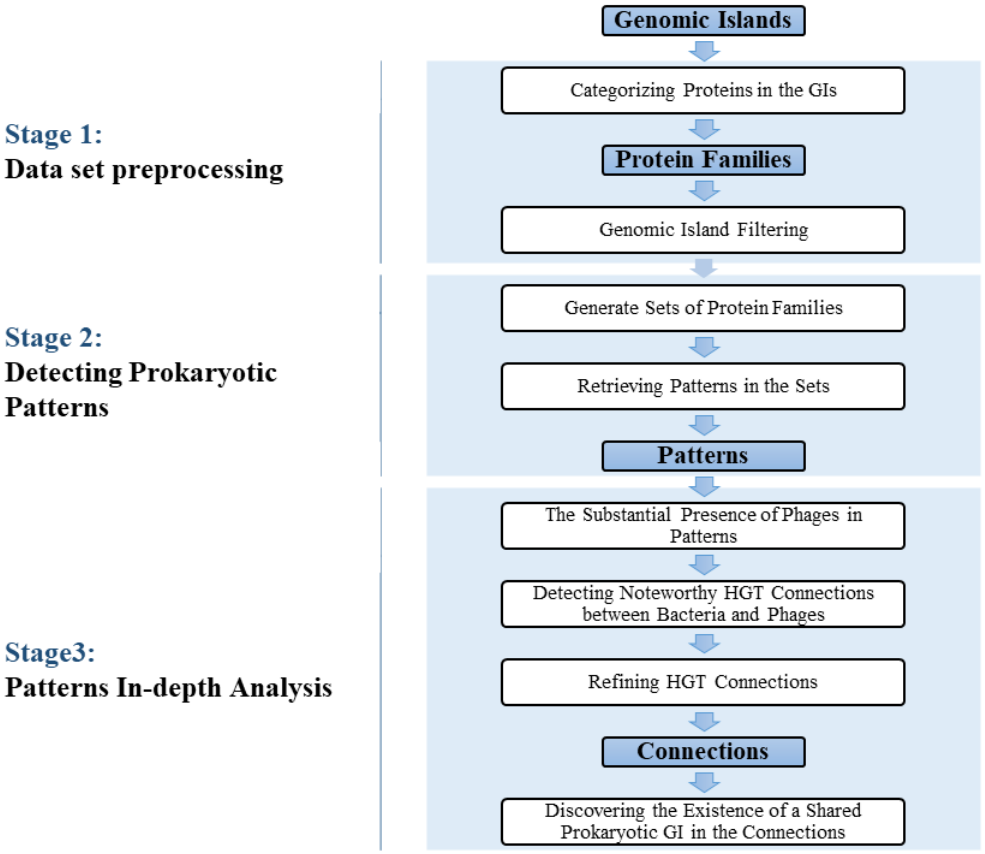


Figure 5.1: The Pattern-Connection Pipeline is divided into three main stages. The first stage is the data set preprocessing. The function of this stage is to generate protein families of the proteins in the GIs and filter the GIs. The second stage is to obtain the patterns by generating sets of protein families using the Apriori algorithm before retrieving patterns from the sets and filtering them to get the promising patterns. The third stage of the pipeline is related to the connections. In this part, the patterns are as input and go through three major steps to get the connections. In these three steps BLAST is used to detect matches for the prokaryotic GIs with the phage database to obtain connections and then filter them to obtain the promising ones.

of protein families in the universe of GIs. Therefore, from all these sets of protein families, the Apriori algorithm identifies the T most frequent sets:

$$\mathcal{S} = \{S_1, S_2, \ldots, S_T\}$$

One set $S_j \in \mathcal{S}$ contains a number Z of protein families that frequently exist in some of the GIs, though not necessarily in the same order or even consecutively.

The second step of stage two of the pipeline is to retrieve patterns in these sets and filter them under certain conditions to obtain promising patterns using BLAST and the new algorithms. A pattern is a number of protein families that frequently exist in the GIs in the same order that they present themselves in the pattern, but not necessarily consecutively. An example of a pattern of length three might be $P = F_1, F_2, F_3$, of protein families, where order matters. Pattern analysis finds that there are a number 20 of patterns in the GIs. Further investigation demonstrates that the patterns exist in several GIs and that these GIs are in genomes from different species.

In stage three of the pipeline, I took a deeper look at the resulting patterns to connect GIs across prokaryotic and phage genomes via protein families. Suppose a particular pattern P exists in a GI in each genome from this set $\{B_1, B_2, \ldots, B_d\}$ of bacterial species. The protein families in the patterns mainly belong to phage protein families, which is evidence of a connection between the bacteria and phages in terms of protein families. Therefore, a pattern in a bacteria GI and phage is proof that there is a connection in terms of protein families between them. The in-depth analysis is mainly performed by using BLAST to find the connection between the bacteria and phages and then using certain conditions to filter the resulting connections to obtain promising ones. This deeper look at the resulting patterns results in a number of connections. A connection C is a triple consisting of a pattern P, a phage genome H, and a set of bacteria species:

$$C \;=\; (P, H, \{B_1, B_2, \ldots, B_d\})$$

The phage and bacterial species share the common pattern P in their genomes and the number d of bacteria will vary. Therefore, the set of connections from the in-depth analysis are a particular set

$$\mathcal{C} = \{C_1, C_2, \ldots, C_L\} \tag{5.2}$$

The resulting patterns and connections lead to the target of this research, which is discovering GI patterns and then using these patterns to connect GIs across prokaryotic and phage genomes via protein families.

5.2 Analytical Flow

This section explains in detail the three stages of the Pattern-Connection pipeline in Figure 5.1, which is presented in the research problem section (Section 5.1).

5.2.1 Data Set Preprocessing

The GIs in the data set are from numerous prokaryotic species that are from various taxonomies. However, it was noticed that there were more GIs in certain species than in other species within the data set, which is demonstrated in Table 5.1 with the five most frequent species that the GIs in the data set belong to. *Escherichia coli* species are the most frequent prokaryotic species in the data set since they are a large and diverse group of prokaryotes. Regarding the connections between the top species, from a taxonomy point of view, all the species were from the same phylum known as Proteobacteria, as shown in the taxonomy tree in Figure 5.2. The first three species (i.e. *Escherichia coli*, *Salmonella enterica*, *Klebsiella pneumoniae*) are from the same family named Enterobacteriaceae, Pseudomonas aeruginosa is from the same class as the previous three bacteria species but from another order called Pseudomonadales, and *Bordetella pertussis* is further from all of the aforementioned species since it is from a different class called Betaproteobacteria. In general, Proteobacteria is a major phylum of Gram-negative bacteria which includes some pathogenic species, and it is the largest and most diverse phylum of bacteria [108].

Table 5.1: Most frequent species in the data set

Species	Genomic Islands
Escherichia coli	51152
Salmonella enterica	29225
Klebsiella pneumoniae	17238
Bordetella pertussis	16382
Pseudomonas aeruginosa	8614

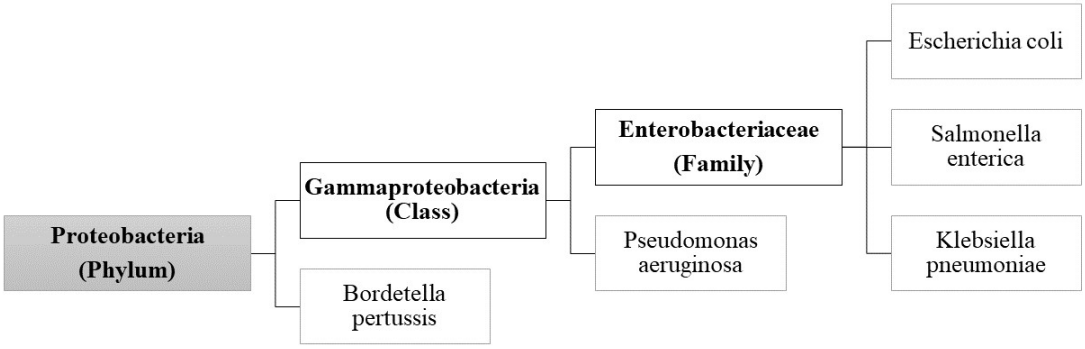


Figure 5.2: The taxonomic tree of the most frequent species in the data set

Table 5.1 and Figure 5.2 show that the data set is mostly about Proteobacteria species. However, it should be mentioned that there are other species that were from different phyla, which were also demonstrated during this research as one of the objectives of this research was to find GI patterns that were present in species and were significantly different in terms of taxonomy. It was noticed that the five most common bacteria species were common in the rod shape (i.e. bacillus) which means GIs can usually exist in bacillus bacteria.

5.2.1.1 Categorizing Proteins in the Genomic Islands

In this step of the pipeline, the focus is on generating the protein families of the protein sequences in the prokaryotic GIs to use them in the next stage of the pipeline in order to detect the patterns. The identification of the protein family of each protein sequence was performed using hmmsearch and pfam [39, 44] with the following command:

```
hmmsearch --cut_nc --tblout Pfam-A.hmm seqfile
```

hmmsearch is a biosequence analysis tool used to search for a protein sequence against a profile-HMM database such as Pfam. The seqfile is the input file that contains a number of sequences. In this research, two options were added to the hmmsearch as shown in the command. The first option is the –cut nc uses the noise cutoff bit score thresholds to set per sequence. The second option is the tblout option that saves the output in tabular format. After running hmmsearch, the result came to 5,584,117 sequences assigned to families. There were some sequences that were assigned to more than one protein family, and in this case, the match that had the lowest E-values was taken, which led to having 3,071,081 protein sequences. This number is derived from the E-value condition and another condition related to ribosomal protein families explained in the Section 5.2.1.2.

5.2.1.2 Genomic Island Filtering

In this section, an early analysis was performed to analyze the prokaryotic GIs to make an overview of the data set and how it would contribute to the research. At this step of the pipeline, GIs were treated as sets of protein families. Therefore, a clustering algorithm (i.e., BiMax algorithm [103]) was used to cluster these sets. There were notable sets in some clusters with almost the same content (i.e., protein families) but in different orders. These protein families are ribosomal protein families. To illustrate this point, the following is one example of these cases. In this example, the pattern is of size 13 protein families and exists in 1188 GIs (i.e., 222 Species):

- The 1188 GIs belong to 222 species, such as *Acholeplasma axanthum* (NZ_LR215048.1 GI_2), *Brevibacillus brevis* (NZ_LR134338.1 GI_4), and *Neisseria elongata* (NZ_CP031255.1 GI_1).
- The 13 protein families were: Ribosomal_L2_C, Ribosomal_L3, Ribosomal_L23, Ribosomal_L22, Ribosomal_L16, Ribosomal_L29 , Ribosomal_L14 , Ribosomal_L17, Ribosomal_S8, Ribosomal_L18p, Ribosomal_L30, SecY, Ribosomal_L27A

It was noticed that within each species, there was one GI that contained all of these protein families and this GI was predicted by the IslandPath-DIMOB tool for the most part. Therefore, based on the gene annotations, this looks like a ribosomal protein operon, so, in theory, this could be present in all cellular life. It is possible that this part of the genome had been incorrectly predicted as a GI on one or a few genomes, and the genes that come from this part of the genome are not horizontally transferred to the host genome, which means that they should not be part of a GI. GIs will frequently have a sequence composition that is considerably dissimilar compared to the host's genome due to the different genome sequence compositions found in various bacterial lineages. Genomic island predictors that apply the sequence composition technique are always heavily dependent on this fact. IslandPath-DIMOB predicts GIs by using the dinucleotide sequence composition bias technique along with the presence of mobility genes. Using the sequence composition bias technique to discover GIs could lead to many flaws. For example, some genes consider highly expressed genes, such as those genes in ribosomal-protein operons. These genes frequently have a sequence composition that is notably different from the rest of the genome. As a result, this leads to a false-positive prediction of the GIs [79]. Therefore, GIs containing more than 50% of ribosomal protein families were eliminated from the data set used in the research. Therefore, the total number of GIs after removing these GIs is 368,339.

5.2.2 Detecting Prokaryotic Patterns

In this section, stage two of the pipeline, the strategy for detecting prokaryotic patterns is presented. The strategy is divided into two steps; generate sets of protein families and retrieve the patterns from the sets.

5.2.2.1 Generate Sets of Protein Families

The first significant step of this research is to generate patterns from the GIs to discover the most frequent patterns in the GIs, which led to obtaining indications regarding the GIs structure. These patterns were generated from the protein families of the protein sequences in the GIs. The protein families of the protein sequences were chosen for two reasons. The first reason is that they are more general than proteins and there is a high possibility to get patterns in many GIs. The second reason is based on one of the research aims, which is to detect the structure of the GIs that made them GIs. This is related to the function of the proteins in the GIs, since there could be two different proteins that give the same function. Therefore, dealing with the protein families of the protein sequences was chosen as it achieves what is desired in this research.

In this section, the first step of generating the patterns was performed by generating the sets of protein families using the Apriori algorithm. The Apriori algorithm is a data mining technique which is a series of steps that are carried out to discover the most frequent itemset in a given database [2]. A minimum support threshold is a parameter used in the algorithm which refers to the frequency of occurrence of items.

To summarize, Apriori algorithm generated a large number of sets that exist in many GIs, but, as known, these sets were not in order. This means, for example, the "HTH_Tnp_1, rve" pattern exists in a number of GIs where rve could be before HTH_Tnp_1 or vice versa. Hence, this research aims to generate patterns that mean "protein families in order". Therefore, in the next section, the resulting sets from this step are used to obtain the patterns.

5.2.2.2 Retrieving Patterns in the Sets

In this section, the aim is to obtain patterns from the sets that were carried out using the Apriori algorithm. This was performed by creating an algorithm to generate all the possible patterns for each set from the generated sets in Section 5.2.2.1. Then, the promising patterns were selected under certain conditions, which is explained in the following paragraph.

Obtaining the promising patterns is performed using the *Patterns* algorithm (Algorithm 1). There are a number of conditions applied to the patterns to obtain the wanted patterns. First, patterns that existed in fewer than ten GIs were removed from the study. The second condition is related to the pattern size or the number of the protein families. It was observed that when patterns become large in size, they are biased towards a particular species, and when the pattern is very small, like a pattern with two protein families, they exist in a vast number of GIs due to them often being very common as well as there not being anything special about them most of the time. Therefore, according to the research data set, the second condition is that patterns in the middle of the different sizes, such as 3 to 5, are more interesting than patterns out of this range. The third condition is a taxonomy condition. Biologically, the taxonomy classification is divided mainly into seven levels from species to kingdom, where the order is in the middle of the classification. For this reason, the order was chosen as a taxonomy condition since it is in a fair position in the taxonomy classification. The most frequent orders were used as a condition for filtering the patterns because most of the GIs in the data set belong to these orders. This condition has the advantage of dealing with GIs that exist in species known to have GIs in their structure. With this condition, we are not aiming for GIs that exist in species that are not common to have GIs in their genomes. The aim is to deal with more reliable GIs. The second part of this condition added for orders is their number, which is seven. The analysis showed that the number of orders less than seven returns a large number of patterns (i.e., Hundreds). Here the aim is to select the most promising patterns. Therefore, the number of orders less than seven

biologically may indicate that these patterns are known to be present in the prokaryotic genomes. Furthermore, using the number of orders greater than seven did not retrieve any pattern. Applying the number seven as a condition for the number of orders leads to twenty patterns. Therefore, the number seven is the cut-off point through which the most promising patterns have been identified. In this step, it has been noticed that the most frequent protein families in the patterns are the phage protein families, which proves that these patterns could have originated from phages. The following section will describe a deep analysis of the generated patterns.

5.2.3 Patterns In-depth Analysis

In this section, stage three of the pipeline, the patterns went through an intensive analysis composed of a number of steps, starting with the first step of the analysis that showed a substantial presence of phages in patterns. Later, noteworthy HGT connections between bacteria and phages were detected. After that, HGT connections were refined to get the promising ones. Finally, the existence of a shared prokaryotic GI in the connections between the species was discovered.

5.2.3.1 The Substantial Presence of Phages in Patterns

In this section, as mentioned previously, there are indications that the origin of the GIs possibly came from phages since in the GIs structure there are proteins that belong to phage protein families. To prove this, the patterns were processed by the *Presence of Phages in the Patterns* algorithm (Algorithm 2). The algorithm is composed of a number of steps that lead to detecting if these patterns exist in phages. First, for each pattern, the algorithm retrieves the GIs that contain this pattern in their structure, which is performed in the same pattern order. In the second step of the algorithm, after collecting the GIs that have this pattern, the algorithm calls BLASTp for each GI. The query in the BLASTp command is the

proteins in the GI against the viral protein database that was downloaded from the NCBI. The BLASTp command is the following command:

```
BLASTp -db Viral_DataBase -query InputFile_GI_Proteins
       -out OutputFile_GI_Viral_Proteins -evalue 10e-10
```

The expectation value threshold (-evalue) for saving hits is 10e-10, which means only hits with an E value of less than this value are reported. In most cases, this command returns a number of matches for each protein or, in rare cases, no matches. A match in a case means that a GI protein matches with a viral protein or more than one viral protein. The BLAST results are saved in separate files based on the pattern and the GI that has this pattern (i.e., Filename: PatternName_GINmber). The BLAST analysis showed that all the bacteria proteins in the patterns have one or more viral protein matches. This is a vital sign and may indicate a strong association of these proteins with phages. These resulting files from this step are used in the following analysis step of the pipeline to detect HGT connections.

5.2.3.2 Detecting Noteworthy Horizontal Gene Transfer connections between Bacteria and Phages

In this step of the analysis, there are two algorithms implemented to generate the HGT connections; the *Patterns (Phage Information)* algorithm (Algorithm 3) and the *Extracting HGT Connections* algorithm (Algorithm 4).

In the *Patterns (Phage Information)* algorithm (Algorithm 3), for each file from the previous step, the proteins of the GI were checked to make sure that all of the proteins that belonged to the protein families of the pattern existed in the GI. The successful GIs, which were the GIs where all of their proteins had a viral match, went through the next step of the algorithm. For each file, all of the proteins in the file were parsed to generate the genome

accession number of the viral proteins using the EFetch function in the Entrez package. After generating the genome accession of all viral proteins, in the next part of this step in the algorithm the aim was to find the common viral genome accession among all these proteins in the GI, i.e., retrieving the viral genome accession that exists in all proteins in this particular GI that have this particular pattern in their structure. The aim of this step is to detect a viral genome accession that has a number of proteins that match the same number of bacteria proteins in the GI. The viral genome accession was chosen instead of the viral genome name because it is more specific, and in the next steps, more analysis was performed on these viral genomes. These viral genome accessions were reported considering the largest Evalue among all the proteins in the GI. While at the beginning of this step there is an input file for each GI for a pattern (i.e. PatternName_GINumber), at the end there is an output file that contains all GIs for a pattern (e.g. PatternName). In each pattern file, there were lines where each one contained a GI and a viral genome that matched the GI with the pattern as well as the E-value, which shows the quality of this match. In each pattern file, a viral genome could exist in more than one line, and, therefore, matches with more than one GI. A viral genome and these GIs are common in a specific pattern (i.e. Pattern file). The output pattern files from the previous step are used as input files in *Extracting HGT Connections* algorithm (Algorithm 4) wherein each pattern file the aim is to list all of the viral genomes in a file, and for each viral genome to list all the matched bacteria GIs along with their E-values. This is because a deep analysis is performed on phage and the related GIs with an aim to detect the most interesting Phage-Bacteria matches. All of this is performed to detect the HGT connection between these species. In general, as shown in *Extracting HGT Connections* algorithm (Algorithm 4), for each pattern file, a new file is generated where each line in the file represents a phage and a list of all GIs that match this phage along with their E values, which was performed to all file patterns.

5.2.3.3 Refining Horizontal Gene Transfer Connections

Considering the previous step, in the pattern file, a line represents a viral genome and a list of all bacteria GIs along with their E-values that matches a viral genome with a specific pattern. This means that each line in the file represents a possible HGT connection between a viral genome and a group of bacteria. This group of bacteria could be closely related. An example of this connection is a phage, which infects a specific bacteria species (i.e. *Bacillus phage AP631* and *Bacillus paralicheniformis* bacteria) and this HGT connection is already known in the literature and, therefore, not the target of this research. On the other hand, this group of bacteria could be distantly related. However, they could be related in terms of genome content and not yet discovered. The objective of this research is to detect the most interesting phages that infect bacteria that are from different lineages. This is performed by using *Filtering HGT Connections [Taxonomy]* algorithm (Algorithm 5), where for each line in the pattern file, the algorithm obtains the common lineage of the bacteria by obtaining the lineage for each bacterium and then computing the intersection between them. Later, the result is filtered to take solely the lines with a lineage length of less than five, which means the bacteria could be from the same order or higher. The order is used in this step for the second time as a condition for the same reason as it being at the middle level of the taxonomy. This process is performed for each pattern file and the result of all files is saved in one file where each line in the file represents:

- The pattern that exists in phage and bacteria GIs.
- The phage name and accession number.
- The list of GI numbers and the species accession numbers that hold these GI.
- The E values that show the quality of the match between the phage and GIs.
- The common lineage of all the bacteria with a lineage length of less than five.

The result from this step of the algorithm showed a large number of connections between phages and bacteria that showed evidence of HGT either directly or indirectly, where the quality of these connections is known according to the E-values. Since the research objective is to focus on the most promising results, another filter was applied to keep solely the bacteria with small E-values because bacteria with small E-values have more evidence of HGT with the phage than other bacteria that have moderate E-values on the same line. In this research, E-values that are smaller than 10e-100 were considered as small E-values, therefore, any line containing a bacteria with an E-value greater than or equal 10e-100 was removed from the line. This research contains a large number of results, however, the focus of this research was on the most promising ones; this is the reason many filters or cutoffs were applied. At the end of this step, there is another file containing the same line structure as the previous step with only bacteria containing E-values of less than 10e-100.

The next step of this section is to dive deep into each pattern by studying their proteins to discover the connection between each phage and group of bacteria in the same connection by using *Connections* algorithm (Algorithm 6). This is performed for each line by retrieving the phage proteins and bacteria proteins before using BLAST to compute the percent Identity, E-values, and the scientific name that holds the subject protein. In BLAST command, the query sequence is the phage proteins and the subject sequence is the bacteria GI proteins. It should be noted that the scientific name could be a specific species name or higher, such as a family or an order. The scientific name of the BLAST result should be the same name as the bacteria GI that has this protein that is provided in the line, otherwise, the protein is eliminated. Therefore, for each specific line, the BLAST result of all the proteins belonging to the pattern with a match between the phage and the bacterium GI should give the scientific name of the bacterium GI and not the name of the family or a higher level that the bacterium belongs to to keep this phage-bacterium GI connection. Furthermore, the percent identity value that describes how similar the query protein sequence is to the subject protein sequence should be greater than or equal to 85%, as the higher the percent identity

is, the more significant the match is. Also, the phage and bacteria should be unrelated, which means the phage should not be a well-known phage that infects the bacteria.

The aforementioned method is the analysis between the phage and the group of bacteria for each specific line in the file to give information about the HGT between them. However, on the other hand, another investigation was also performed between the group of bacteria on the same line. In this case, the proteins for each two bacteria are BLASTed sequentially and in reverse to make sure that the resulting scientific name is the same as the bacteria that hold the two GIs containing the two proteins. The aim of this step of the analysis is to study the connection between all of the species on the line to discover how they are related and which species are more related to each other. All of the species in the same line are studied together because the research input, which is the GIs that have evidence that is horizontally transferred from another species, is already published. Therefore, the line that exists in the file is also evidence that something has been discovered in this research, where there is a high possibility of HGT between these species.

5.2.3.4 Discovering the Existence of a Shared Prokaryotic Genomic Island in the connections

At this step of the pipeline, a BLAST analysis is performed to study the connection species to understand what they have in common by analyzing their genomes, not just the GIs. The BLAST analysis shows that the resulting alignment sequence between a phage and a bacteria genome has almost the same coordinates as the coordinates of the GI in the bacteria genome on the IslandViewer4 website. To further investigate the similarity between the GIs in the same connection, the multiple sequence alignment was performed. MUSCLE is the multiple sequence alignment tool that has been used in this stage of the analysis. The *Connections (Extract Species Subsequences)* algorithm (Algorithm 7) was used to do the BLAST analysis and get the MUSCLE input sequences in each connection.

5.3 Results

In this section, the results are divided into three main subsections. The first part shows the results of sets of protein families that were generated from the GIs. Later, the patterns that are extracted from these sets are presented. Finally, the discovered connections from the patterns are illustrated.

5.3.1 Protein Families

This section presents a deep analysis of the data set protein families. Furthermore, it shows the results of sets of protein families that were generated using the Apriori algorithm to be used in discovering the patterns.

5.3.1.1 A Deeper look at Protein Families

In this section, the protein families in the data set were analyzed. The GIs in the data set are composed of protein sequences and these protein sequences belong to protein families from 9,121 different protein families.

Figure 5.3, shows that the GIs in the data set have protein sequences that belong to protein families ranging from one protein family to 194 protein families. This means the GIs in their structure have a number of protein families ranges from one protein family to 194 protein families. Furthermore, the figure shows that most of the GIs in the data set have a number of protein families ranging from 1 to 9 in their structure. It is worth mentioning that most of the GIs, which equates to 51,883 GIs in the data set, have three protein families in their structure. Moreover, there is an outlier GI in the data set that has 947 protein families and this GI belongs to the Lactobacillus curvatus KG6 strain (NZ_CP022475.1). Lactobacillus curvatus KG6 demonstrated phenotypic novobiocin resistance when it was isolated on de Man-Rogosa-Sharpe agar from a fermented meat product similar to salami that was bought

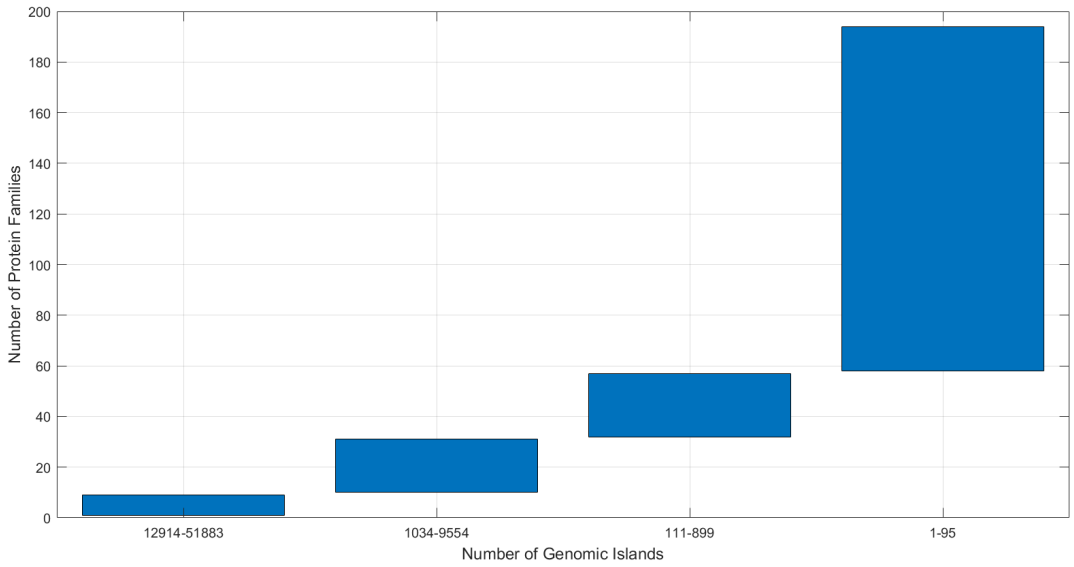


Figure 5.3: The bars represent the range of protein families among the categories that are represented by ranges of GIs

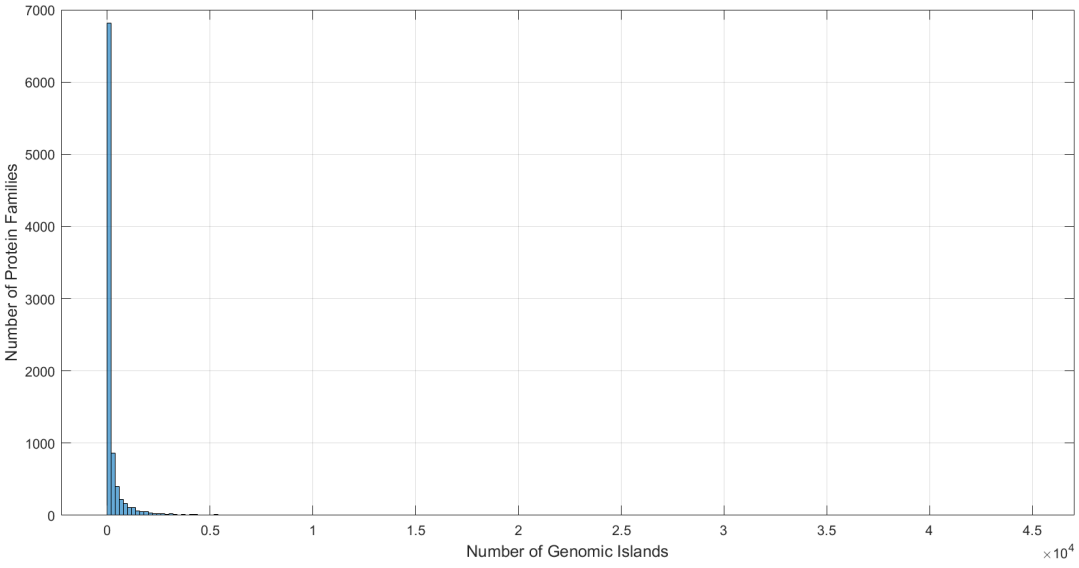


Figure 5.4: The histogram represents the distribution of protein families in genomic islands. Each bar represents the number of protein families in the genomic islands. In the beginning, there is a peak representing the number of 6,816 protein families present in a number of genomic islands ranging from one to 200. In general, the chart shows that most protein families are present in a number of genomic islands, less than five thousand.

from a retailer in Switzerland in 1999 [47]. In general, it has been noticed in the data set that the majority of bacteria genera that have species containing GIs in the range between

58 and 194 are Salmonella, Escherichia, and Bacillus (in order).

The data set has a large number of protein families that are diverse in function. Figure 5.4 shows a histogram that represents the occurrence of protein families in GIs. In Figure 5.4, most of the protein families in the data set exist in 2,500 GIs or less. The protein families that are interesting are the ones that exist in most GIs.

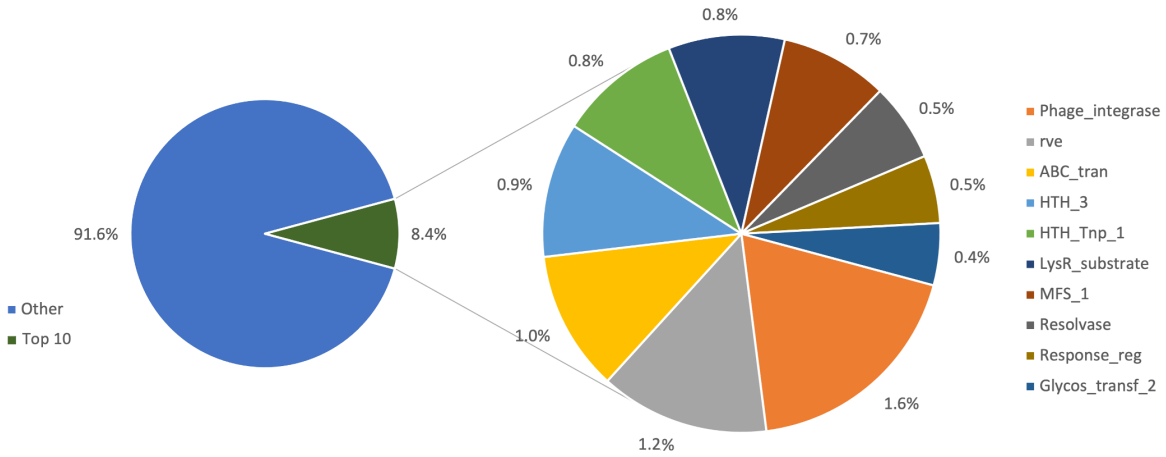


Figure 5.5: The ten most frequent protein families that exist in the GIs. Each percentage represents the proportion of occurrences of the protein family in the genomic islands.

Table 5.2: Information about the ten most frequent protein families in the GIs

Protein Family	GIs	Pfam Clan	Gene Ontology (Molecular Function)
Phage_integrase	44,605	DNA breaking-rejoining enzyme	DNA_binding
rve	32,624	Ribonuclease H-like	-
ABC_tran	26,987	P-loop containing nucleoside triphosphate hydrolase	ATP_binding
HTH_3	26,010	Helix-turn-helix	-
HTH_Tnp_1	23,636	Helix-turn-helix	DNA_binding, Transposase_activity
LysR_substrate	22,401	Periplasmic binding protein	-
MFS_1	20,781	Major Facilitator	Transmembrane_transporter_activity
Resolvase	15,074	-	DNA_binding, DNA_strand_exchange_activity
Response_reg	13,068	CheY-like	-
Glycos_transf_2	11,942	Glycosyl transferase	-

Figure 5.5 shows the top ten frequent protein families in the data set. More information about the top ten protein families can be found in Table 5.2. The table shows the name of the protein family, the number of the GIs that have this protein family in their structure, the

pfam clan (i.e. superfamily) of the protein family, and the gene ontology about the protein family.

According to the table, there are ten protein families with numerous functions, starting with the phage_integres protein family that exists in 44,605 GIs. Phage integrases are enzymes whose main function is to catalyze the site-specific recombination between two sequences (i.e., DNA); the bacterial and phage attachment sites. Phage integrases are now well renowned due to the impact they have made on the scientific community, especially in the in vitro GATEWAY cloning [50]. Moreover, the ability that phage integrases possess for efficiently and specifically recombining sequences (i.e. DNA) in living cells makes it likely that they would be useful in numerous genetic engineering applications [50].

Below that, the rve protein family, short for the retroviral integrase. The retroviral integrase enzyme plays a significant role in a crucial phase in the replication cycle of viruses. By means, It is responsible for inserting a copy of the viral genome into the DNA of the host [8]. Under high-salt conditions, retroviral intasomes have a great deal of resistance to challenges [87].

Under the rve protein family, there is the ABC_tran protein family. The ATP-binding cassette (ABC) transporters are known for being a large-sized superfamily of membrane proteins that contain diverse functions [58]. ABC transporters can be found in prokaryotes and play numerous roles for them. They represent efflux proteins and influx proteins. Efflux proteins are responsible for removing toxins from the cell, whereas influx proteins are responsible for transporting nutrients into the cell [42]. To the present day, chemotherapy failure caused by ABC drug efflux is considered an ongoing research topic that frequently contributes additional evidence on multiple drug resistance, allowing scientists to tackle and overcome this matter. The efflux of drugs from the cell is the main mechanism of resistance performed by the membrane transporters that is found among all organisms. The membrane transporters are proteins that belong to the ABC transporter superfamily [138]. Therefore, one of the fundamental causes of chemotherapy's unsuccess is the drug resistance that the prokaryotic

ABC transporter family causes. Furthermore, bacterial cells also contain various families of ABC transporters which contribute to resistance to antibiotics [35].

The HTH_3 domain and HTH_Tnp_1 protein family both fall under the same superfamily called HTH, which consists of a large variety of principally DNA binding domains that include a helix-turn-helix motif. Many derivatives of the helix-turn-helix motif play a role in multiple antibiotic resistance, including the DNA-binding domain discovered in multiple antibiotic resistance regulators that form winged helix-turn-helix. Regarding HTH_Tnp_1, this protein family stands for helix-turn-helix transposase. The HTH structure is related to DNA binding, while transposase is needed for DNA transposition. Simply, HTH_Tnp_1 proteins bind to the nucleotide and assist in transposition of DNA. The HTH_Tnp_1 family contains many *E. coli* Insertion Elements (IS) along with various other bacterial transposases where some are members of the IS3 family, which can operate as a mobile promoter in E. coli [29].

The next protein family is the LysR_substrate protein family. The LysR_substrate family is a member of the clan Periplasmic Binding Proteins (PBPs). PBPs are nonenzymatic receptors that are used by bacteria to pick up small molecules and carry them into the cytoplasm. Most PBPs take part in transporting solute molecules to the cytoplasm by way of ABC transporters [125], where they aim for critical nutrients such as vitamins, amino acids, carbohydrates, and ions. Macheboeuf et al in [85] mentioned that PBPs play role in drug resistance.

Below that, the MFS_1 protein family can be seen, which belongs to the Major Facilitator Superfamily (MFS), it is one of the largest families of membrane transporters [89]. Since they take part in an essential role in several diseases by means of drug transportation, drug resistance, or aberrant action, members of the MFS family are central to the physiology of humans. The issue of resistance to antibiotics is often due to the action of MFS resistance genes [45]. It has also been observed that mutations in MFS transporters can cause diseases

such as neurodegenerative disease [5], and glucose storage diseases [100]. Regarding MFS_1 in bacteria, the Arsenic Resistant (AR) bacteria species have evolved numerous efflux systems for AR. One of the arsenic efflux proteins that has been reported is MSF_1 [115]. Many MFS proteins are submitted to UniProtKB, such as efflux pump antibiotic resistance (TARUN_5198) and MFS multidrug transporter (MFS_1).

Regarding the Resolvase protein family, in bacteria, resolvase proteins are commonly found on mobile DNA elements, such as plasmids and transposons. Antibiotic resistance genes are frequently associated with Tn3 family transposons [94]. Tn3 family transposons also contain a tnpR resolvase gene [99].

Regarding the Response_reg protein family, the response regulators protein families form the central family of signaling proteins in prokaryotes. A protein from the response regulator's protein family assists the cell of the bacteria in responding to the environment changes. This is performed by enabling the bacteria to sense, respond, and adapt to numerous environments [117]. There are some types of response regulators that are related to the resistance to antibiotics and protonophores. Concerning antibiotic resistance, QseB and BfmR response regulator proteins are responsible for degrees of antibiotic resistance in *Acinetobacter baumannii* and *Francisella novicida*, respectively [90]. Regarding protonophores resistance, the response regulator YcbB in *Bacillus subtilis* is one of the proteins whose disruption raises cell resistance to protonophores [105].

Finally, the Glycos_transf_2 protein family, glycosyl transferases (GTs), catalyse the formation of several kinds of glycoproteins with crucial roles in cell-to-cell recognition and communication [18]. One of their important roles is to catalyze the transfer of sugar onto aglycons, and this has a substantial association for the synthesis of natural products with high value [110]. Glycans play significant roles in many biological processes in disease and health. There is a close relation between human and bacteria in the intestine, and this relation can be pathogenic or symbiotic. Bacterial GTs are considered as one of the virulence

factors to humans. Therefore, understanding GTs has a significant role in vaccine production to protect against bacterial infections. Liposaccharide antibiotic moenomycin that inhibits bacterial glycosyltransferases is considered as a promising lead [149].

Regarding the Gene ontology, as shown in the table, not all Pfam families map directly to GO terms. As shown in the table, from the gene ontology view, the common molecular functions of the protein families are binding and transpose.

Table 5.3: The two most frequent protein families that exist together in the GIs

Protein Families	**The Number of GIs**
HTH_Tnp_1, rve	11,521
HTH_3, Phage_integrase	9,730
BPD_transp_1, ABC_tran	7,100
ABC_tran, rve	5,882
HATPase_c, Response_reg	5,512
IstB_IS21, rve	5,218
MFS_1, LysR_substrate	5,209
HTH_17, Phage_integrase	4,744
Glycos_transf_1, Glycos_transf_2	4,699
rve, Phage_integrase	4,666

Table 5.3 shows the top two protein families of the GIs in the data set, which means the most frequent two protein families that exist in the GIs together. Simply, as shown in Table 5.3, proteins that usually belong to transpose protein families and binding or insertion protein families come together, such as in (HTH_Tnp_1, rve), (HTH_3, Phage_integrase), (ABC_tran, rve), (MFS_1, LysR_substrate). Furthermore, another case of the two protein families with the transpose and insertion or binding functionalities is the BPD_transp_1 and ABC_tran. BPD_transp_1 is a family that is a member of the clan BPD_transp_1, which is a clan containing families that are included in the transport of molecules across membranes. Regarding IstB_IS21, rve protein families, proteins belonging to the protein family IstB_IS21 contain an ATP/GTP binding P-loop motif. This motif is found to be linked to the IS21 family insertion sequences [120]. In general, The protein function is unknown, but there is the possibility that it can perform a transposase function [139]. There

are some cases when there are two protein families with almost the same molecular function, such as binding or insertion in the two protein families (rve, Phage_integrase), and (HTH_17, Phage_integrase).

Furthermore, in the (HATPase_c, Response_reg) case, the two protein families mainly function in sensing any changes in the environment and responding to these changes. HATPase_c belongs to the Histidine kinases clan and members of this superfamily are essential components of regulatory systems that allow bacteria to react to changes in their environment. Finally, Glycos_transf_1, Glycos_transf_2 protein families, overall, are from a large family of enzymes called Glycosyl transferases, as mentioned previously, they catalyze the transfer of sugars to a numerous of accept or molecules, which are important in all domains of life.

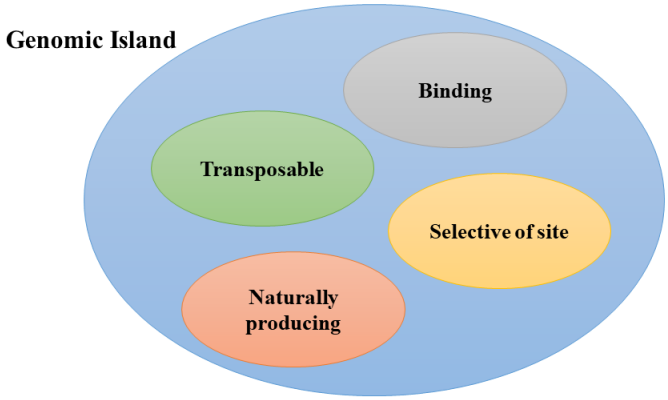


Figure 5.6: The main functional components in the GIs

To summarise, from all of the aforementioned information about the protein families that exist in the GIs in the data set, it is highly possible that the main components of the GI are Transposable, Selective of site, Binding, and Naturally producing, as shown in Figure 5.6. This makes sense as these are the main components that need to transfer a subsequence from one genome to another. The transpose function that is found in proteins belonging to protein families transposes facilities such as the MFS_1 protein family. After, there is the selection of site. The function of the proteins in these protein families is to specify the location of the subsequence in the host genome. An example of such a protein family is the rve. Regarding binding, which assists in binding the subsequence in the host genome,

proteins with this function belong to protein families such as phage_integres protein family. Finally, there is the naturally producing, which could be possibly required for the success of the whole transfer process. An example of a protein family is Glycos_transf_2. It should be noted that the most important components are transposing and binding.

5.3.1.2 Protein Families Sets

The Apriori algorithm was chosen for this research to obtain the most frequent sets in the GIs with a support value equating to 0.006. This support value was chosen according to system space capabilities and another reason related to the type of species in the data set, which will be explained in the following paragraph.

According to this support value, there were six hundred forty sets of protein families in GIs with sizes between 2 and 7 protein families resulting from Apriori algorithm. It was noticed from the results that when a set gets larger, the number of species that have this set gets smaller. Therefore, the smaller the support value, the smaller the number of species, which could lead to a specific species. This means the generated sets with a very small support value could be in the most frequent species in the data set, such as *Escherichia coli*. This conflicts with the target of the research, which is to detect patterns that exist in numerous species, not specific species. Therefore, the chosen support value is sufficient for this research and data set.

Table 5.4, shows the most frequent two sets of each size and the support value for each pattern along with the number of species and genera that have this set. For example, the first row represents a set of two protein families that exists in 3% GIs out of 368,339 GIs, along with 1,102 species and 438 genuses.

In general, Table 5.4 shows numerous protein families in the patterns that are from different

Table 5.4: The two most frequent protein family sets of each size

Support	Size	Sets of Protein Families	Species	Genera
0.031277115	2	HTH_Tnp_1, rve	1,102	438
0.02641525	2	Phage_integrase, HTH_3	1598	548
0.012076719	3	PapD_N, Fimbrial, Usher	206	42
0.01195907	3	Phage_tail_L, Lambda_tail_I, Phage_min_tail	190	60
0.008897242	4	Terminase_1, HNH, Phage_portal, Phage_capsid	465	178
0.008344289	4	Terminase_1, Phage_H_T_join, Phage_portal, Phage_capsid	471	186
0.007664864	5	Phage_tail_U, Phage_tail_T, Minor_tail_Z, Phage_TTP_12, Phage_TAC_2	48	20
0.00701485	5	Phage_H_T_join, HNH, Terminase_1, Phage_portal, Phage_capsid	371	148
0.006526605	6	Phage_tail_U, Phage_min_tail, Phage_tail_T, Minor_tail_Z, Phage_TTP_12, Phage_TAC_2	42	16
0.006438368	6	TMP_2, Phage_tail_U, Phage_tail_T, Minor_tail_Z, Phage_TTP_12, Phage_TAC_2	45	18
0.006097184	7	Phage_TTP_12, Phage_tail_U, Phage_tail_T, Minor_tail_Z, Phage_min_tail, Phage_tail_L, Phage_TAC_2	41	16
0.006100126	7	Phage_TTP_12, TMP_2, Phage_min_tail, Phage_tail_U, Minor_tail_Z, Phage_tail_T, Phage_TAC_2	42	16

molecular functions. However, this table shows the protein families that most, if not all, of them have a very significant role in the horizontal gene transfer process. Starting with the phage tail protein families, the tail protein of phages is significant for the interaction between host bacteria and phages. Phage tail proteins are responsible for host cell recognition and delivery of the viral genome to the host cytoplasm. For example, there are a number of minor tail proteins that have enzymatic activity. This type of activity assists the phage in recognizing the correct host. Furthermore, pass through the cell wall or surface to inject the DNA. The other important protein family is the Phage_portal family. The proteins of this protein family form a portal (i.e., hole or channel) that enables DNA passage during packaging and ejection. During the bacteria infection process, the protein family Phage_capsid plays a significant role in the success of this process since Phage_capsid proteins protect the viral genome during entry and exit from the host cells. Regarding Phage_integrase and rve proteins, they are responsible for the integration of a DNA copy of the viral genome into the host genome. Terminase_1 or Phage Terminase, the majority of the members of this family are phage proteins. In general, Terminase protein is a key component of the DNA packaging machine found in phages. Finally, the His-Asn-His (HNH) protein family is a common protein family that is mainly associated with endonuclease activity. It is a protein

family that is composed of small nucleic acid-binding proteins. Endonucleases are enzymes that cleave the phosphodiester bond within a polynucleotide chain.

Table 5.5: The ten most frequent protein families in the sets

Protein Family	The Number of Sets Exist in
Phage_min_tail	152
Phage_TAC_2	137
Phage_tail_T	134
Phage_TTP_12	128
Phage_tail_L	123
Phage_tail_U	121
TMP_2	100
Minor_tail_Z	100
Lambda_tail_I	80
Phage_capsid	52

It was observed that the most frequent protein families in all sets are phage protein families, as shown in Table 5.5. This could give an indication that the origin of the GIs could be from the phages.

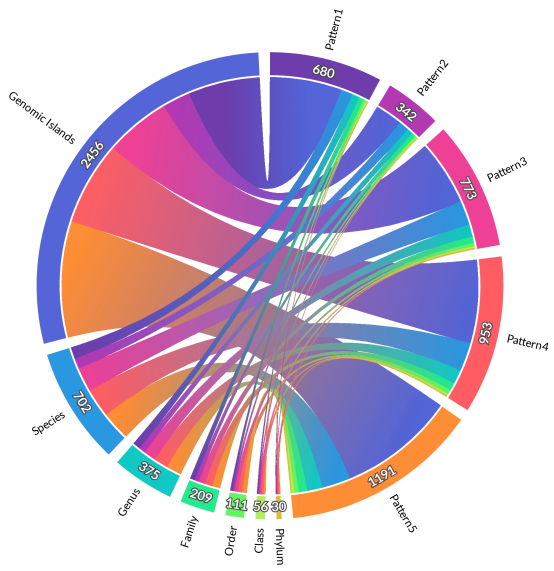
5.3.2 Patterns

The initial resulting patterns at this part of the analysis was 33,448 patterns from the 640 original sets. The *Patterns* algorithm (Algorithm 1) filters the resulting patterns to get the most promising ones. There were 20 resulting patterns from the algorithm, as shown in Table 5.6. The interesting patterns are from size three except for the last one, which is of size four. This number of patterns agrees with the initial analysis of the protein families in Section 5.3.1.1. The protein analysis section shows that the largest proportion of GIs in the data set contains three protein families in their structure. The table shows that phage protein families exist in almost all the patterns, which proves that these patterns could have originated from bacterophages.

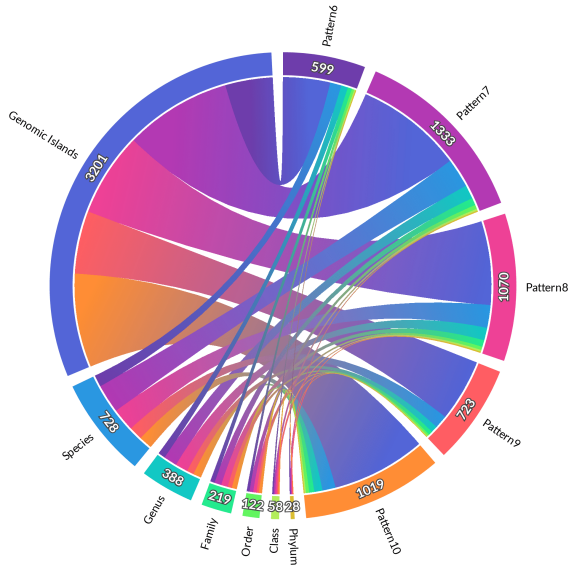
Table 5.6: Patterns

Index	Pattern Size	Pattern
1	3	Phage_capsid HK97-gp10_like HNH
2	3	Terminase_1 HK97-gp10_like HNH
3	3	Terminase_1 Phage_capsid HK97-gp10_like
4	3	Phage_H_T_join Terminase_1 HNH
5	3	HNH Phage_portal Phage_capsid
6	3	Phage_capsid HNH Phage_portal
7	3	Phage_capsid Phage_portal HNH
8	3	HNH Terminase_1 Phage_capsid
9	3	Terminase_1 Phage_capsid HNH
10	3	Phage_capsid HNH Terminase_1
11	3	Terminase_4 Phage_capsid HNH
12	3	Phage_connect_1 Phage_portal HNH
13	3	HNH Terminase_1 Phage_portal
14	3	Terminase_4 Terminase_1 HNH
15	3	HTH_Tnp_1 rve Phage_integrase
16	3	Phage_integrase HTH_Tnp_1 rve
17	3	Terminase_1 Phage_portal Phage_capsid
18	3	Terminase_1 Phage_capsid Phage_portal
19	3	Phage_capsid Phage_portal Terminase_1
20	4	Terminase_1 Phage_capsid HK97-gp10_like HNH

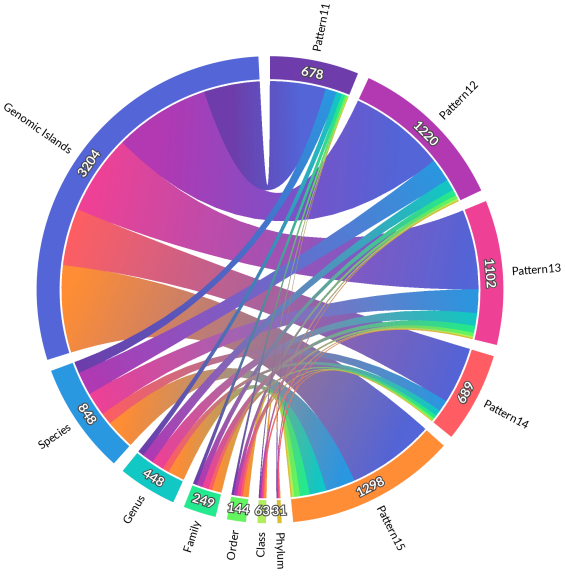
Table 5.6 shows that the most frequent protein families in the patterns are HNH, Phage_capsid, Terminase_1, and Phage_portal. The HNH protein family exists in many phages, and the HNH proteins location in the phage genome is next to the terminase proteins, and it is highly conserved [143]. There are many studies that have shown that the presence of HNH and terminase proteins together is essential for phage activities [68]. This proves that they also have a significant role together in the HGT process. Regarding Phage_capsid and Phage_portal protein families, they both play an essential role during the phage infection process, as mentioned previously in Section 5.3.1.2. Therefore, it is natural if one of them is present in a sequence, then the other protein is also exists. The table shows that when these protein families exist in a particular pattern, they are present in a different arrangements, as shown in the colored rows. This may indicate that their presence is vital in the GI in any order. In general, the presence of these four protein families in the patterns may indicate that these are the essential protein families that lead to the success of the HGT process.



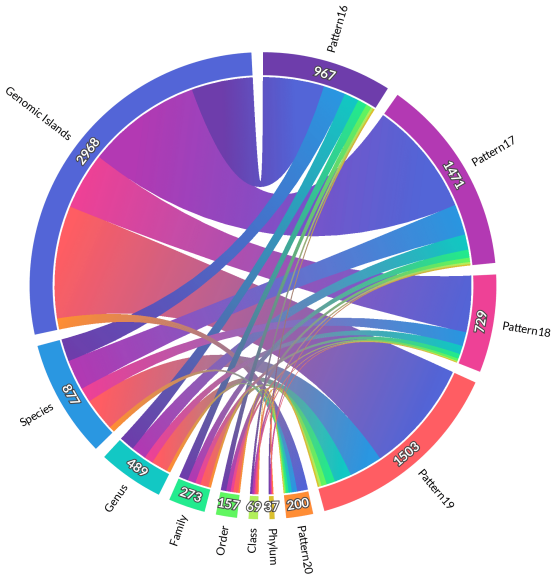
(a) **Patterns 1 - 5**



(b) **Patterns 6 - 10**



(c) **Patterns 11 - 15**



(d) **Patterns 16 - 20**

Figure 5.7: The chord diagrams show information about the taxonomy levels for each pattern

In regards to taxonomy, Figure 5.7 shows chord diagrams of the taxonomy levels for each pattern. Each chord diagram shows the taxonomy information for a number of the patterns. For each pattern, the diagram illustrates the number of GIs that have the pattern in their structure and the number of species, genera, families, orders, classes, and phyla in these

GIs. The superkingdom value was omitted since it is one for all the patterns. Overall, the figure shows that these patterns are diverse since they exist in numerous phyla, classes, orders, … etc. This figure could indicate that the discovered patterns may strongly represent an essential part of the GIs. This could be because the patterns contain phage protein families, and these patterns exist in many organisms as shown in the chord diagrams and also illustrated in more detail in Figure 5.8.

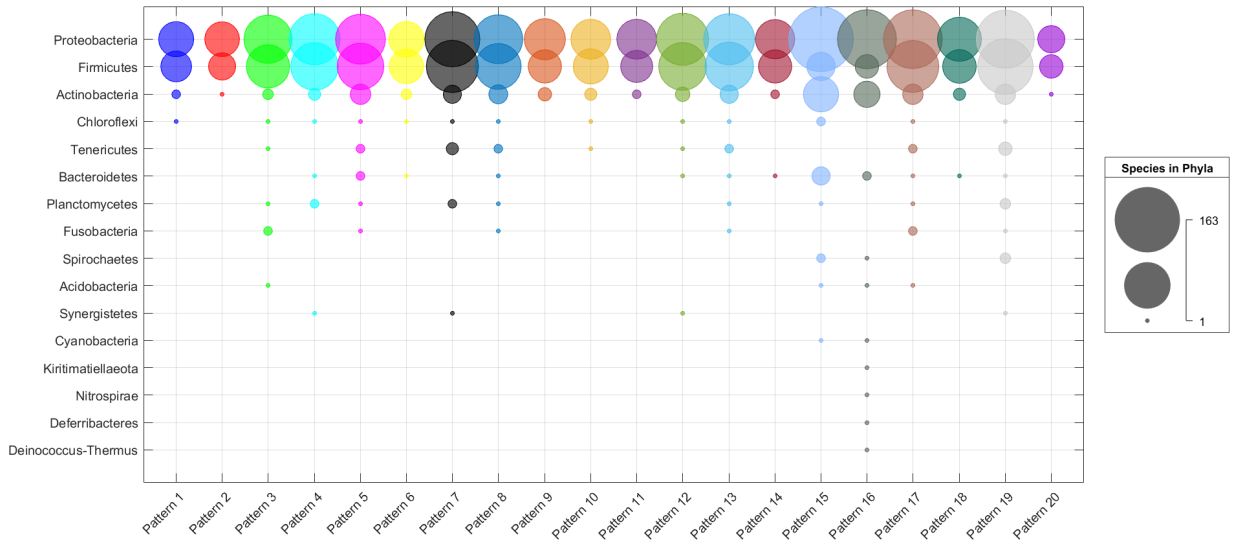


Figure 5.8: The bubbles represent the number of species in a phylum in each pattern

Figure 5.8, is a bubble chart that shows all the patterns in the x-axis and the phylum name in the y-axis. A bubble for a pattern P and phylum Phy represents the number of species (i.e. GIs) that belong to phylum Phy and have the pattern P in their structure. It is evident from the chart that all the patterns exist in species from the Firmicutes, Proteobacteria, and Actinobacteria phyla. The presence of the Proteobacteria phylum is expected because, as previously mentioned most of the GIs in the data set are from species that belong to the Proteobacteria phylum. Furthermore, the Proteobacteria phylum includes various pathogens, such as Escherichia and Salmonella, which play a significant role in the HGT process in prokaryotic. The presence of Firmicutes and Actinobacteria in all patterns may indicate that their GIs have a structure similar to the Proteobacteria GIs structure. This proves that

there is a specific structure of GIs in prokaryotic. It is worth mentioning that numerous Firmicutes produce endospores. Endospores are highly resistant to ultraviolet light, desiccation, heat, and chemicals. They can survive extreme conditions. Firmicutes bacteria can be found in various environments, and they include some pathogens. Regarding Actinobacteria, species belonging to this phylum can be found in many environments with extreme conditions such as salty seas, boiling hot springs, and extreme arctic cold. Some members of the Actinobacteria are pathogenic such as Streptomyces that consider as infrequent pathogens [30] Figure 5.8 shows that pattern number sixteen (Phage_integrase, HTH_Tnp_1, rve) exists in many GIs that are from eleven phyla, and pattern number nineteen exists in many GIs that are from ten different phyla. These two patterns could be essential in GIs since they exist in numerous phyla.

One of the patterns is (Phage_integrase, HTH_Tnp_1, rve). This pattern exists in species that belong to eleven different phyla, and that leads to them being dissimilar from one another, with each one having its own unique characteristics and living in different environments. Most of the species that have this pattern belong to the Proteobacteria, Actinobacteria, and Firmicutes phyla as shown in Figure 5.8. Firstly, there is the *Candidatus Solibacter usitatus* which belongs to the Acidobacteria phylum. In the data set the *Candidatus Solibacter usitatus* strain that has this pattern is Ellin6076. *Candidatus Solibacter usitatus Ellin6076*, showed an abundance of genes that are affiliated with mobile genetic elements [26]. Furthermore, after comparative genome analyses, it was revealed that the Ellin6076 large genome came into being by HGT through the ancient phages or/and other processes [27]. The next phylum is Spirochaetes, where from this phylum, the species *Leptospira mayottensis* has this pattern in its genome. The *Leptospira mayottensis* species is a pathogenic species that comes from the genus Leptospira and is isolated from humans [17]. It was discovered in 2014 and is linked to various diseases [33]. Another species that has this pattern is *Kiritimatiella glycovorans*, which is from the Kiritimatiellaeota phylum. *Kiritimatiella glycovorans* is a species discovered in 2016. In general, the Kiritimatiellaceae family is composed of mainly

bacteria which is found in environments that are hypersaline and anoxic. The species of Kiritimatiellaceae have a gram-negative cell wall that contains peptidoglycan (PNG) [121]. Gram-negative bacteria are the most prevalent primary pathogens. They possess multiple cell surface glycans that have demonstrated their importance in the process of the biosynthesis and regulation of the cell wall of pathogenic Gram-negative bacteria [69]. The mechanical strength and shape of bacterial cell is owed to the peptidoglycan, which and can also play a role in the pathogenesis [14]. As a greatly conserved and vital constituent of almost every bacterial cell, PG is identified as a pathogen-associated molecular pattern by the eukaryotic immune system and is a potent activator of innate immunity [119]. Next, there is the species *Pleurocapsa* sp. PCC 7327 that belongs to the Cyanobacteria phyla. In general, the Cyanobacteria are found to be amongst the most diverse and widely dispersed phyla of bacteria [116]. Moreover, they are essential contributors to global nitrogen fixation [142]. Found in a variety of marine environments, the genus Pleurocapsa is a nitrogen-fixing, spore-forming cyanobacterium, which can grow in either freshwater or saline environments. The other phylum that has species with this pattern is Nitrospirae, whose official name is *Nitrospira moscoviensis*. *Nitrospira moscoviensis* was discovered in 1995 and is a non-motile, non-marine, gram-negative, nitrite-oxidizing bacterium that has a curved rod shape [40]. The cytoplasm of *Nitrospira moscoviensis* contains polyhydroxybutyrate (granules), which are crucial storage compounds for carbon and energy in a number of prokaryotes, allowing for the survival of the cells when there is an absence of appropriate carbon sources [40]. The *Salinivirga cyanobacteriivorans* species that belongs to the Bacteroidetes phylum is another species that has this pattern. The species named *Salinivirga cyanobacteriivorans* was initially described by Ben Hania et al. 2017 [56]. It is a species of bacteria which preys on cyanobacteria. In general, the bacteria species of the Cyanobacteria phylum are gram-negative that obtain their energy by means of photosynthesis. Cyanobacteria also known as Cyanophyta, Deferribacteres phylum has a species named *Denitrovibrio acetiphilus* that has this pattern in its genome sequence. The species *Denitrovibrio acetiphilus* belong to

the bacterial family Deferribacteraceae and bacterial genus Denitrovibrio [75]. *Denitrovibrio acetiphilus* has a curved rod-shaped structure and is a gram-negative, mesophilic, marine, anaerobic bacterium that respires by means of nitrate reduction. Its capacity to reduce nitrate is of economic significance, as it eradicates the requirement for expensive biocides that are presently used in the treatment of oil reserves [92]. Next, there is the *Deinococcus psychrotolerans* species that belongs to the Deinococcus-Thermus phylum. In 2019, *Deinococcus psychrotolerans* was identified [127]. It is a coccus-shaped, non-motile, gram-negative, strictly aerobic bacterium. The species of the genus Deinococcus are characterized by their resistance to ionizing radiation.

After that, there is the species *Achromobacter denitrificans* from the Proteobacteria phylum. Formerly, the *Achromobacter denitrificans* species was known as *Alcaligenes denitrificans*. It was recently reclassified under the Achromobacter genus [24]. The *Achromobacter denitrificans* species are gram-negative, motile, strictly aerobic, and ubiquitous bacterium. The strains of *Alcaligenes denitrificans* exist in soil; however, they can sometimes exist in human clinical samples since they can cause human infections [32]. Later, in the Actinobacteria phylum it has been discovered that the *Gordonia bronchialis* species have this pattern. Gordonia species are Gram-positive and aerobic actinomycetes that have recently been recognized for causing human disease. The species that belong to the Gordonia genus are gram-positive, catalase-positive, nonmotile, and aerobic [9]. There are numerous Gordonia species that have been isolated from soil [114]. Furthermore, the bacterium species *Gordonia bronchialis* was discovered in a number of patients with sternal wounds, dogs, and skin [137] [66]. Numerous laboratories for clinical microbiology could misidentify Gordonia species for Rhodococcus and Nocardia species. This is due to the fact that the species that belong to the Gordonia genus are closely related to the species that belong to the Rhodococcus genus and Nocardia genus. In the data set there species from Nocardia or Rhodococcus that have this pattern.

Finally, there is the species *Halanaerobium hydrogeninformans* from the phylum Firmicutes. *Halanaerobium hydrogeninformans* is an obligatory anaerobic, gram positive, non motile,

elongated rod-shaped bacterium, that is able to tolerate exceptionally high salinity and high alkalinity conditions within its environment [19]. Most commonly found in halo alkaline lakes, *Halanaerobium hydrogenoformans* is an alkaliphilic bacterium that can carry out biohydrogen production. Microorganisms are currently being explored as a means of chemical and biofuel production, such as future biohydrogen generation at industrial scales. Owing to the increase in price of fossil fuels and diminishing of reserves, biofuel production is viewed as a viable contribution to present and future energy demands. *Halanaerobium hydrogeninformans* are able to use a number of pure sugars for hydrogen production, and, consequently, this bacterium is potentially capable of raising the level of efficiency and efficacy of biohydrogen production that comes from renewable biomass resources [19].

In conclusion, instead of demonstrating every single pattern in detail, this one example serves to demonstrate that, like in all other patterns, a pattern contains species that have differences on many points. In general, the Figure 5.7 and Figure 5.8 show that the patterns can be found in various species and each pattern is present in at least three different phyla of bacteria.

5.3.3 Connections

A connection reflects an HGT relation between a phage and a number of prokaryotic species. This section presents the connections and the analysis of the phages in the connections.

5.3.3.1 Details on Identified Connections

This section show the resulting connections resulted from *Presence of Phages in the Patterns* algorithm (Algorithm 2), *Patterns (Phage Information)* algorithm (Algorithm 3), *Extracting HGT Connections* algorithm (Algorithm 4), *Filtering HGT Connections [Taxonomy]* algorithm (Algorithm 5) and *Connections* algorithm (Algorithm 6). Table 5.7 contains the

information about the resulting connections.

Table 5.7: Each row represents a connection. Rows colored with the same color mean that there are different phages have a connection with the same group of bacteria.

Index	Pattern	Phage Name	Phage Genome	Bacteria Genomes
44	Terminase_1 Phage_portal Phage_capsid	*Bacteriophage* sp.	MN855837.1	NZ_CP026116.1, NZ_AP017931.1
54	Terminase_1 Phage_portal Phage_capsid	*Enterobacteria* phage SfI	NC_027339.1	NZ_CP030787.1
58	Terminase_1 Phage_portal Phage_capsid	*Enterobacteria* phage mEp235	NC_019708.1	NC_009778.1
64	Terminase_1 Phage_portal Phage_capsid	*Escherichia* phage Henu7	LR881103.1	NZ_CP035214.1, NC_015968.1, NZ_CP020388.1
73	Terminase_1 Phage_portal Phage_capsid	*Klebsiella* phage ST13-OXA48phi12.2	MK422452.1	NZ_LT556084.1
92	Terminase_1 Phage_portal Phage_capsid	*Salmonella* phage ST64B	NC_004313.1	NZ_CP030787.1
94	Terminase_1 Phage_portal Phage_capsid	*Shigella* phage SfII	NC_021857.1	NZ_CP030787.1
96	Terminase_1 Phage_portal Phage_capsid	*Shigella* phage SfIV	NC_022749.1	NZ_CP030787.1
200	Terminase_1 Phage_portal Phage_capsid	*Uncultured Caudovirales* phage clone 7S_14	MF417956.1	NZ_CP035214.1, NC_015968.1, NZ_CP020388.1
201	Terminase_1 Phage_portal Phage_capsid	*Uncultured Caudovirales* phage clone 2AX_6	MF417957.1	NZ_CP035214.1, NC_015968.1, NZ_CP020388.1
383	Phage_capsid Phage_portal Terminase_1	*Klebsiella* phage ST13-OXA48phi12.2	MK422452.1	NZ_CP037894.1, NZ_CP032841.1
385	Phage_capsid Phage_portal Terminase_1	*Klebsiella* phage ST846-OXA48phi9.1	MK416021.1	NZ_CP020448.2, NZ_CP038469.1
571	Terminase_1 Phage_portal Phage_capsid	*Klebsiella* phage KPP5665-2	MF695815.1	NZ_LS992183.1, NZ_CP020820.1
573	Terminase_1 Phage_portal Phage_capsid	*Klebsiella* phage ST16-OXA48phi5.3	MK416014.1	NZ_LS992183.1, NZ_CP020820.1

Continued on next page

Table 5.7 – continued from previous page

Index	Pattern	Phage Name	Phage Genome	Bacteria Genomes
1385	HNH Phage_portal Phage_capsid	*Lactobacillus* phage Sha1	NC_019489.1	NZ_CP021927.1

Table 5.7 shows the pattern index number, the pattern protein families, the phage name, the phage genome accession number, and the bacteria genome accession numbers that have this pattern in common.

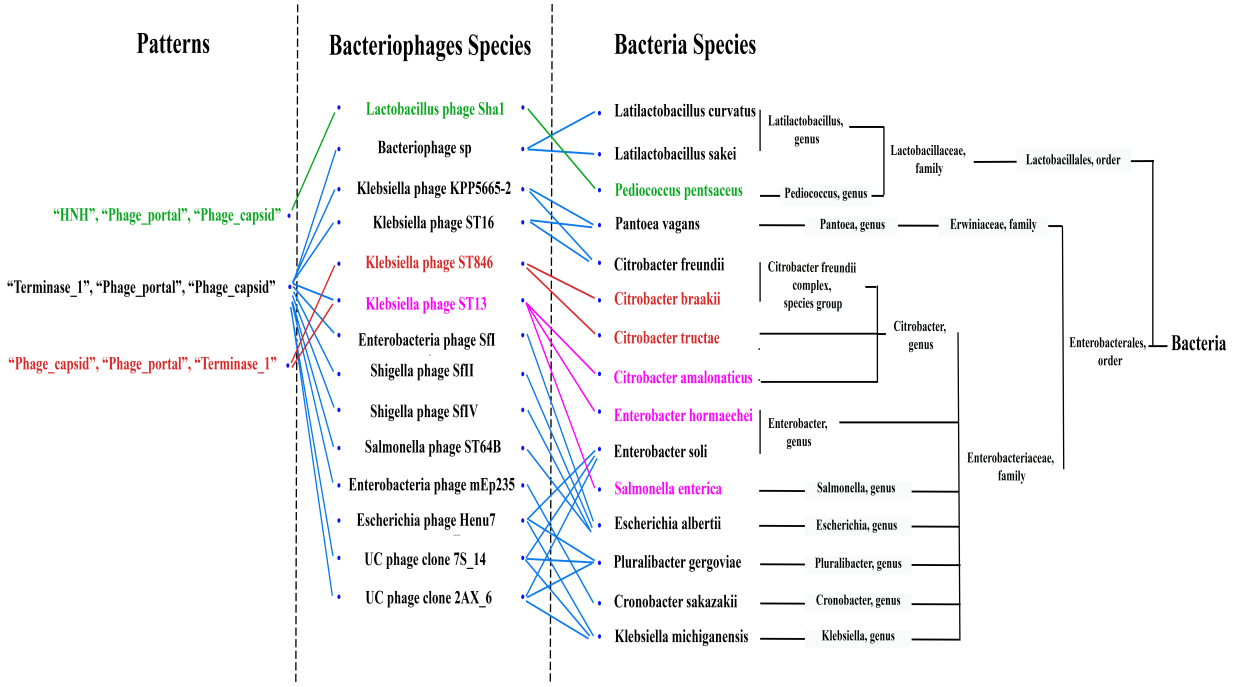


Figure 5.9: A tripartite graph of the connections. Going from left to right, the first set represents the patterns, the next set represents the phage species, and the following set shows the bacteria species. In the patterns part, each pattern has a color and a colored link between any two sets shows that these species have the same pattern. For example, the red pattern exists in one phage, which means one connection that is composed of one phage and two bacteria that are from different species. A phylogenetic tree of the bacteria species is added on the right side that represents the taxonomy relation between the bacteria.

Figure 5.9, shows Table 5.7 content as a tripartite graph with three independent sets that represent the three main parts of each connection. Formally, let Tri_G=(Tri_P, Tri_Ph,

Tri_B, E_PPh, E_PhB) be a tripartite graph with the following sets:

$$Tri\_P = \{tri\_p1, tri\_p2, tri\_p3\}$$
$$Tri\_Ph = \{tri\_ph, \ldots, tri\_ph14\}$$
$$Tri\_B = \{tri\_b1, \ldots, tri\_b15\}$$

The Tri_P set represents the patterns, next comes the phage species set (Tri_Ph), then the bacteria species set (Tri_B). The edges between the sets are represented as E_PPh and E_PhB, where E_PPh represents the edges between the patterns and the phages, while E_PhB represents the edges between the phages and bacteria. A connection in the graph is represented as an edge between the pattern set and the phage set as well as another edge(s) between this phage and a bacterium or more than one bacterium. For example, the pattern ("HNH", "Phage_portal", "Phage_capsid") found in the *Lactobacillus phage Sha1* phage and the *Pediococcus pentosaceus* bacterium is represented in the color green. The taxonomy of bacteria species has been added to the figure to show the phylogenetic relation between the bacteria. As shown in the phylogenetic tree of bacteria species, for some connections, their bacteria are distantly related. For instance, in the pattern (Pattern: "Terminase_1", "Phage_portal", "Phage_capsid".) the bacteria *Pantoea vagans* and *Citrobacter freundii* are from two different families but in the same connection with *Klebsiella phage* KPP5665-2.

In this research, there are many connections, as shown in Figure 5.9,where a lot of exciting information can be seen in each one. One of the connections that deserves research and scrutiny is one that consists of two compounded connections where the phage named *Klebsiella phage* ST13-OXA48phi12.2 can be found, colored in pink as a result of the two patterns blue and red being combined. The *Klebsiella phage* and bacteria in this connection share two patterns in their genome. However, despite belonging to the same bacteria family, these bacteria species are from different genera.

The bacteria in this compound connection are *Citrobacter amalonaticus*, *Enterobacter hormaechei*, *Salmonella enterica* and all have their own special characteristics. The *Citrobacter*

amalonaticus belongs to the Citrobacter genus, which is a genus of aerobic, gram-negative bacterium that is part of the Enterobacteriaceae family. Citrobacter species have been found in sewage, water bodies, the human gut, as well as in animal intestines [91] [109]. The existence of Citrobacter species in sewage water is seen as a significant public health threat, especially in areas heavily populated with humans, due to the capacity of these bacteria to spread efficiently and quickly once an infection takes place. Infection from Citrobacter species is associated with urinary tract infections, gastroenteritis, and neonatal meningitis [82]. *Citrobacter amalonaticus*, or *Levinea amalonatica*, was initially described by Young et al. in 1971. The *Citrobacter amalonaticus* Y19 strain was reported to have the capability of producing hydrogen from oxidizing toxic carbon monoxide [4]. Furthermore, it was demonstrated that, like *E. coli*, *Citrobacter amalonaticus* Y19 also possesses an FHL complex [74].

Regarding *Enterobacter hormaechei*, it is a species of gram-negative, oxidase-negative bacteria which is extensively found in the majority of temperate soils and waters [97]. Overall, Enterobacter species are widely scattered across many areas such as soil, water, vertebrate and invertebrate hosts, and in the feces of animals and humans. *Enterobacter hormaechei* is also an opportunistic infectious pathogen, as it can be present within the intestinal tract of humans causing diseases in immunocompromised hospital patients [36] [98]. Moreover, *Enterobacter hormaechei* infections have been discovered in a variety of mammals, such as pets with respiratory disease complex [73], and calves with respiratory diseases [134].

Finally, the *Salmonella enterica*, in general, Salmonella is a gram-negative bacteria with a rod-shaped appearance and facultative anaerobes that belong to the Enterobacteriaceae family. Salmonella species have the ability to multiply and grow under numerous environmental conditions outside the living hosts. Salmonella genus, which is closely related to the genus Escherichia, is known as the most prevalent foodborne pathogen that is often detached from food-producing animals, causing zoonotic infections in humans and various animal species including birds. Thus, infections from Salmonella are a major concern to public health, the health of animals, as well as the global food industry [6]. The intestinal tract of humans and

farm animals is Salmonella serovars' main niche. However, it can also be found in the intestinal tract of reptiles, wild birds and less commonly insects [7]. Salmonella is divided into two different groups, *Salmonella bongori* and *Salmonella enterica*. *Salmonella enterica* species is a food-borne pathogen known for causing mild to serious human diseases, such as mild gastroenteritis and severe systemic infections. One of the major disease burdens worldwide is the human infections caused by *Salmonella enterica* through contaminated water or food. *Salmonella enterica* is like the other salmonella species that cause foodborne illness globally and is accountable for significant public health as well as economic damages. When testing for *Salmonella enterica* in food, there are usually issues caused by the existence of background microflora that could present themselves as Salmonella which is false-positive [46]. It is often very challenging to tell the difference between false-positive isolates that belong to the Citrobacter genus and Salmonella genus owing to the similarities in their cell surface antigens, genetics, as well as other phenotypes [102]. The core genome analysis shows that the Citrobacter and *Salmonella enterica* populations investigated appear to share a common evolutionary history. As with Salmonella, Citrobacter is often isolated from water, soil, and animals' digestive tract [15]. In general, an ambiguous connection between *Escherichia coli*, *Citrobacter*, and *Salmonella* is apparent, which presents a practical issue in the field of medical diagnosis and food safety testing, since the identification of accurate species is crucial to confirm the existence of pathogens [102].

Overall, in this compound connection, it is obvious that what bacteria species have in common is their presence in the same locations, such as water or some types of organisms (i.e. intestinal tract). It is important to mention that these bacteria have a serious effect on many kinds of living organisms including humans and various kinds of animals. Furthermore, when one of such bacteria species is present in the body of a living being, whether it be a human or an animal, it is usually concentrated in a specific area of the body. From these points that were mentioned, it is clear that these bacteria in this compound connection may have a strong relation, which deserves to be studied in depth. This deep analysis may assist answer

questions about specific diseases related to these bacteria because these bacteria are in the same connection with a particular phage (i.e. *Klebsiella phage* ST13-OXA48phi12).

5.3.3.2 Analysis of the Phages in the connections

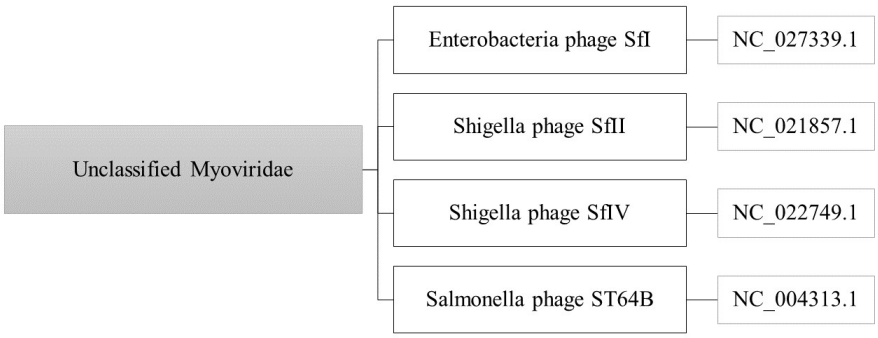


Figure 5.10: Case 1 phages phylogeny

In this section, the phages in the connections were analyzed. Table 5.7 shows that there is a group of phages in a connection with the same bacteria species. For example, there is a connection between Phage MF695815.1 and bacteria NZ_LS992183.1, NZ_CP020820.1. Furthermore, the same bacteria exists in another connection with phage MK416014.1. In Table 5.7 there are three cases where a group of phages is in HGT connection with the same bacteria species. Therefore, phylogenetic trees and BLAST analysis were performed to analyze the phages and understand the relation between these phages. Starting with the first case, blue rows, Figure 5.10 shows the phylogeny connection between the phages in the first case in the table. The figure shows the four phages are from different species. All the species belong to the same taxonomy level named unclassified Myoviridae. Unclassified Myoviridae is a level with no rank under the Myoviridae family.

Connections (Phages BLAST Analysis) table (Table A.1) in the appendix shows the BLAST analysis for all the phages in Table 5.7. The BLAST analysis shows that the E-value is zero between all the phages regarding the first case. The second case (i.e., orange) in Table 5.7 includes two partial genomes that belong to the same species; *uncultured Caudovirales*

phage. In the *Connections (Phages BLAST Analysis)* table (Table A.1), the idetity value is 99.982% and the E-value equated to zero. Moreover, the query coverage is around 100%. It is worth mentioning that the difference between the length of the two phage genomes (MF417956.1=16570 bp, MF417957.1=16537 bp) is very small. Finally, the last case include species from the same taxonomy level named unclassified Siphoviridae. In the BLAST analysis, as shown in the *Connections (Phages BLAST Analysis)* table (Table A.1), the identity values above 98% and the E-values equated to zero.

Overall, it is obvious from the coverage values in each case that there is a common subsequence between the phages in each case. This sequence could be the GI in each connection in the phage and bacteria. More information about the shared subsequence (i.e. GI) is presented in Section 5.3.4.

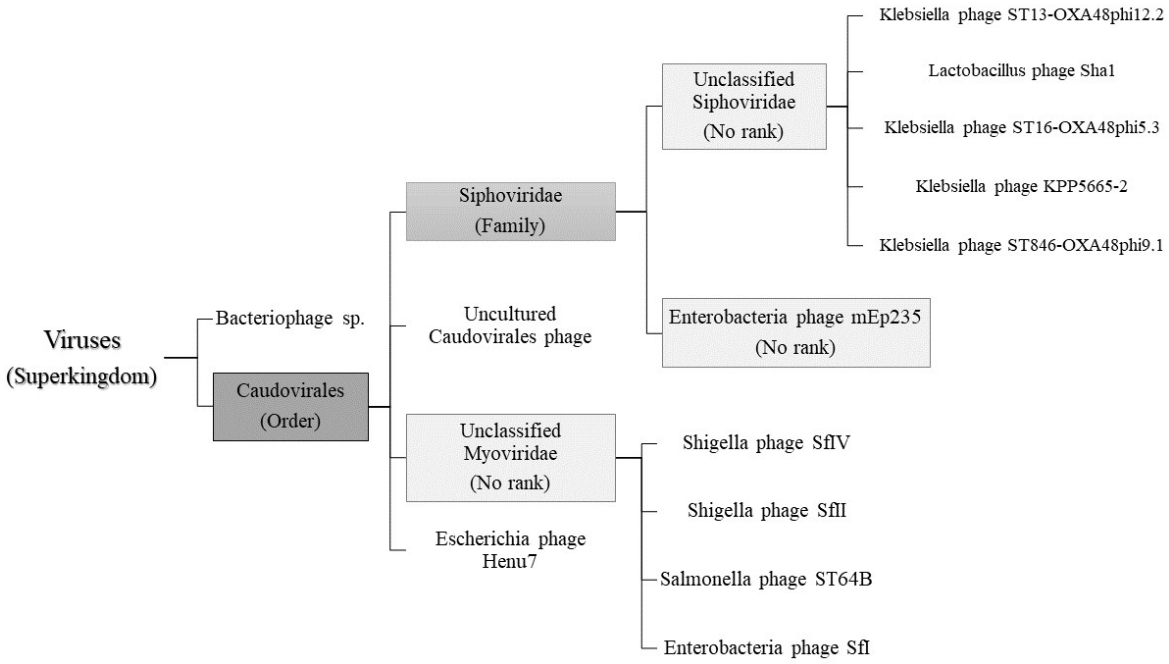


Figure 5.11: Phylogeny of the phages in the connections

The phylogeny of the phages in the connections is presented in Figure 5.11. The phylogeny shows that most of the phages belong to the *Caudovirales* order except *Bacteriophage sp.*. In general, it is clear that there is a connection between phages in each case, and this connection

deserves further investigation in the future.

5.3.4 The Existence of a Shared Prokaryotic Genomic Island in the connection Species

The HGT connections analysis was performed using BLAST and MUSCLE. The analysis aims to investigate more about the connections.

The *Connections (Phages and Bacteria BLAST Analysis)* table (Table A.2) in the appendix shows the BLAST result between the genomes of the species within each connection. The E-value for all the BLAST results is zero, therefore, it is not included in the table. The identity values in general are around the 80's and 90's, however, there are few 70's cases. The *Connections (GIs Coordinates Information)* table (Table A.3) in the appendix also shows the BLAST between the genomes for each connection with more information regarding the alignment included. The *Connections (GIs Coordinates Information)* table (Table A.3) includes the length of the two genomes; the query length (Q length) and subject length (S length), the overlap or alignment length (A Len), the start and the and of query alignment (Q Start, Q End), the start and the and of subject alignment (S Start, S End), the start and the end of the GI that is discovered in the bacteria genome (i.e. subject genome) as mentioned in the IslandViewer4 website.

Connections (GIs Coordinates Information) table (Table A.3) shows that the resulting alignment sequence between the phage and the bacteria genome has almost the same coordinates as the coordinates of the GI in the bacteria genome in the IslandViewer4 website. In more details, the analysis shows that for each connection C, when BLASTing the phage genome H against the genome of bacterium B_z, the resulting alignment (i.e. subsequence) coordinates are similar to the GI coordinates in the bacterium B_z. Moreover, for a connection C the resulting phage H coordinates against the bacteria $\{B_1, B_2, \ldots, B_d\}$ in most of the cases

is almost the same coordinates. Therefore, from the aforementioned two observations, an assumption is created, which is that the resulting subsequences between the phage H and the bacteria $\{B_1, B_2, \ldots, B_d\}$ are actually only one GI that is common between them. This means the genomes in the same connection C share the same GI, but the GI is not identical in all genomes in the same connection because the GI has been changed over time, which could be due to the mutations and movement of the GI between organisms.

The MUSCLE result of the GIs shows a good alignment between the sequences in each connection. This is evidence that the sequences in the connection are the same sequence, which is the GI. The phylogenetic trees of the connections that composed of more than one bacteria species are presented in figures in the appendix. The phylogenetic trees show the evolutionary relationship among the species GIs.

Overall, in most of the cases, in the *Connections (Phages and Bacteria BLAST Analysis)* table (Table A.2) the results are similar to the phylogenetic trees. For example, the table shows that if the phage X genome is more related to the bacteria species Y genome than other species in the same connection. Therefore, in the phylogenetic tree, the GI of phage X also exists in the same clade with the GI of species Y. The table result could mean that there are also other common areas between the phage X and bacteria species Y. Furthermore, the result in the phylogenetic tree is evidence that there is an HGT between the phage X and bacteria species Y. This HGT event could be direct from the phage X to the bacteria Y, or another bacteria Z in between. Regarding bacteria Z, the phage X could be the phage of the bacteria Z, and the GI of this bacteria is not shown in the connection could be because it was removed from the filter in Section 5.2.3.3.

5.4 Discussion

The prokaryotic GIs in this chapter were analyzed in order to investigate their structure in terms of PFs in a way that has not been previously addressed intensively. This chapter aims to study GIs structure in-depth to extract patterns from them that were used to obtain biological connections between prokaryotes and phages.

In this study, the protein families were used as the basic unit for the research instead of proteins because they are more informative than proteins, and the GIs can be distinguished according to the biological function that they give the host genome. Furthermore, there is also a high probability of knowing the GI pattern when dealing with protein families rather than something more specific like proteins because more than one protein can give them the same biological function since they belong to the same protein family. The analysis on protein families showed that most GIs contain three protein families in their structure, which may indicate that there are three essential components or protein functions that must be present in GIs. Furthermore, the results showed that the main components of the GIs might be binding, transposable, selective of site, and naturally producing. This makes sense as these are the main components that could be needed to transfer a GI from one genome to another. Moreover, the analysis showed that most GIs exist in rod shape species. It is known that among all the shapes of bacteria species, each one demonstrates different physical features and appearances to the outside world. Therefore, there could be a relation between the HGT and the rod shape. A possible factor could be the surface area to volume ratio as the rod shape gives a broad surface area per unit volume, which could assist in facilitating the attachment to bacteria and subsequently help to transfer the genetic material [28].

When looking into the literature on the five most common bacteria species in the data set (i.e. Table 5.1), it was noticed that in Africa they were common pathogens shuttling antibiotic resistance genes. It was observed that there was transmission of the same strains across borders and between macroscopic creatures [112]. This is consistent with the data set

used in this research and they are highly probable pathogens shuttling antibiotic resistance GIs and not just genes.

The patterns show that the GIs have a specific structure that makes them sub-sequences of particular genes that have a crucial role in changing the biological function of the host species. Furthermore, patterns are found in bacteria species that are distantly related. This means that the HGT frequently happens across a substantial part of the bacterial communities between closely related species and distantly related species. Most of the generated patterns are of size three. This number of protein families, or pattern size, is expected because in the initial analysis of protein families, it was proven that the most significant proportion of GIs in the data set contains three protein families in their structure. It's worth mentioning that all patterns contain several species that differ in phyla, and in each pattern, there are at least three species that each is from different bacteria phyla. It has also been observed that most of the resulting patterns contain phage protein families, and this is a significant sign that these GIs could originate highly from phages. Phages play a critical role in regulating bacterial populations and nutrient cycling. Bacteria and phages contain a sophisticated network of interactions where continued coexistence and survival of phages and their host bacteria is permitted. Bacteria and phages constantly battle in a manner where bacteria build up a resistance to phages, while phages develop methods to tackle that resistance [22]. Evidence of such battles dates back to three billion years ago, which has given rise to a vast amount of diversity in phage and bacterial genomes by way of HGT and novel mutations [122].

In the in-depth analysis of a pattern, the resulting connection consists of a phage, a group of bacteria, and a common pattern between them. The group of bacteria in each connection are distantly related. However, the BLAST analysis showed that they are similar in genome content. Furthermore, the common pattern between the phage and the group of bacteria assists in discovering a GI that is common between them. This GI is the same as the GI that is provided in IslandViewer4 for each bacteria in the connection. Therefore, species in the same connection have a very high probability of HGT between them, even if they are

distantly related. This means the implemented pipeline discovered the GIs in the bacteria species by using the patterns and connections, and also discovered the phages that have HGT relations with the bacteria. Therefore, the pipeline discovered patterns and used these patterns to discover connections that led to discover the GIs. Discovering the patterns and connections was expected since we are dealing with an ecosystem where everything happens for a reason, but discovering the GIs from the connections was a positive an unexpected find.

The resulted connections can be used in future work to determine the origin and direction of the GIs. A connection containing species of different lineages that share the same content is uncommon and requires an intense investigation to find out the mystery of this connection. It is essential to know that these connections were reached through a large number of filters that led to the deletion of a large number of species to achieve a solid biological connection composed of phage and a group of bacteria. These connections also contain other species related to them but not strongly associated with them. They were omitted because their genomic content did not meet the strict filtering conditions. This filtering was performed to discover the species that are very close to each other in terms of genome structure. Therefore, understanding the connections is understanding the relationship of many other species, not just the species in the connection.

A deep analysis of the connections could assist in understanding the relationship between bacteria species and the phage as well as the relationship between the bacteria together. This means this deep analysis produces promising results by answering many biological questions. For example, in Section 5.3.3, the investigation of the compound connection that contains three bacteria species (i.e., *Citrobacter amalonaticus*, *Enterobacter hormaechei*, *Salmonella enteric*) showed that these bacteria could have severe effects (i.e., diseases) on humans and animals. Furthermore, when one of such bacteria species is present in the body of a living being, it usually exists in a specific area of the body. Therefore, a deep analysis of this connection could assist in understanding the cause of the diseases from these bacterial species and finding the proper solution. As future work, the common GI in the HGT connections

could be studied further to bring to light the origin and direction of the GI in each connection. This analysis will play a significant role in understanding prokaryotic evolution and, more specifically, in discovering the origin species of the prokaryotic GIs. Knowing the species that are the origin of GIs assists in comprehending the species that play a vital role in the HGT in prokaryotes. This will assist in focusing future research on these species since these species could play a significant role in diseases and drug and antibiotic resistance. In almost all cases, treating the source of the problem is always the most effective way of solving the mystery behind unsolved natural questions. Biologically, the protein families in the patterns and the species in the connections could be studied further to answer many of these questions, especially the ones related to prokaryotic evolution.

Chapter 6

Investigating the Nature of Genomic Islands' Locations within a Genome

6.1 Problem Definition

In the second research direction, the aim is to study the GIs' structures in terms of their location in the genome. This is performed by analyzing the GIs to understand their location features from different perspectives.

As mentioned in Section 5.1, in the data set there are a number, N, of genomes where each genome contains a number of GIs. Furthermore, a GI j in a genome i, represented as $I_{\{i,j\}}$, therefore, a genome i contains a number u of GIs:

$$G_i = \{I_{\{i,1\}}, I_{\{i,2\}}, \ldots, I_{\{i,u\}}\} \tag{6.1}$$

The number of GIs, u, vary from one genome to another. It is worth mentioning that the GIs in the genome could overlap. Therefore, an assumption is added in this direction, which is that GIs that overlap are considered as one GI in order to avoid redundant GIs in the genome. For this reason, in this direction, GIs are reformulated, and the resulting GIs are used in the analysis. The first step of the analysis is to study the GIs' location in relation to the origin of replication. Later, Section 6.3.2 presents an investigation about the nature of the distances between the GIs. Finally, in Section 6.3.3, the location distribution of GIs

in the genome is discovered.

6.2 Genomic Islands Preprocessing

In this section, the GIs have been analyzed and reformulated before performing the investigation performed in Section 6.3.

6.2.1 Genomic Islands Analysis

In this section, the GIs are briefly analysed in order to understand the broad view of the GIs in the data set. In the first step of the analysis, the aim is to discover the number of the GIs in the genomes by understanding what the maximum number of GIs in a genome is as well as the minimum number of GIs within a genome. Furthermore, the aim is to discover the length of the genomes and GIs in the data set.

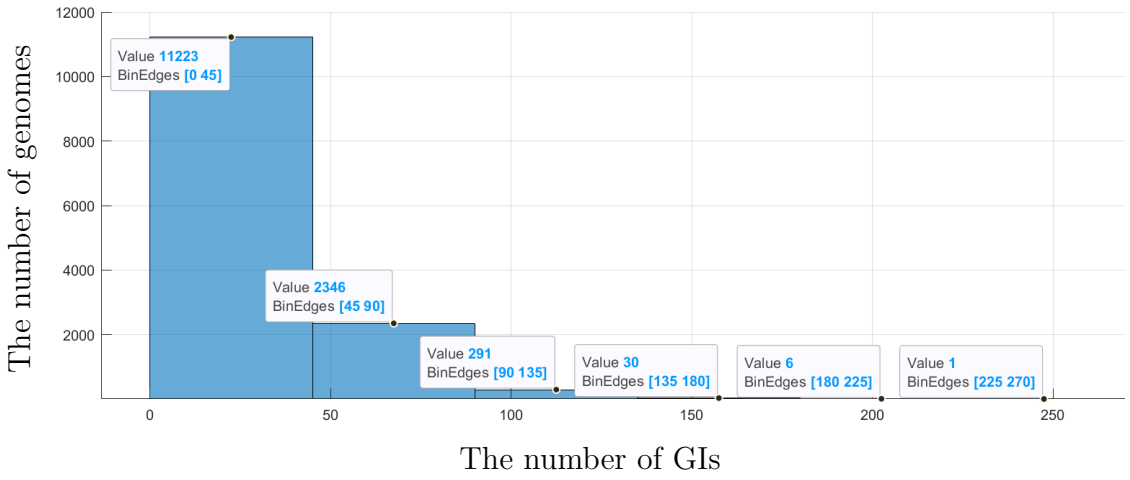


Figure 6.1: The number of the GIs in the genomes

The histogram in Figure 6.1 answers the first aforementioned question. As shown in Figure 6.1, the x-axis demonstrates that the range of the GIs is from 1 to 270 in the genomes, whereas the y-axis demonstrates the number of genomes. This means the minimum number

of GIs in the genome is one, while the maximum number of GIs in a genome is 270. Overall, it is obvious from the figure that there are many genomes with few GIs, while there are also many GIs in few genomes. For example, the fist bar shows that there are 11,223 genomes that have GIs in the range of 1 and 45. Furthermore, there is only one genome that has a number of GIs in the range of 225 and 270, which is actually the genome NC_011894.1 containing 270 GIs in its structure.

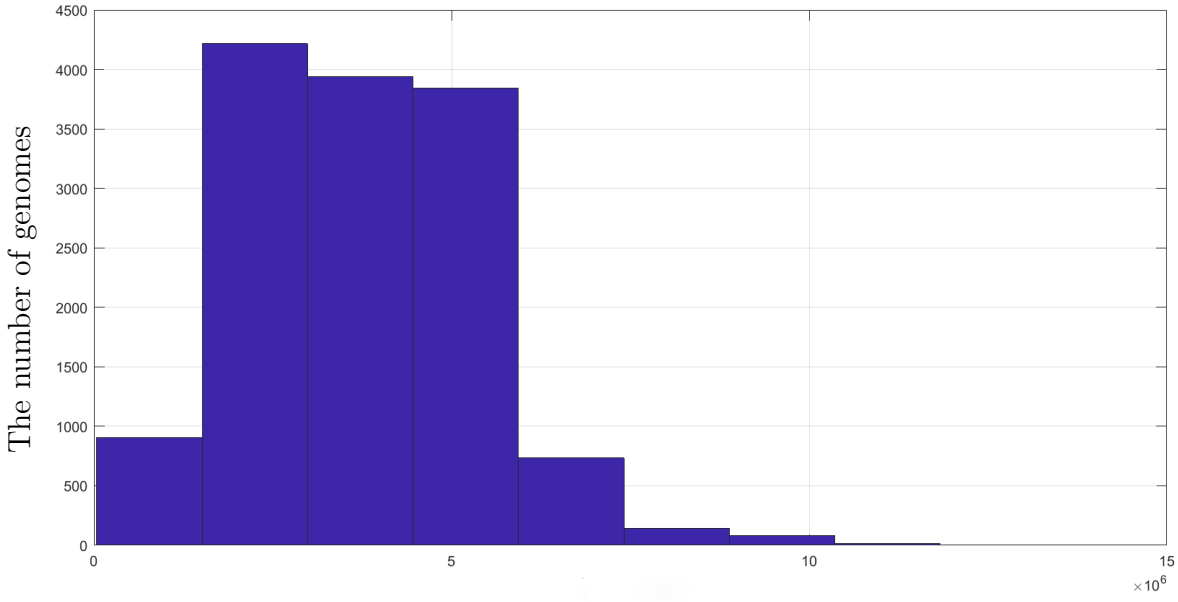


Figure 6.2: The length of the genomes in the data set

In the data set, the length of the genomes and GIs vary. Therefore, the other step of the GI analysis is to discover the lengths of the genomes in the data set as well as the lengths of the GIs and see how they are related. As shown in Figure 6.2, most of the genomes in the data set have a length range from $1,510,000$ to $5,940,000$, as shown in the tallest three bars in the middle of the histogram chart. On the other hand, it has been observed that the length of most GIs after normalisation is less than 0.1 as shown in figure 6.3. This means that most GIs make up approximately 10% or less of the length of the genome.

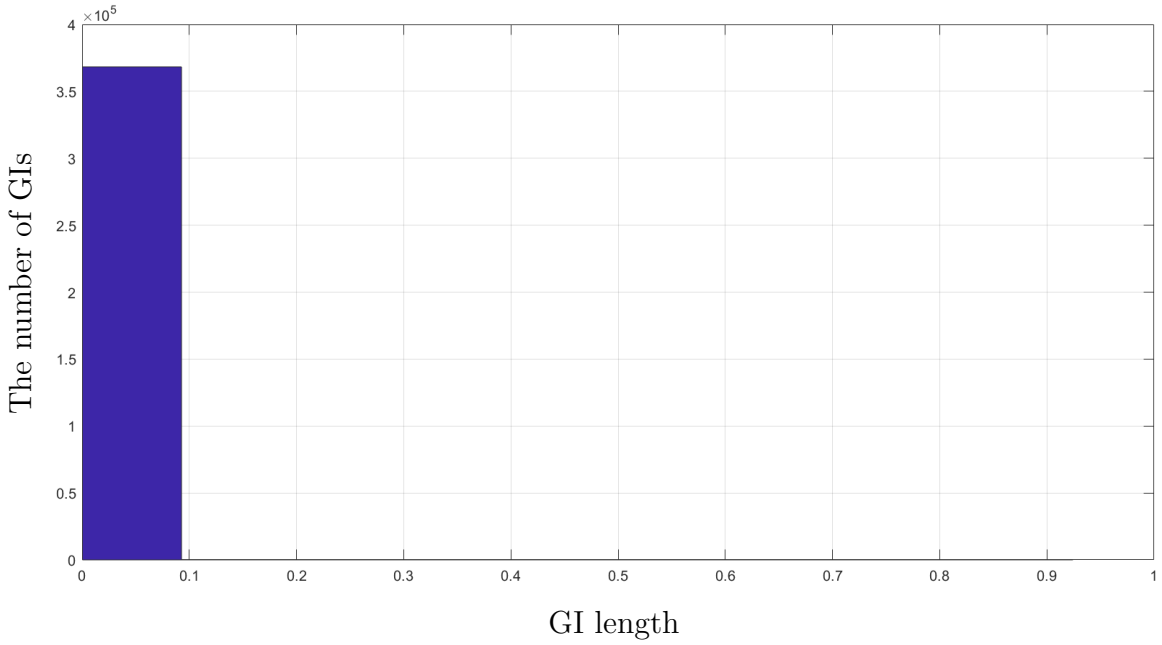


Figure 6.3: The length of the GIs in the data set

6.2.2 The Forming of Genomic Islands

The data set of GIs used in this research from the IslandViewer4 database is predicted by different tools. Therefore, some of the GIs overlap, whereas others do not. The overlapped GIs are GIs predicted by different tools and share roughly the same location within the genome. However, the start and the end coordinates from the overlapped GIs can be different. In this section, the overlapped GIs have been merged to form a new GI that covers all of the area where the overlapping occurs.

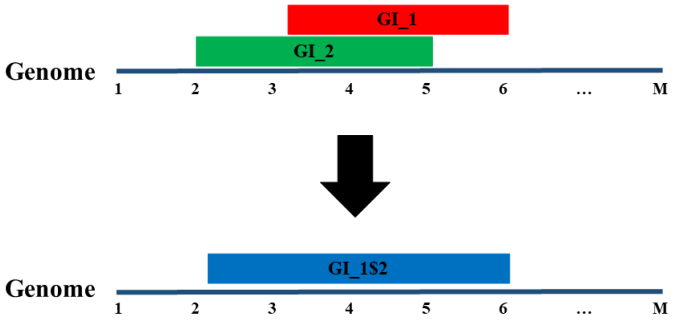


Figure 6.4: Genomic island overlap

The merge was performed to avoid the redundant GIs that are predicted from different GI prediction tools. Moreover, since this part of the genome has been determined as a GI by several GI prediction tools, it is more practical to deal with one GI, as all of these GIs share roughly the same area. However, the start and the end can be different for these overlapped GIs since different ones are predicted by different GI prediction tools. Therefore, the overlapped and non-overlapped areas of these GIs form a big GI, starting with the smallest coordinate out of all of the starting points, from the GIs in question and ending with the largest coordinate out of all of the ending points from the GIs as shown in Figure 6.4. Moreover, Algorithm 8 shows the strategy followed to form these GIs. The total number of the GIs before the overlap equates to 368,339, whereas after the overlap it equates to 257,050.

6.3 Methods

This section presents the analysis that has been performed on the GIs in order to discover the nature of GIs' locations within a genome. The first step of the analysis is to study the GIs' location in relation to the origin of replication (oriC) [37, 41]. Later, an investigation is presented about the nature of the distances between the GIs. Finally, the location distribution of GIs in the genome is determined.

The input of the analysis is the GIs and these GIs have been divided into circular group and linear group according to the genome structure that they belong to. Therefore, the aforementioned three analysis has been performed to each group of GIs separately. Furthermore, each of the three analyses was performed using three cases according to the data input utilized: first, all of the GIs in the data set; second, the GIs in one genome from each species; and third, the GIs of the most frequent species in the data set. The latter two cases were performed to avoid bias, since, within the data set, there are more GIs for a few specific

species than for others.

6.3.1 Genomic Islands location in Relation to the Origin of Replication

The prokaryotic GIs are in numerous locations within the genome and their preferred location is still unknown. One possible way of discovering the features of these GIs locations is by knowing their location in relation to the start of the genome, which is the location of the origin of replication. Origin of replication is a specific sequence found within a genome where the start of DNA replication takes place. DNA replication is the process by which the DNA is copied. In this research, the starting point of the GIs, which refers to the location of the GI in relation to the location of the oriC, has been plotted as histograms using Algorithm 9 to generate the numbers and Matlab to draw the figures.

6.3.2 The Nature of the Distances between the Genomic Islands

In this section, one of the objectives of the research is to understand the nature of the distance between the GIs. This is performed by normalizing the GIs coordinates then computing the distance between the GIs. The aim here is to plot the distances to discover if the GIs are located systematically or randomly, as mentioned previously in the problem definition section. Algorithm 10 has been used to compute the distances between the consecutive GIs in each genome. The distance between GIs is the length of space between every two GIs, as shown in Figure 6.5. This strategy in computing the distances is applied to the GIs that belong to circular and linear genomes. In circular genomes, the distance between the last GI in the genome and the first GI in the genome is also computed. The resulting distances were plotted using Matlab.

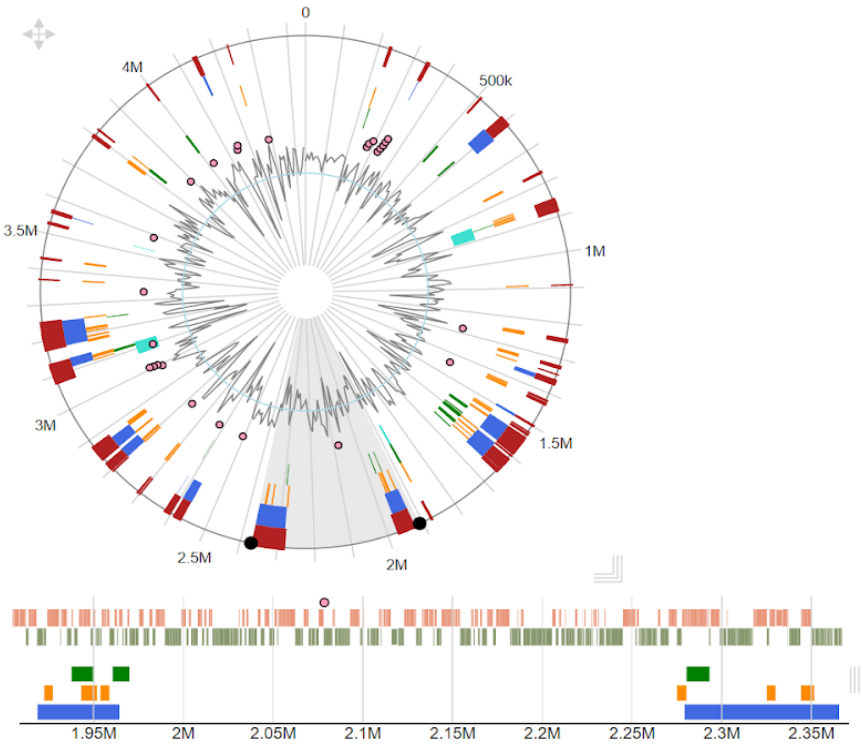


Figure 6.5: The circle shows the complete genome of the Acidovorax sp. (strain JS42) and the resulting GIs that come from IslandPick (Green), SIGI-HMM (Orange), and IslandPath-DIMOB (blue) [12]. The horizontal plot shows the part of the genome shaded in gray between the two black dots in the lower part of the circle. The blue triangle represents the GIs predicted by the IslandPath-DIMOB method, and the rest of the GIs are represented by the color of the predicted method. The distance between GIs is the length of space between every two GIs.

6.3.3 Location Distribution of Genomic Islands in the Genome

The prokaryotic GIs are distributed in the genome over different locations. The results of Section 6.3.1 and Section 6.3.2 proved that there were specific locations in which GIs are abundant; the middle of the genome (i.e., termiuns) and the oriC. Therefore, in this section, the locations of the GIs is studied from a different perspective by dividing the genome into n sections or slices and studying the GIs density in each section. Algorithm 11 is one of the algorithms that has been used in this step of the research to perform the analysis. As shown in Algorithm 11, the algorithm divides the genome into n sections then computes the number of the GIs in each section. A GI is considered located in a section if part of this

GIs is in this section.

6.4 Results

This section presents the investigation results of the GIs' locations using the three different analysis techniques.

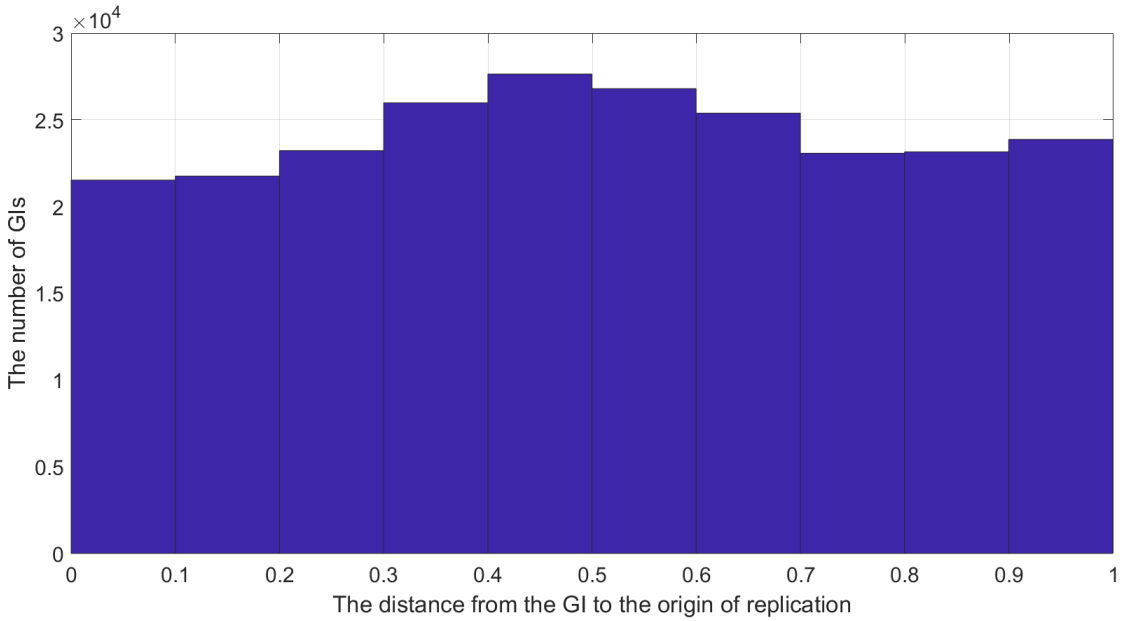
6.4.1 Genomic Islands location in Relation to the Origin of Replication

The location of the GI in relation to the location of the oriC, has been plotted as a histogram, as shown in Figure 6.6, where Part (a) of the figure represents all of the GIs in the data set that were located in circular genomes, and Part (b) represents all of the GIs in the data set that were located in linear genomes.

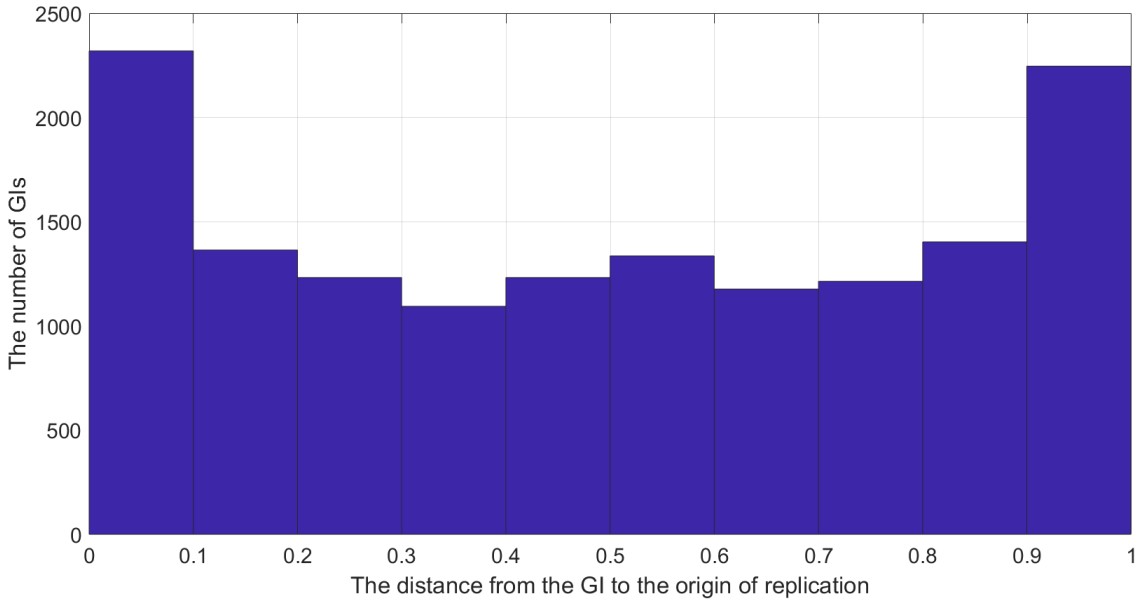
Figure (a) shows that there is a slight elevation in the termiuns. In Part (b), GIs in liner genomes, there is more presence of the GIs at start and the end points (i.e, in the origin and terminus). The high presence of GIs in the oriC area could be due to the presence of the binding proteins that assist in the replication process in the oriC area. Furthermore, another factor that facilitates the replication process is the presence of many A-T base pairs, which have weaker hydrogen bonding than G-C base pairs. The presence of weak hydrogen bonds could assist in the attachment of the GIs.

Figure 6.7 represents the analysis of the GIs form each species. The result of this analysis showed the same observations as in the "all the GIs" case in Figure 6.6.

As mentioned in Section 5.2.1, the top species in the data set which make up the largest proportion of the data set are *Escherichia coli*, *Salmonella enterica*, *Klebsiella pneumoniae*, *Bordetella pertussis*, and *Pseudomonas aeruginosa*. A histogram is built for each of the top

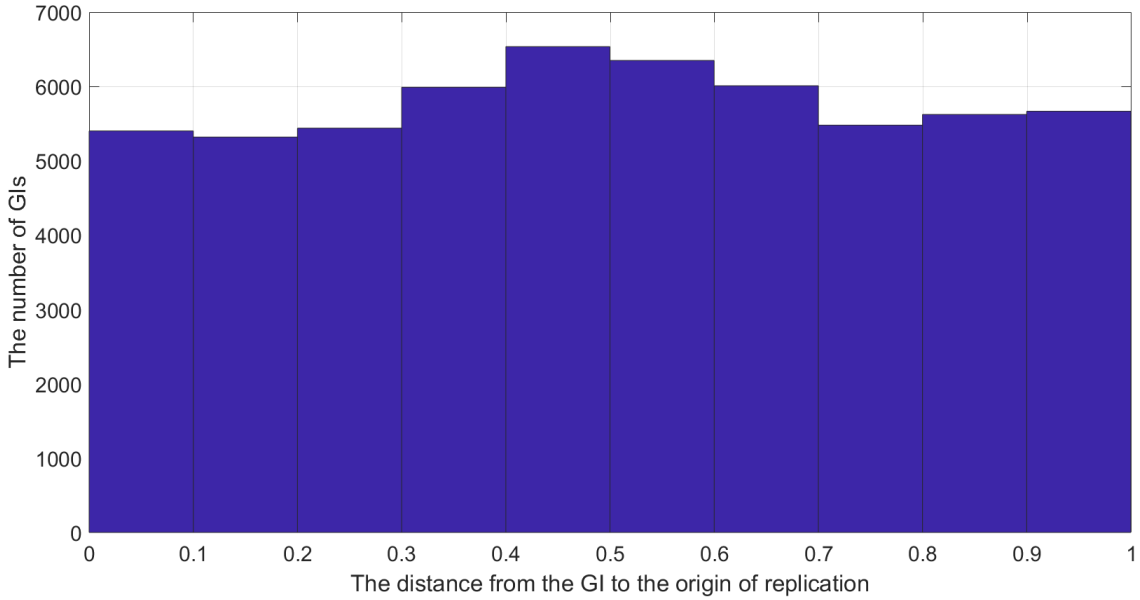


(a) GIs in circular genomes

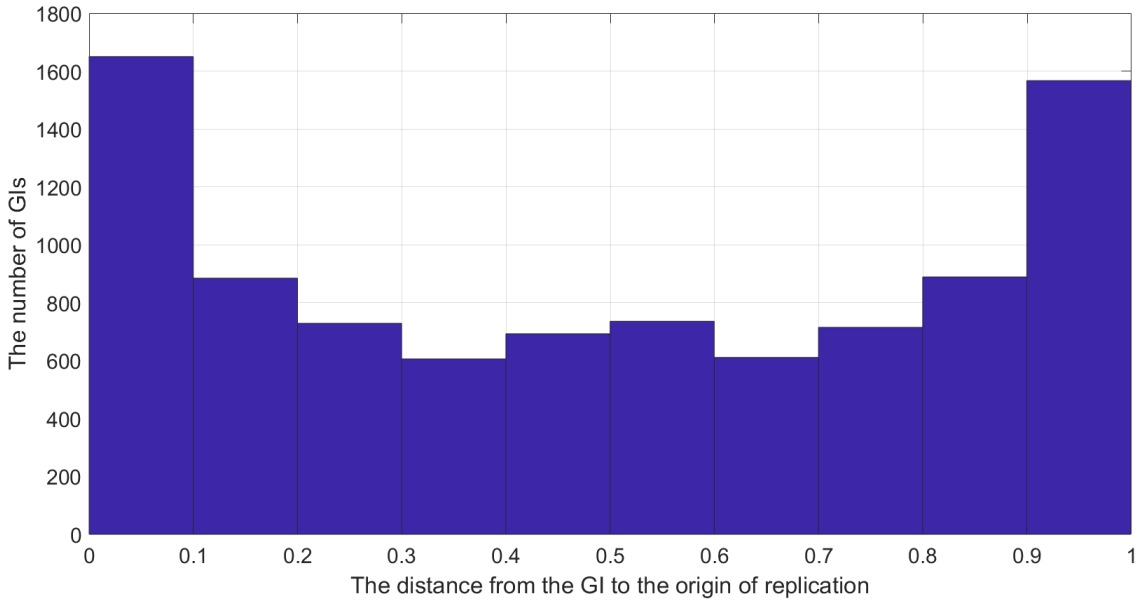


(b) GIs in linear genomes

Figure 6.6: The location of the GIs in relation to the oriC using all of the GIs in the data set

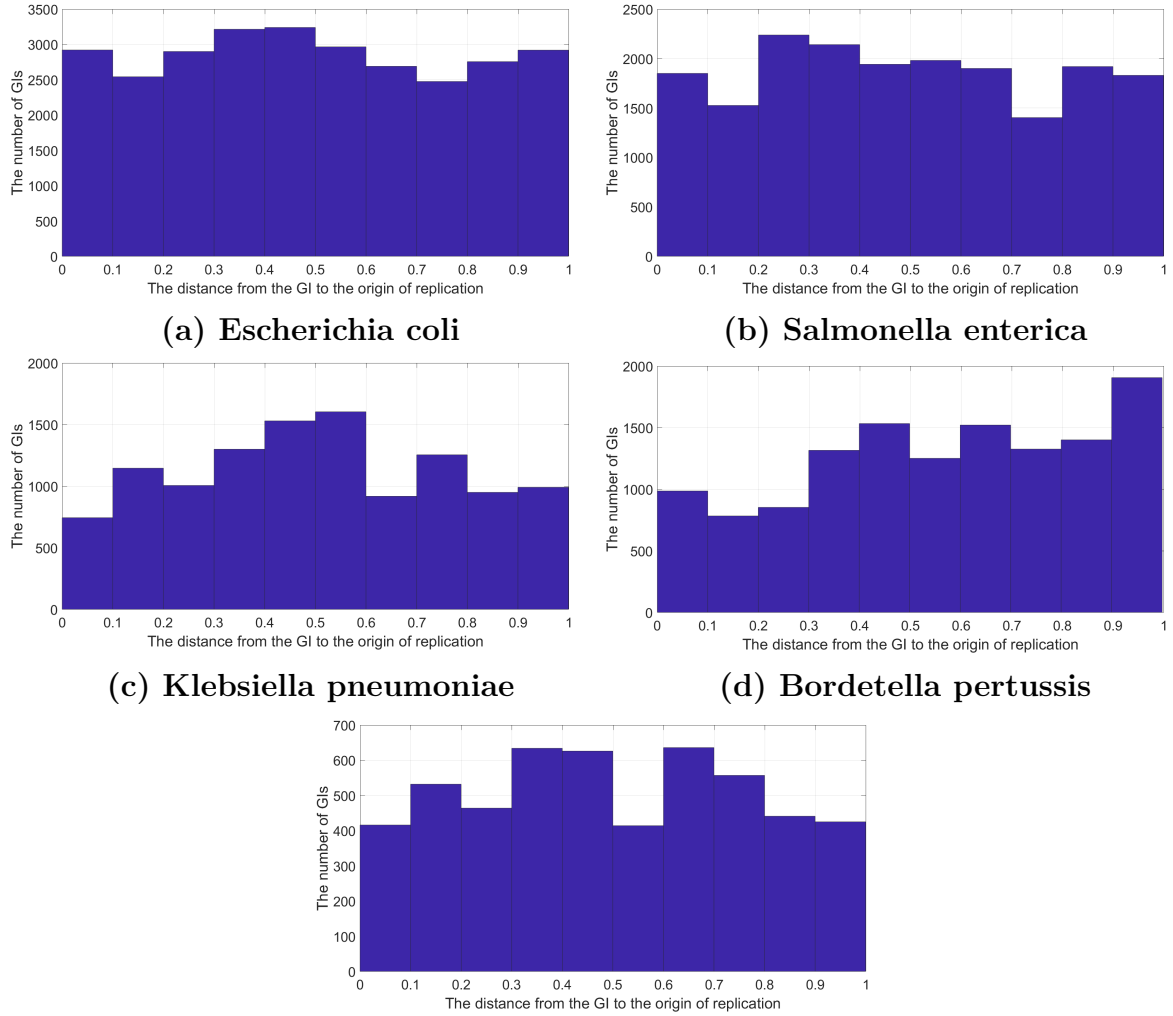


(a) GIs in circular genomes



(b) GIs in linear genomes

Figure 6.7: The location of the GIs in relation to the oriC using a genome from each species in the data set



(a) Escherichia coli

(b) Salmonella enterica

(c) Klebsiella pneumoniae

(d) Bordetella pertussis

(d) Pseudomonas aeruginosa

Figure 6.8: The location of the GIs of the most frequent species in the data set in relation to the origin of replication in circular genomes

species in the data set that reflect the true location of the GIs in relation to the oriC. This is carried out with no bias because the histogram represents only the GIs in a specific species. Figure 6.8 shows the histograms of the top five species in the data set. This analysis is performed for the GIs in circular genomes since, in the data set, very few GIs are located in linear genomes that belong to the most frequent species. Starting with the *Escherichia coli*, there is more presence of the GIs in the termiuns and at the end points (i.e., oriC). In *Salmonella enterica* there is a high presence of GIs in the termiuns and oriC. In both species, there are few GIs in the genome between 0.1 and 0.2 and between 0.7 and 0.8. The *Klebsiella pneumoniae* histogram shows a big presence of the GIs in the termiuns, unlike the other top species that show that their difference in location is not big. In the *Bordetella pertussis* histogram, GIs's presence generally increase at the end of the chart. There is less of a presence of GIs between 0.1 and 0.2. Finally, *Pseudomonas aeruginosa* shows a high presence in the termiuns, where there is a slight drop that is located in the genome between 0.5 and 0.6.

Table 6.1: Uniformity test for the GIs location in relation to the oriC

GI Category	**P-value**
All the GIs	1.405e-145
Genome from each species	1.021e-24
Escherichia coli	1.993e-15
Salmonella enterica	1.939e-20
Klebsiella pneumoniae	3.002e-26
Bordetella pertussis	1.749e-137
Pseudomonas aeruginosa	1.406e-09

The Kolmogorov-Smirnov uniformity test has been used to investigate the distribution of the GIs in relation to the oriC in all of the figures mentioned above, whether it is a uniform distribution or not. In statistics, a uniform distribution means every possible result (i.e. distance) has an equal chance of occurring. Table 6.1, shows that all the figures do not follow the uniform distribution since all the p-values are less than 0.05; therefore, this leads

to rejecting the null hypothesis. The null hypothesis is rejected if the p-value is at the 5% significance level or less.

Overall, there are different places, as shown in the diagrams mentioned above, containing GIs, but generally they are concentrated in the terminus or the oriC area. In the literature, there are some observations about how HGT can occur more often around the origin, and terminus [80, 86]. Regarding the termiuns, the density of GIs in circular genomes is more in the termiuns than in other areas. Consequently, an investigation has been performed in order to investigate further about the density of the GIs in the termiuns and the oriC area, since the oriC and terminus locations in linear genomes are the oriC location in circular genomes.

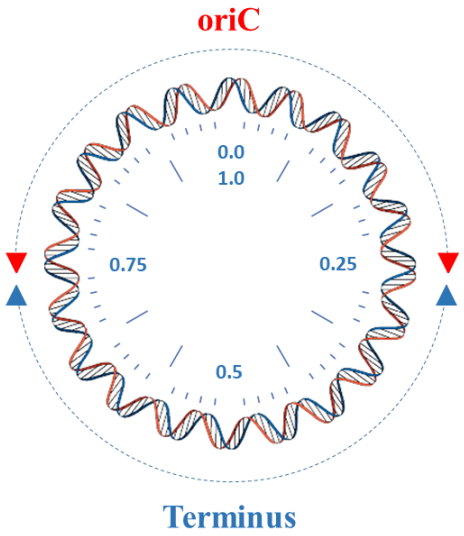


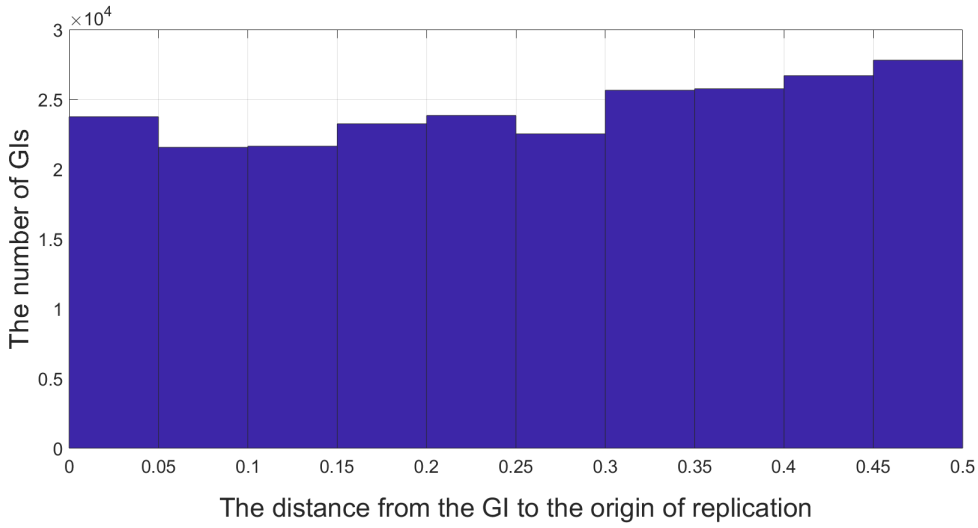
Figure 6.9: A circular genome divided into two equal size arcs; oriC arc (from 0.0 to 0.25 and from 0.75 to 1.0) and termiuns arc (from 0.25 to 0.75).

In this analysis step, the circular genomes have been divided into two equal sections (i.e., arcs), as shown in Figure 6.9. GIs in the area from 0.0 to 0.25 or 0.75 to 1.0 are considered

Table 6.2: The number of GIs in circular genomes, in the oriC area and the termiuns, for each data input utilized.

GI Category	oriC	Termiuns
All the GIs	114,025	128,408
Genome from each species	27,482	30,335
Escherichia_coli	14108	14516
Salmonella_enterica	9330	9400
Klebsiella_pneumoniae	4968	6476
Bordetella_pertussis	6142	6734
Pseudomonas_aeruginosa	2239	2906

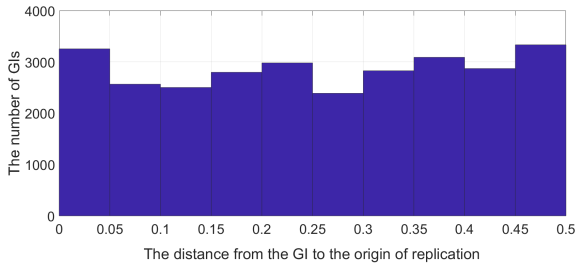
located in the oriC area. Furthermore, GIs located between 0.25 and 0.75 are considered to belong to the termiuns area. The number of the GIs has been counted in both areas, as shown in Table 6.2. In all the GI types, the number of GIs in the termiuns is more significant than in the oriC area. This also can be seen in Figure 6.11 and Figure 6.10. More precisely, in circular genomes, a GI that is 0.1 away from the oriC (oriC locus = 0.0) is the same as a GI 0.9 from the oriC. Therefore, at this step, the x-axis scale in Figure 6.6, 6.7, and Figure 6.8 is changed from (0 to 1) to (0 to 0.5). This is performed by subtracting any GI located in a locus greater than 0.5 by 1, and keeping the other GIs the same if they are located between 0 and 0.5. For this reason a GI located between 0 and 0.25 are in the oriC area, and GIs between 0.25 and 0.5 are in the termiuns area.



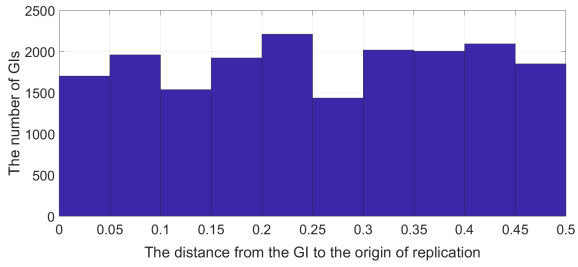
(a) All of the GIs in the data set

(b) A genome from a species

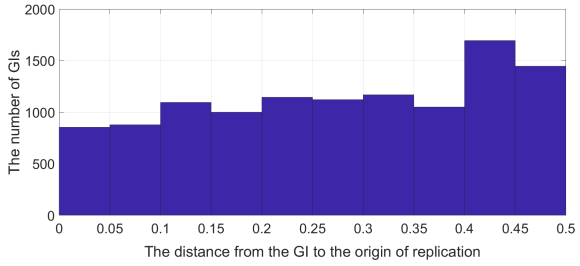
Figure 6.10: The location of the GIs in relation to the origin of replication in circular genomes with a range of 0 to 0.5



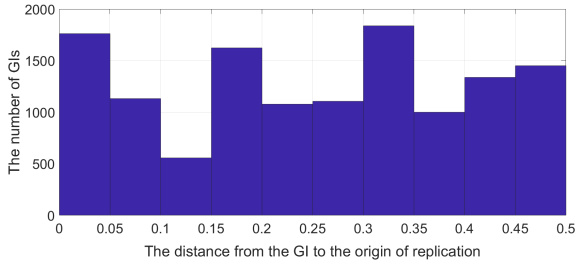
(c) Escherichia coli



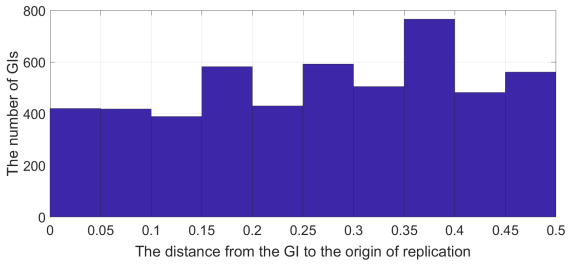
(d) Salmonella enterica



(c) Klebsiella pneumoniae



(d) Bordetella pertussis

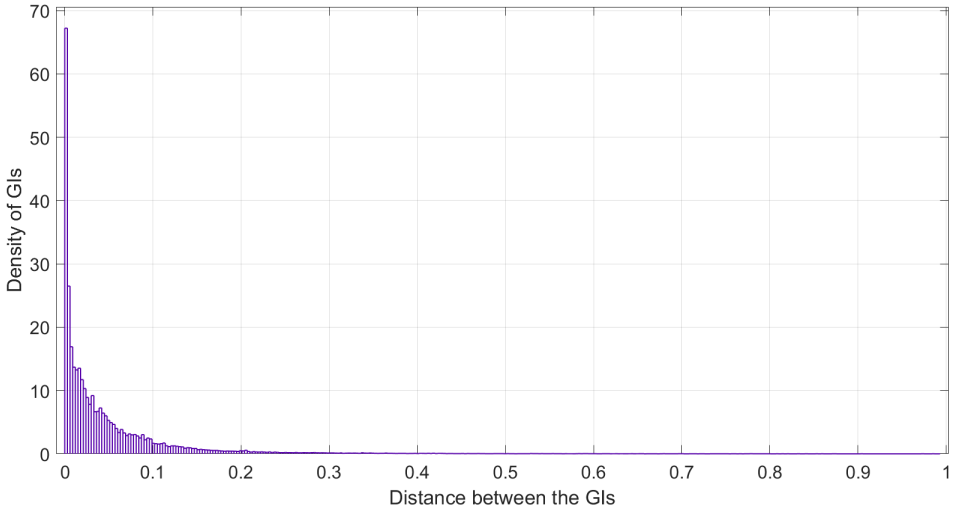


(d) Pseudomonas aeruginosa

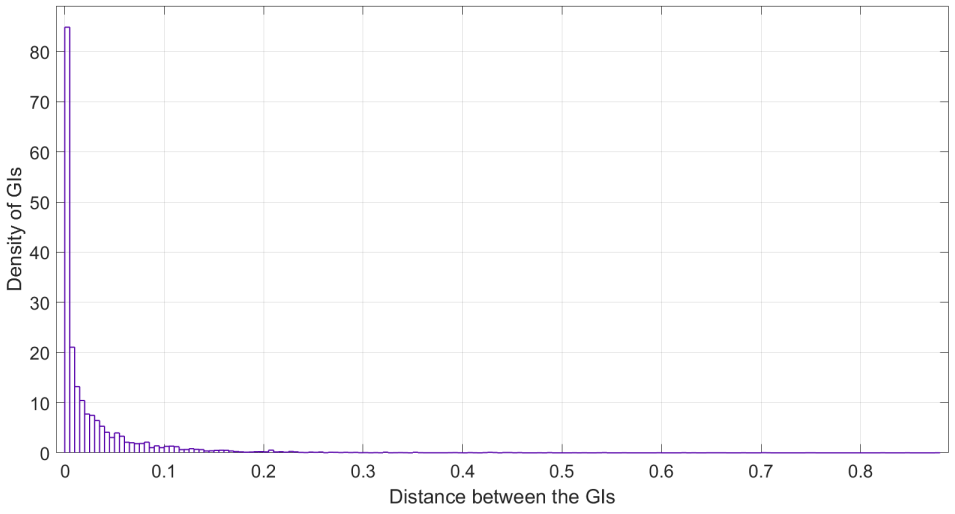
Figure 6.11: The location of the GIs in relation to the origin of replication in circular genomes ranges from 0 to 0.5

6.4.2 The Nature of the Distances between the Genomic Islands

In this section, the resulting distances have been plotted using Matlab. The distribution fitter application from the statistical and machine learning toolbox in Matlab has been used to depict the distances. Moreover, the parameter Log likelihood has been used as a measure of selecting the distribution, where the higher value, the better.

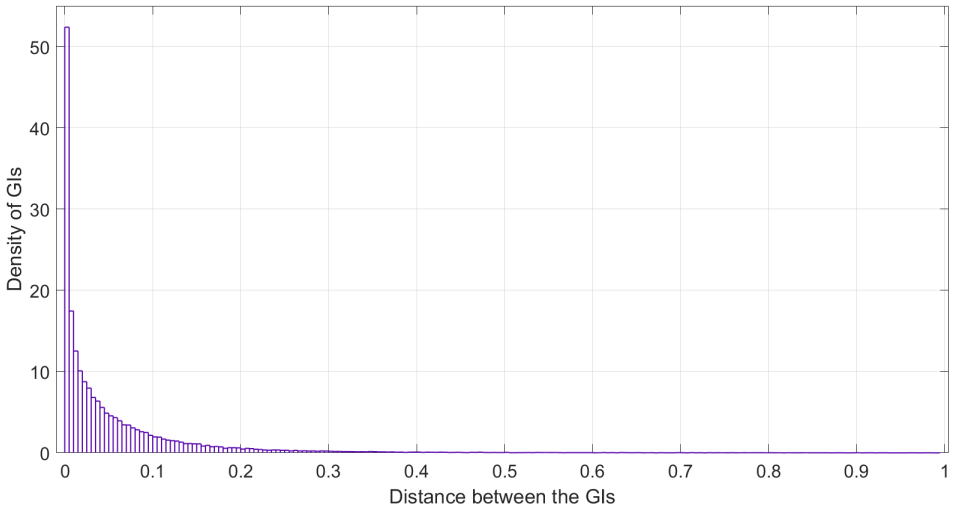


(a) GIs in circular genomes

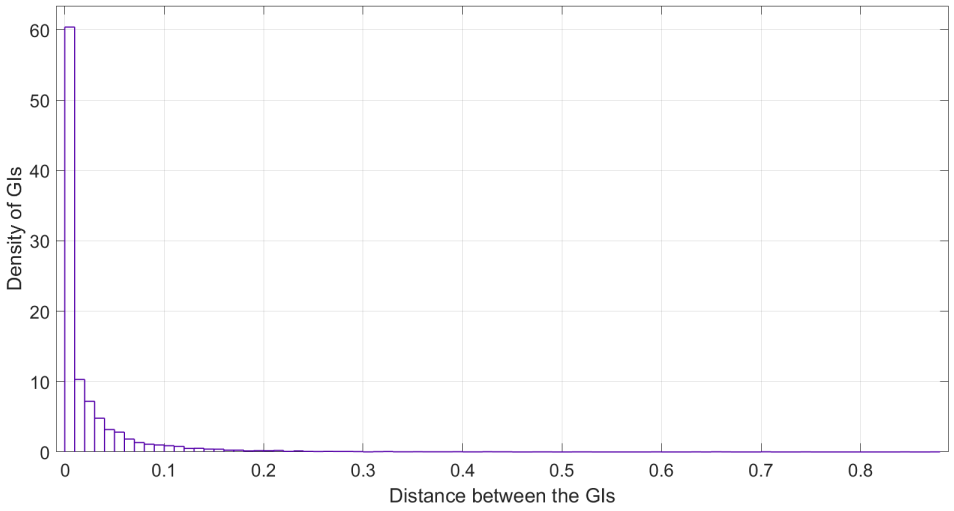


(b) GIs in linear genomes

Figure 6.12: The distance between the GIs using all of the GIs in the data set

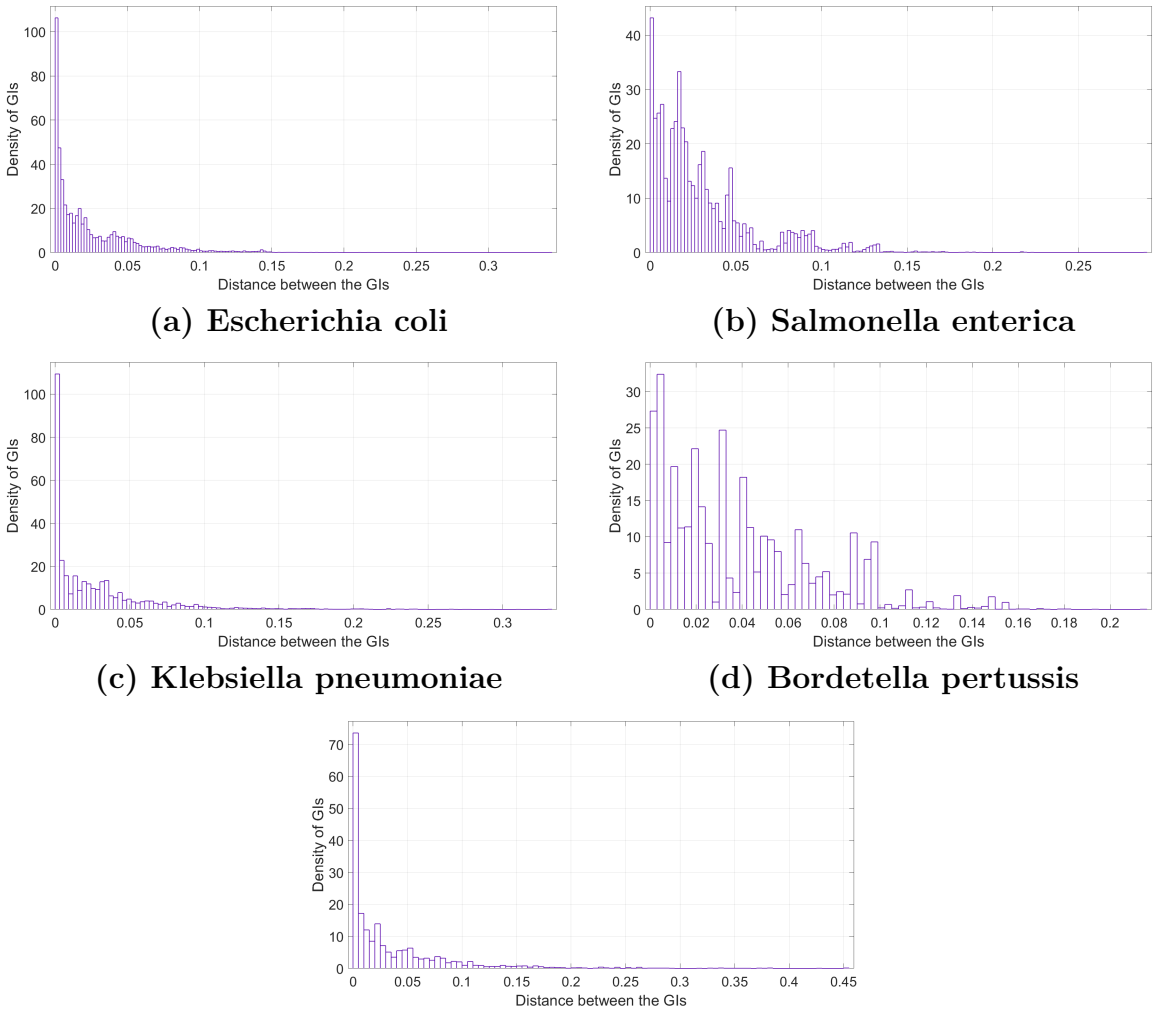


(a) GIs in circular genomes



(b) GIs in linear genomes

Figure 6.13: The distance between the GIs using the GIs from each species in the data set



(a) Escherichia coli

(b) Salmonella enterica

(c) Klebsiella pneumoniae

(d) Bordetella pertussis

(c) Pseudomonas aeruginosa

Figure 6.14: The distance between the GIs in the most frequent species in the data set

Overall, the resulting distance charts look almost like an exponential probability distribution for most of the charts, as shown in Figure 6.12, Figure 6.13, and Figure 6.14. From these charts, it is obvious that most of the GIs are usually close to each other and this supports the previous results in Section 6.3.1 where there is high presence of the GIs in relation to the oriC in specific locations in the genome. In Figure 6.14, it can be seen that in *Salmonella enterica* and *Bordetella pertussis* the distribution is not overly exponential.

6.4.3 Location Distribution of Genomic Islands in the Genome

In the location distribution of GIs in the genome analysis, the analysis is performed to the GIs for n=10. This means each genome is divided into ten equal parts then the number of the GIs in each section is computed. This analysis aims to know the location of the GIs more precisely in the termiuns and oriC area.

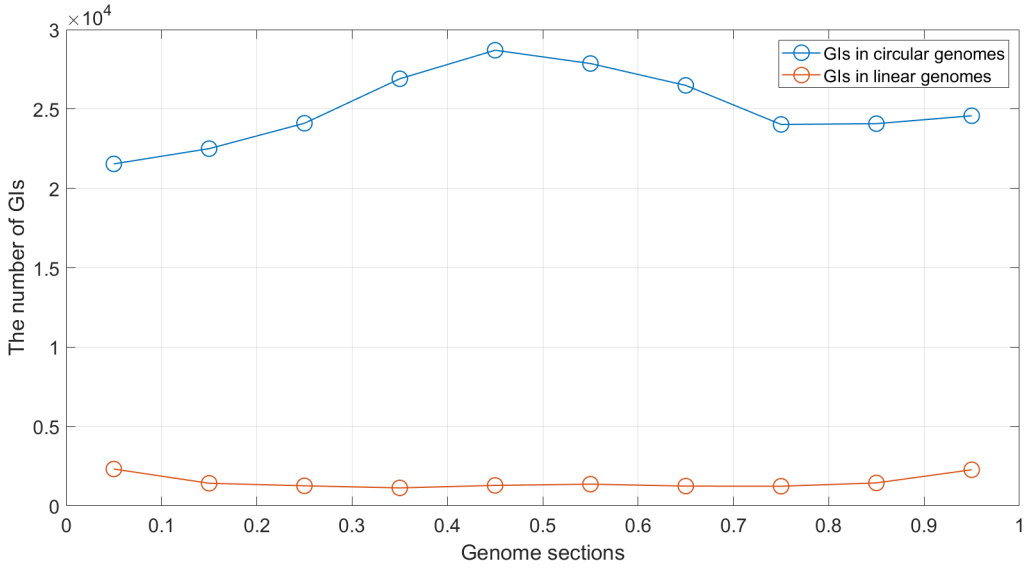


Figure 6.15: GIs' distribution using all of the GIs in the data set

Figure 6.15 shows the distribution using all of the GIs in the data set. In the figure, there are two lines; the blue line represents the GIs that are located in the circular genomes, whereas the orange line represents the GIs that are located in the linear genomes. Regarding all the

GIs in the data set, it is clear that there is a notable elevation in the termiuns. In regards to the GIs in linear genomes, overall, there is no significant difference but there is a slight elevation in the oriC and terminus.

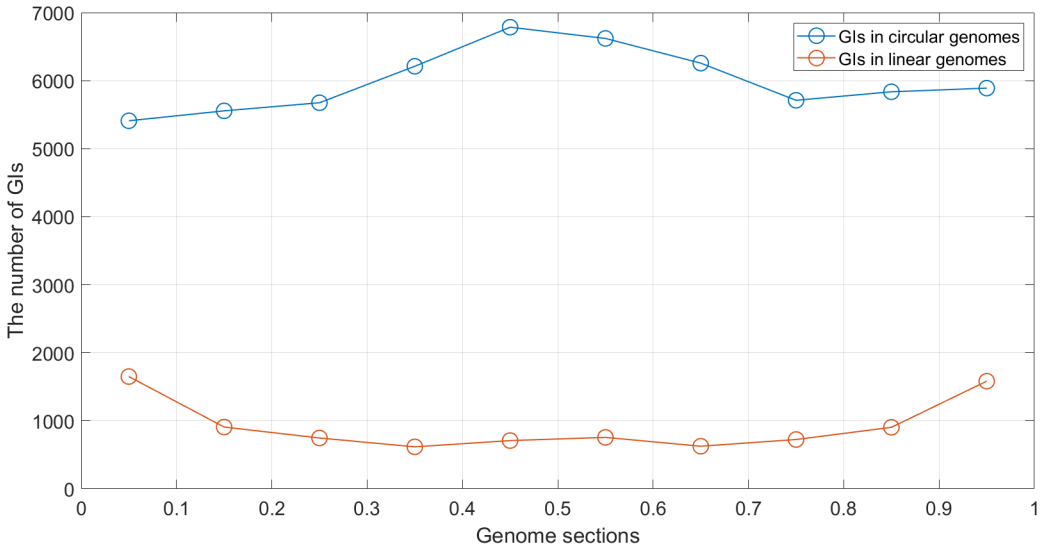


Figure 6.16: GIs' distribution using the GIs from each species in the data set

In Figure 6.16, the analysis of the GIs from each species shows the same results as in the "all the GIs" case. The only difference is that in the GIs in the linear genomes, the elevation in the oriC and terminus is more evident than in the "all the GIs" case.

The analysis has also been performed on the most frequent species in the data set; *Escherichia coli*, *Salmonella enterica*, *Klebsiella pneumoniae*, *Bordetella pertussis*, and *Pseudomonas aeruginosa*. Figure 6.17 shows the most frequent species GIs distribution for circular genomes.

In *Escherichia coli*, the elevation of the GIs is in the terminus and the oriC area. For the *Salmonella enterica*, the density of the GIs in the genome is in locus between 0.2 and 0.3. Regarding the *Klebsiella pneumoniae*, the GIs more located in the terminus in the locus between 0.5 and 0.6. The *Bordetella pertussis* GIs are more located at the end between 0.9 and 1. Finally, the *Pseudomonas aeruginosa*, the distribution of the GIs is almost uniform. It's worth mentioning that there is less occurrence of the GIs in two location; between 0.1

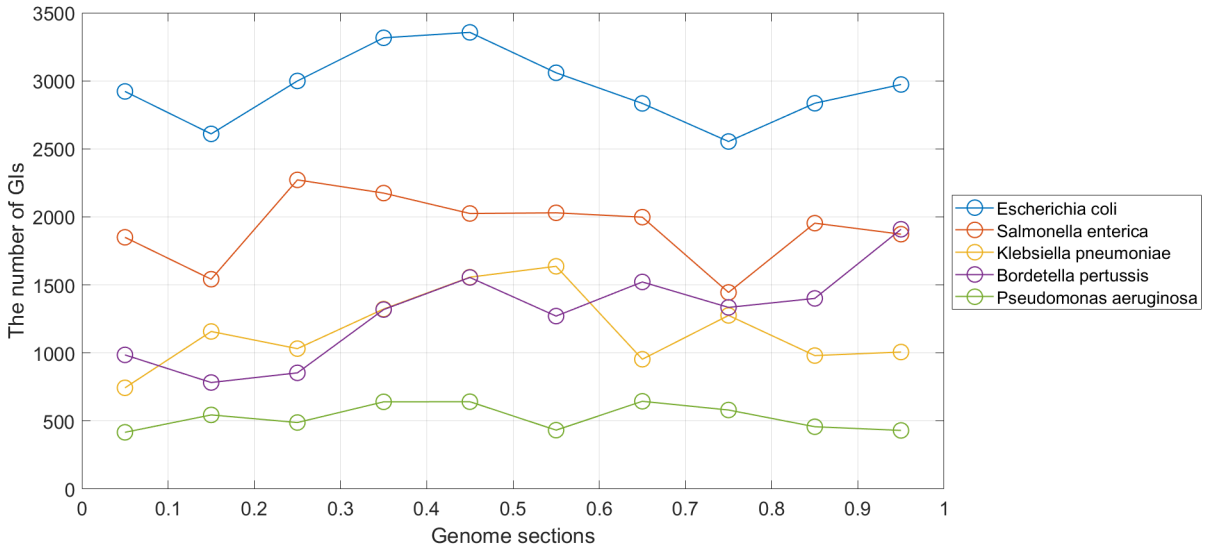


Figure 6.17: GIs' distribution using the most frequent species in the data set

and 0.2, and between 0.7 and 0.8, and this can be seen in the GIs of *Escherichia coli* and *Salmonella enterica* species.

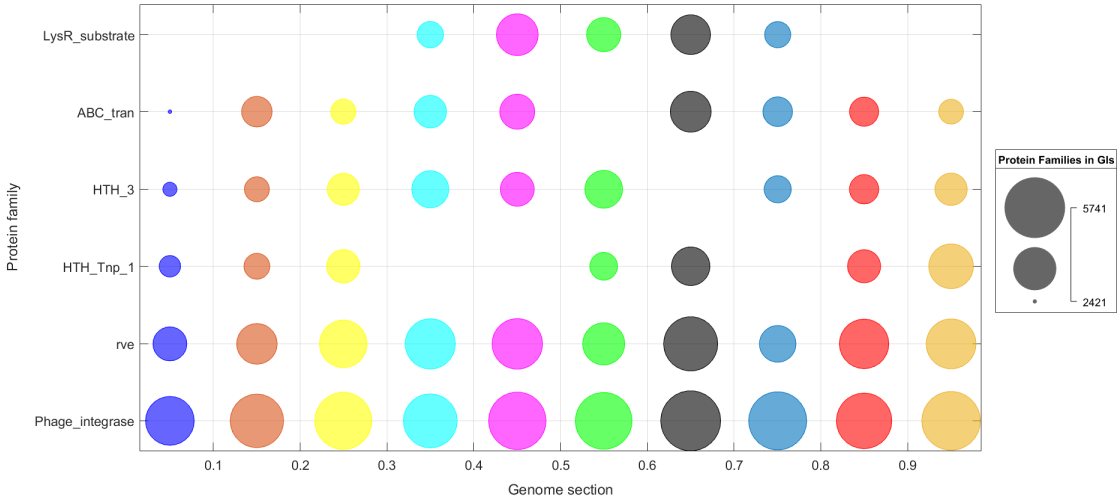


Figure 6.18: The most frequent protein families in each section in the genomes

In general, most prokaryotic GIs share the same fact that they are almost abundant in some areas. Therefore, it is significant to discover the content of the GIs in these areas. The GIs content is the protein families. Figure 6.18 shows that the Phage_integrase and rve protein families always exist in each section, and Phage_integrase is the most abundant protein

family. After that, in no specific order, comes the other protein families; HTH_Tnp_1, HTH_3, ABC_tran, and LysR_substrate. The aforementioned protein families are the most frequent protein families in the GIs in each section in the genome.

6.5 Discussion

In this chapter, the prokaryotic GIs' structure has been studied in terms of their location in the genome. The GIs used in this research from the IslandViewer4 have been preprocessed to overcome the overlapped GIs. The overlapped GIs have been merged to form new GIs covering all areas where the overlapping occurs. The input of the analysis is the resulting GIs, which have been divided into circular and linear groups according to the genome structure they belong to. Each group is analyzed according to the data input utilized: first, all of the GIs in the data set; second, the GIs in one genome from each species; and third, the GIs of the most frequent species in the data set. The latter two cases were performed to avoid bias. The analysis is performed to study the GIs' location in relation to the oriC, investigating the nature of the distances between the GIs, and determining the distribution of GIs in the genome.

Overall, the locations of GIs in relation to the oriC analysis and the distribution of the GIs in the genome analysis demonstrated that the middle of the genome (i.e., termiuns) is a preferred site for the GIs to reside in the circular. Moreover, the oriC and terminus are the preferable locations for the GIs in linear genomes. This observation is supported by the results of the analysis of the distance between the GIs. The analysis results showed that the distribution of the distances between the GIs almost follows an exponential distribution. This could lead to there being preferable sites for the GIs in the genomes. The oriC is a preferred site for the GIs and the natural explanation for this is that it could be related to the content of the oriC area that facilitates the attachment process of the GI, since it

contains binding proteins and also contains more AT base pairs that have weak hydrogen bonds and that are easy to break [107]. The termiuns requires deep investigations in order to discover the characteristics of this part of the genome in terms of proteins that make it a preferable site for GIs since, in the literature, there are not sufficient studies regarding the content of the termiuns in prokaryotic species.

The content of the GIs in terms of protein families has been studied to discover the most frequent protein families in each section of the genome when dividing the genome into ten sections. The results showed that the most frequent protein families are almost identical in each section.

Chapter 7

Conclusions and Future Work

This dissertation discussed the structure of the prokaryotic GIs in numerous perspectives that have not been addressed intensively in the literature. These directions could have a great impact in the field of prokaryotic evolution. Islandviewer4 is the database used in this research that is very diverse since it contains a vast number of GIs collected from a different number of GI prediction tools. The two main directions of this research are to study the GIs in terms of the content of the GIs and also the location of the GIs.

In this first direction, the in-depth analysis aims to extract patterns of protein families from the GIs, which are then used to obtain biological connections between prokaryotes and phages. The HGT in bacterial communities can occur in closely and distantly related species, and this has been seen in the patterns that exist in distantly related species. There could be a relation between the HGT and the shape of the prokaryotic where the GI is located since the analysis showed that most GIs exist in rod shape species. The GIs have specific structures and that has been shown in the patterns that have a crucial role in changing the biological function of the host species. This structure is mainly composed of three protein families, that could indicate that the main prokaryotic GI patterns are composed of three protein families that are necessary for the success of the HGT process. As a future work, in the first research direction, these GI could be studied further to study the possibility that they originate from phages since it has been observed that most of the resulting patterns contain phage protein families.

The analysis in Section 5.3.4, showed that the common pattern between the phage and

the group of bacteria assists in discovering a GI that is common between them. This GI is the same as the GI that is provided in IslandViewer4 for each bacteria in the connection. Therefore, species in the same connection have a very high probability of HGT between them, even if they are distantly related. Therefore, an investigation could be performed in order to discover the origin and direction of the GIs using the patterns and connections.

In the second direction, in general, most prokaryotic GIs share the same characteristic, that they are abundant in specific locations in the genome. Analyzing the GIs' location leads to the conclusion that the prokaryotic GIs are frequently located in the termiuns in circular genomes. In the linear genomes, the GIs can usually be seen in the oriC area and terminus. This could indicate that the prokaryotic GIs are possibly located biologically in a systematic way as opposed to randomly. In the literature, there are some observations that the HGT occurs in the oriC and terminus [80, 86]. Regarding the termiuns, it has not been intensively studied, therefore, studying its sequence could assist in discovering the possibility of being the preferred location of GIs. Another future direction could be studying the distribution of the GIs' protein families in the genome that could assist in understanding more about the localization strategy of the GIs in the genome. One of the objectives of this analysis is to discover if there are preferred locations for specific GIs' genes within the genome.

In general, analyzing the GIs using techniques that have not been addressed in the literature play a significant role in understanding the HGT process, prokaryotic evolution, disease, drug resistance, and antibiotic resistance.

Bibliography

[1] H.-W. Ackermann, *Phage classification and characterization*, Methods in Molecular Biology, 501 (2009), p. 14 pages.

[2] R. Agrawal, R. Srikant, et al., *Fast algorithms for mining association rules*, in Proceedings of 20th International Conference on Very Large Data Bases, VLDB, vol. 1215, Citeseer, Sep 1994, pp. 487–499.

[3] S. E. Ahnert, T. M. Fink, and A. Zinovyev, *How much non-coding DNA do eukaryotes require?*, Journal of Theoretical Biology, 252 (2008), pp. 587–592.

[4] S. K. Ainala, E. Seol, and S. Park, *Complete genome sequence of novel carbon monoxide oxidizing bacteria* Citrobacter amalonaticus *Y19, assembled* de novo, Journal of Biotechnology, 211 (2015), pp. 79–80.

[5] M. A. Aldahmesh, Z. N. Al-Hassnan, M. Aldosari, and F. S. Alkuraya, *Neuronal ceroid lipofuscinosis caused by MFSD8 mutations: a common theme emerging*, Neurogenetics, 10 (2009), pp. 307–311.

[6] A. B. Alzwghaibi, R. Yahyaraeyat, B. N. Fasaei, A. G. Langeroudi, and T. Z. Salehi, *Rapid molecular identification and differentiation of common* Salmonella *serovars isolated from poultry, domestic animals and foodstuff using multiplex PCR assay*, Archives of Microbiology, 200 (2018), pp. 1009–1016.

[7] A. Andino and I. Hanning, Salmonella enterica*: Survival, Colonization, and Virulence Differences among Serovars*, Scientific World Journal, 2015 (2015), p. 520179.

[8] M. D. Andrake and A. M. Skalka, *Retroviral Integrase: Then and Now*, Annual Review of Virology, 2 (2015), p. 241.

[9] M. Arenskötter, D. Bröker, and A. Steinbüchel, *Biology of the metabolically diverse genus* Gordonia, Applied and Environmental Microbiology, 70 (2004), pp. 3195–3204.

[10] R. Assaf, F. Xia, and R. Stevens, *Identifying genomic islands with deep neural networks*, bioRxiv, (2019).

[11] J. Bazan, I. Całkosiński, and A. Gamian, *Phage display–a powerful technique for immunotherapy: 1. Introduction and potential of therapeutic applications*, Hum. Vaccin. Immunother., 8 (2012), pp. 1817–1828.

[12] C. Bertelli, M. R. Laird, K. P. Williams, B. Y. Lau, G. Hoad, G. L. Winsor, and F. S. B. and, *IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets*, Nucleic Acids Research, 45 (2017), pp. W30–W35.

[13] C. Bertelli, K. E. Tilley, and F. S. L. Brinkman, *Microbial genomic island discovery, visualization and analysis*, Briefings in Bioinformatics, 20 (2018), pp. 1685–1698.

[14] I. G. Boneca, *The role of peptidoglycan in pathogenesis*, Current Opinion in Microbiology, 8 (2005), pp. 46–53.

[15] D. Borenshtein, M. E. McBee, and D. B. Schauer, *Utility of the* Citrobacter rodentium *infection model in laboratory mice*, Current Opinion in Gastroenterology, 24 (2008), pp. 32–37.

[16] L. Boto, *Horizontal gene transfer in evolution: facts and challenges*, Proceedings of the Royal Society B: Biological Sciences, 277 (2009), pp. 819–827.

[17] P. Bourhy, L. Collet, S. Brisse, and M. Picardeau, Leptospira mayottensis *sp. nov., a pathogenic species of the genus* Leptospira *isolated from humans*, International Journal of Systematic and Evolutionary Microbiology, 64 (2014), p. 4061.

[18] K. Brew, P. Tumbale, and K. R. Acharya, *Family 6 Glycosyltransferases in Vertebrates and Bacteria: Inactivation and Horizontal Gene Transfer May Enhance Mutualism*

between Vertebrates and Bacteria, The Journal of Biological Chemistry, 285 (2010), p. 37121.

[19] S. D. Brown, M. B. Begemann, M. R. Mormile, J. D. Wall, C. S. Han, L. A. Goodwin, S. Pitluck, M. L. Land, L. J. Hauser, and D. A. Elias, *Complete genome sequence of the haloalkaliphilic, hydrogen-producing bacterium* Halanaerobium hydrogeniformans, Journal of Bacteriology, 193 (2011), pp. 3682–3683.

[20] M. Brown-Jaque, M. Muniesa, and F. Navarro, *Bacteriophages in clinical samples can interfere with microbiological diagnostic tools*, Scientific Reports, 6 (2016).

[21] H. Brüssow and R. W. Hendrix, *Phage genomics: small is beautiful*, Cell, 108 (2002), pp. 13–16.

[22] A. Buckling and P. B. Rainey, *Antagonistic coevolution between a bacterium and a bacteriophage*, Proceedings: Biological Sciences, 269 (2002), pp. 931–936.

[23] E. C. Bush, A. E. Clark, C. A. DeRanek, A. Eng, J. Forman, K. Heath, A. B. Lee, D. M. Stoebel, Z. Wang, M. Wilber, and H. Wu, *xenoGI: reconstructing the history of genomic island insertions in clades of closely related bacteria*, BMC Bioinformatics, 19 (2018), pp. 1–11.

[24] E. Cankaya, M. Keles, E. Gulcan, A. Uyanik, and H. Uyanik, *A Rare Cause of Peritoneal Dialysis-Related Peritonitis:* Achromobacter denitrificans, Peritoneal Dialysis International, 34 (2014), p. 135.

[25] S. Casjens, *Prophages and bacterial genomics: what have we learned so far?*, Mol. Microbiol., 49 (2003), pp. 277–300.

[26] J. Challacombe and C. Kuske, *Mobile genetic elements in the bacterial phylum Acidobacteria*, Mobile Genetic Elements, 2 (2012), p. 179.

[27] J. F. Challacombe, S. A. Eichorst, L. Hauser, M. Land, G. Xie, and C. R. Kuske, *Biological Consequences of Ancient Gene Acquisition and Duplication in the Large Genome of* Candidatus *Solibacter usitatus Ellin6076*, PLOS One, 6 (2011), p. e24882.

[28] F. Chang and K. C. Huang, *How and why cells grow as rods*, BMC Biology, 12 (2014), pp. 1–11.

[29] D. Charlier, J. Piette, and N. Glansdorff, *IS3 can function as a mobile promoter in* E. coli., Nucleic Acids Research, 10 (1982), p. 5935.

[30] K. Chawla and P. Y. Prakash, *Biology of Pathogenic Actinobacteria:* Nocardia *and Allied Genera*, in New and Future Developments in Microbial Biotechnology and Bioengineering, Elsevier, Walthm, MA, USA, Jan 2018, pp. 225–233.

[31] H. Y. Chu, K. Sprouffske, and A. Wagner, *Assessing the benefits of horizontal gene transfer by laboratory evolution and genome sequencing*, BMC Evolutionary Biology, 18 (2018).

[32] T. Coenye, M. Vancanneyt, M. C. Cnockaert, E. Falsen, J. Swings, and P. Vandamme, Kerstersia gyiorum *gen. nov., sp. nov., a novel* Alcaligenes faecalis*-like organism isolated from human clinical samples, and reclassification of* Alcaligenes denitrificans *Rüger and Tan 1983 as* Achromobacter denitrificans *comb. nov.*, International Journal of Systematic and Evolutionary Microbiology, 53 (2003), pp. 1825–1831.

[33] C. Cordonin, M. Turpin, M. Bringart, J.-L. Bascands, O. Flores, K. Dellagi, P. Mavingui, M. Roche, and P. Tortosa, *Pathogenic* Leptospira *and their animal reservoirs: testing host specificity through experimental infection*, Scientific Reports, 10 (2020), pp. 1–8.

[34] C. Darwin, *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life (london, 1859)*, Faraday to Agassiz, 20 (1863).

[35] A. L. Davidson, E. Dassa, C. Orelle, and J. Chen, *Structure, function, and evolution of bacterial ATP-binding cassette systems*, Microbiology and Molecular Biology Reviews, 72 (2008), pp. 317–364.

[36] A. Davin-Regli, C. Bosi, R. Charrel, E. Ageron, L. Papazian, P. A. Grimont, A. Cremieux, and C. Bollet, *A nosocomial outbreak due to Enterobacter cloacae strains with the E. hormaechei genotype in patients treated with fluoroquinolones*, Journal of Clinical Microbiology, 35 (1997), pp. 1008–1010.

[37] G. del Solar, R. Giraldo, M. J. Ruiz-Echevarría, M. Espinosa, and R. Díaz-Orejas, *Replication and Control of Circular Bacterial Plasmids*, Microbiology and Molecular Biology Reviews, 62 (1998), p. 434.

[38] U. Dobrindt, B. Hochhut, U. Hentschel, and J. Hacker, *Genomic islands in pathogenic and environmental microorganisms*, Nature Reviews Microbiology, 2 (2004), pp. 414–424.

[39] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, England, UK, Apr 1998.

[40] S. Ehrich, D. Behrens, E. Lebedeva, W. Ludwig, and E. Bock, *A new obligately chemolithoautotrophic, nitrite-oxidizing bacterium,*Nitrospira moscoviensis *sp. nov. and its phylogenetic relationship*, Archives of Microbiology, 164 (1995), pp. 16–23.

[41] B. Ekundayo and F. Bleichert, *Origins of DNA replication*, PLOS Genetics, 15 (2019), p. 21 pages.

[42] R. El-Awady, E. Saleh, A. Hashim, N. Soliman, A. Dallah, A. Elrasheed, and G. Elakraa, *The Role of Eukaryotic and Prokaryotic ABC Transporter Family in Failure of Chemotherapy*, Frontiers in Pharmacology, 7 (2017), p. 15 pages.

[43] L. Fernández, D. Gutiérrez, A. Rodríguez, and P. García, *Application of Bacteriophages in the Agro-Food Sector: A Long Way Toward Approval*, Frontiers in Cellular and Infection Microbiology, 0 (2018).

[44] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, *The Pfam protein families database*, Nucleic Acids Research, 40 (2011), pp. D290–D301.

[45] N. Fluman and E. Bibi, *Bacterial multidrug transport through the lens of the major facilitator superfamily*, Biochimica et Biophysica Acta, 1794 (2009), pp. 738–747.

[46] O. Gaillot, P. di Camillo, P. Berche, R. Courcol, and C. Savage, *Comparison of CHROMagar Salmonella medium and hektoen enteric agar for isolation of salmonellae from stool samples*, Journal of Clinical Microbiology, 37 (1999), pp. 762–765.

[47] K. Y. Gfeller, *Molecular analysis of antimicrobial resistance determinants of commensal lactobacilli*, PhD thesis, s.n., Huelva, Spain, 2003.

[48] S. E. M. Ginty, D. J. Rankin, and S. P. Brown, *Horizontal Gene Transfer and The Evolution of Bacterial Cooperation*, Evolution, 65 (2011), pp. 21–32.

[49] F. Griffith, *The significance of pneumococcal types*, Epidemiology and Infection, 27 (1928), pp. 113–159.

[50] A. C. Groth and M. P. Calos, *Phage integrases: biology and applications*, Journal of Molecular Biology, 335 (2004), pp. 667–678.

[51] F.-B. Guo, W. Wei, X. Wang, H. Lin, H. Ding, J. Huang, and N. Rao, *Co-evolution of genomic islands and their bacterial hosts revealed through phylogenetic analyses of 17 groups of homologous genomic islands*, Genetics and Molecular Research, 11 (2012), pp. 3735–3743.

[52] F.-B. Guo, Z.-K. Xia, W. Wei, and H.-L. Zhao, *Statistical analyses of conserved features of genomic islands in bacteria*, Genetics and Molecular Research, 13 (2014), pp. 1782–1793.

[53] C. Gyles and P. Boerlin, *Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease*, Veterinary Pathology, 51 (2014), pp. 328–340.

[54] J. Hacker, G. Blum-Oehler, I. Mühldorfer, and H. Tschäpe, *Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution*, Molecular Microbiology, 23 (1997), pp. 1089–1097.

[55] J. Hacker and J. B. Kaper, *Pathogenicity islands and the evolution of microbes*, Annual Reviews in Microbiology, 54 (2000), pp. 641–679.

[56] W. B. Hania, M. Joseph, B. Bunk, C. Spröer, H.-P. Klenk, M.-L. Fardeau, and S. Spring, *Characterization of the first cultured representative of a* Bacteroidetes *clade specialized on the scavenging of cyanobacteria*, Environmental Microbiology, 19 (2017), pp. 1134–1148.

[57] G. F. Hatfull and R. W. Hendrix, *Bacteriophages and their genomes*, Current Opinion in Virology, 1 (2011), pp. 298–303.

[58] B. Holland, S. P. C. Cole, K. Kuchler, and C. F. Higgins, *ABC proteins: from bacteria to man*, Academic Press, London, UK, 2003.

[59] M. Homann, P. Sansjofre, M. V. Zuilen, C. Heubeck, J. Gong, B. Killingsworth, I. S. Foster, A. Airo, M. J. V. Kranendonk, M. Ader, and S. V. Lalonde, *Author correction: Microbial life and biogeochemical cycling on land 3,220 million years ago*, Nature Geoscience, 11 (2018), pp. 965–965.

[60] W. Hsiao, I. Wan, S. J. Jones, and F. S. L. Brinkman, *IslandPath: aiding detection of genomic islands in prokaryotes*, Bioinformatics, 19 (2003), pp. 418–420.

[61] W. W.-L. Hsiao, *Computational characterization of prokaryotic genomic islands*, PhD thesis, Dept. of Molecular Biology and Biochemistry - Simon Fraser University, 2007.

[62] W. W. L. Hsiao, K. Ung, D. Aeschliman, J. Bryan, B. B. Finlay, and F. S. L. Brinkman, *Evidence of a large novel gene pool associated with prokaryotic genomic islands*, PLOS Genetics, 1 (2005), p. e62.

[63] C. M. Hudson, B. Y. Lau, and K. P. Williams, *Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes*, Nucleic Acids Research, 43 (2015), pp. D48–D53.

[64] Y. Hur, M. Chalita, S. min Ha, I. Baek, and J. Chun, *VCGIDB: A database and web resource for the genomic islands from vibrio cholerae*, Pathogens, 8 (2019), p. 261.

[65] B. L. Hurwitz, J. M. U'Ren, and K. Youens-Clark, *Computational prospecting the great viral unknown*, FEMS Microbiol. Lett., 363 (2016), p. fnw077.

[66] J. A. Johnson, A. B. Onderdonk, L. A. Cosimi, S. Yawetz, B. A. Lasker, S. J. Bolcen, J. M. Brown, and F. M. Marty, Gordonia bronchialis *Bacteremia and Pleural Infection: Case Report and Review of the Literature*, Journal of Clinical Microbiology, 49 (2011), p. 1662.

[67] D. Jones and P. Sneath, *Genetic transfer and bacterial taxonomy.*, Bacteriological Reviews, 34 (1970), p. 40.

[68] S. Kala, N. Cumby, P. D. Sadowski, B. Z. Hyder, V. Kanelis, A. R. Davidson, and K. L. Maxwell, *HNH proteins are a widespread component of phage DNA packaging machines*, Proceedings of the National Academy of Sciences of the United States of America, 111 (2014), pp. 6022–6027.

[69] M. I. Kanipes and P. Guerry, *Role of microbial glycosylation in host cell invasion*, in Microbial Glycobiology, Academic Press, Cambridge, MA, USA, Jan 2010, pp. 871–883.

[70] S. Karlin, *Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes*, Trends in Microbiology, 9 (2001), pp. 335–343.

[71] E. Kay, T. M. Vogel, F. Bertolla, R. Nalin, and P. Simonet, *In situ transfer of antibiotic resistance genes from transgenic (transplastomic) tobacco plants to bacteria*, Applied and Environmental Microbiology, 68 (2002), pp. 3345–3351.

[72] E. C. Keen, V. V. Bliskovsky, F. Malagon, J. D. Baker, J. S. Prince, J. S. Klaus, and S. L. Adhya, *Novel superspreader bacteriophages promote horizontal gene transfer by transformation*, mBio, 8 (2017).

[73] H. O. Khalifa, A. F. Oreiby, A. A. A. El-Hafeez, T. Okanda, A. Haque, K. S. Anwar, M. Tanaka, K. Miyako, S. Tsuji, Y. Kato, and T. Matsumoto, *First Report of Multidrug-Resistant Carbapenemase-Producing Bacteria Coharboring* mcr-9 *Associated with Respiratory Disease Complex in Pets: Potential of Animal-Human Transmission*, Antimicrobial Agents and Chemotherapy, 65 (2020), pp. 01890–01820.

[74] S. Kim, E. Seol, S. Mohan Raj, S. Park, Y.-K. Oh, and D. D. Y. Ryu, *Various hydrogenases and formate-dependent hydrogen production in* Citrobacter amalonaticus *Y19*, International Journal of Hydrogen Energy, 33 (2008), pp. 1509–1515.

[75] H. Kiss, E. Lang, A. Lapidus, A. Copeland, M. Nolan, T. G. Del Rio, F. Chen, S. Lucas, H. Tice, J.-F. Cheng, C. Han, L. Goodwin, S. Pitluck, K. Liolios, A. Pati, N. Ivanova, K. Mavromatis, A. Chen, K. Palaniappan, M. Land, L. Hauser, Y.-J. Chang, C. D. Jeffries, J. C. Detter, T. Brettin, S. Spring, M. Rohde, M. Göker, T. Woyke, J. Bristow, J. A. Eisen, V. Markowitz, P. Hugenholtz, N. C. Kyrpides, and H.-P. Klenk, *Com-*

plete genome sequence of Denitrovibrio acetiphilus *type strain (N2460T)*, Standards in Genomic Sciences, 2 (2010), p. 270.

[76] E. V. Koonin, K. S. Makarova, and L. Aravind, *Horizontal gene transfer in prokaryotes: Quantification and classification*, Annual Review of Microbiology, 55 (2001), pp. 709–742.

[77] E. Kutter and A. Sulakvelidze, *Bacteriophages Biology and Applications*, CRC Press, 2004.

[78] M. G. Langille, W. W. Hsiao, and F. S. Brinkman, *Evaluation of genomic island predictors using a comparative genomics approach*, BMC Bioinformatics, 9 (2008), p. 329.

[79] M. G. I. Langille, W. W. L. Hsiao, and F. S. L. Brinkman, *Detecting genomic islands using bioinformatics approaches*, Nature Reviews Microbiology, 8 (2010), pp. 373–382.

[80] D. F. Lato and G. B. Golding, *Spatial Patterns of Gene Expression in Bacterial Genomes*, Journal of Molecular Evolution, 88 (2020), p. 510.

[81] D. M. Lin, B. Koskella, and H. C. Lin, *Phage therapy: An alternative to antibiotics in the age of multi-drug resistance*, World J. Gastrointest. Pharmacol. Ther., 8 (2017), p. 162.

[82] B. A. Lipsky, E. W. Hook, A. A. Smith, and J. J. Plorde, *Citrobacter Infections in Humans: Experience at the Seattle Veterans Administration Medical Center and a Review of the Literature*, Reviews of Infectious Diseases, 2 (1980), pp. 746–760.

[83] G. Litwack, *Human Biochemistry*, Elsevier, 2018.

[84] C. Loc-Carrillo and S. T. Abedon, *Pros and cons of phage therapy*, Bacteriophage, 1 (2011), p. 111.

[85] P. Macheboeuf, C. Contreras-Martel, V. Job, O. Dideberg, and A. Dessen, *Penicillin Binding Proteins: key players in bacterial cell cycle and drug resistance processes*, FEMS Microbiology Reviews, 30 (2006), pp. 673–691.

[86] P. Mackiewicz, D. Mackiewicz, M. Kowalczuk, and S. Cebrat, *Flip-flop around the origin and terminus of replication in prokaryotic genomes*, Genome Biology, 2 (2001), p. interactions1004.1.

[87] G. N. Maertens, A. N. Engelman, and P. Cherepanov, *Structure and function of retroviral integrase - Nature Reviews Microbiology*, Nature Reviews Microbiology, 20 (2022), pp. 20–34.

[88] Y. Mantri, *Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities*, Nucleic Acids Research, 32 (2004), pp. 55D–58.

[89] M. D. Marger and M. H. Saier, Jr., *A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport*, Trends in Biochemical Sciences, 18 (1993), pp. 13–20.

[90] M. E. Milton, B. M. Minrovic, D. L. Harris, B. Kang, D. Jung, C. P. Lewis, R. J. Thompson, R. J. Melander, D. Zeng, C. Melander, and J. Cavanagh, *Re-sensitizing Multidrug Resistant Bacteria to Antibiotics by Targeting Bacterial Response Regulators: Characterization and Comparison of Interactions between 2-Aminoimidazoles and the Response Regulators BfmR from* Acinetobacter baumannii *and QseB from* Francisella *spp.*, Frontiers in Molecular Biosciences, 5 (2018), p. 12 pages.

[91] C. M. Mizuno, T. Luong, R. Cederstrom, M. Krupovic, L. Debarbieux, and D. R. Roach, *Isolation and Characterization of Bacteriophages That Infect* Citrobacter rodentium*, a Model Pathogen for Intestinal Diseases*, Viruses, 12 (2020), p. 737.

[92] S. Myhr and T. Torsvik, Denitrovibrio acetiphilus*, a novel genus and species of dissimilatory nitrate-reducing bacterium isolated from an oil reservoir model column.*, International Journal of Systematic and Evolutionary Microbiology, 50 (2000), pp. 1611–1619.

[93] H. Neve, U. Kemper, A. Geis, and K. J. Heller, *Monitoring and characterization of lactococcal bacteriophages in a dairy plant*, Kieler Milchwirtschaftliche Forschungsberichte, 46 (1994), pp. 167–178.

[94] E. Nicolas, M. Lambin, D. Dandoy, C. Galloy, N. Nguyen, C. A. Oger, and B. Hallet, *The Tn3-family of Replicative Transposons*, Microbiology Spectrum, 3 (2015), p. 3.4.14.

[95] Y. D. Niu, T. A. McAllister, J. H. E. Nash, A. M. Kropinski, and K. Stanford, *Four Escherichia coli O157:H7 Phages: A New Bacteriophage Genus and Taxonomic Classification of T1-Like Phages*, PLoS One, 9 (2014), p. e100426.

[96] H. Ochman, J. G. Lawrence, and E. A. Groisman, *Lateral gene transfer and the nature of bacterial innovation*, Nature, 405 (2000), pp. 299–304.

[97] C. M. O'Hara, A. G. Steigerwalt, B. C. Hill, r. J J. Farmer, G. R. Fanning, and D. J. Brenner, Enterobacter hormaechei*, a new species of the family* Enterobacteriaceae *formerly known as enteric group 75.*, Journal of Clinical Microbiology, 27 (1989), p. 2046.

[98] A. Paauw, M. P. M. Caspers, M. A. L.-v. Hall, F. H. J. Schuren, R. C. Montijn, J. Verhoef, and A. C. Fluit, *Identification of resistance and virulence factors in an epidemic* Enterobacter hormaechei *outbreak strain*, Microbiology, 155 (2009), pp. 1478–1488.

[99] S. R. Partridge, S. M. Kwong, N. Firth, and S. O. Jensen, *Mobile Genetic Elements Associated with Antimicrobial Resistance*, Clinical Microbiology Reviews, 31 (2018), pp. e00088–17.

[100] J. M. Pascual, D. Wang, B. Lecumberri, H. Yang, X. Mao, R. Yang, and D. C. De Vivo, *GLUT1 deficiency and other glucose transporter diseases*, European Journal of Endocrinology, 150 (2004), pp. 627–633.

[101] R. Pierneef, L. Cronje, O. Bezuidt, and O. N. Reva, *Pre_GI: a global map of ontological links between horizontally transferred genomic islands in bacterial and archaeal genomes*, Database, 2015 (2015).

[102] A. V. C. Pilar, N. Petronella, F. M. Dussault, A. J. Verster, S. Bekal, R. C. Levesque, L. Goodridge, and S. Tamber, *Similar yet different: phylogenomic analysis to delineate* Salmonella *and* Citrobacter *species boundaries*, BMC Genomics, 21 (2020), pp. 1–13.

[103] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, *A systematic comparison and evaluation of biclustering methods for gene expression data*, Bioinformatics, 22 (2006), pp. 1122–1129.

[104] N. Principi, E. Silvestri, and S. Esposito, *Advantages and Limitations of Bacteriophages for the Treatment of Bacterial Infections*, Front. Pharmacol., 0 (2019).

[105] P. G. Quirk, A. A. Guffanti, S. Clejan, J. Cheng, and T. A. Krulwich, *Isolation of Tn*917 *insertional mutants of* Bacillus subtilis *that are resistant to the protonophore carbonyl cyanide m-chlorophenylhydrazone*, Biochimica et Biophysica Acta, 1186 (1994), pp. 27–34.

[106] I. Rajan, S. Aravamuthan, and S. S. Mande, *Identification of compositionally distinct regions in genomes using the centroid method*, Bioinformatics, 23 (2007), pp. 2672–2677.

[107] M. Rajewska, K. Wegrzyn, and I. Konieczny, *AT-rich region and repeated sequences – the essential elements of replication origins of bacterial replicons*, FEMS Microbiol. Rev., 36 (2012), pp. 408–434.

[108] G. Rizzatti, L. R. Lopetuso, G. Gibiino, C. Binda, and A. Gasbarrini, *Proteobacteria: A Common Factor in Human Diseases*, BioMed Research International, 2017 (2017), p. 7 pages.

[109] M. M. Royam and R. Nachimuthu, *Isolation, characterization, and efficacy of bacteriophages isolated against* Citrobacter *spp. an* in vivo *approach in a zebrafish model (*Danio rerio*)*, Research in Microbiology, 171 (2020), pp. 341–350.

[110] J. Schmid, D. Heider, N. J. Wendel, N. Sperl, and V. Sieber, *Bacterial Glycosyltransferases: Challenges and Opportunities of a Highly Diverse Enzyme Class Toward Tailoring Natural Products*, Frontiers in Microbiology, 7 (2016), p. 7 pages.

[111] H. Schmidt and M. Hensel, *Pathogenicity islands in bacterial pathogenesis*, Clinical Microbiology Reviews, 19 (2006), p. 257.

[112] J. O. Sekyere and M. A. Reta, *Genomic and Resistance Epidemiology of Gram-Negative Bacteria in Africa: a Systematic Review and Phylogenomic Analyses from a One Health Perspective*, mSystems, 5 (2020), pp. e00897–20.

[113] G. Sevillya, O. Adato, and S. Snir, *Detecting horizontal gene transfer: a probabilistic approach*, BMC Genomics, 21 (2020).

[114] F.-T. Shen, M. Goodfellow, A. L. Jones, Y.-P. Chen, A. B. Arun, W.-A. Lai, P. D. Rekha, and C.-C. Young, Gordonia soli *sp. nov., a novel actinomycete isolated from soil*, International Journal of Systematic and Evolutionary Microbiology, 56 (2006), pp. 2597–2601.

[115] K. Shi, C. Li, C. Rensing, X. Dai, X. Fan, and G. Wang, *Efflux Transporter ArsK Is Responsible for Bacterial Resistance to Arsenite, Antimonite, Trivalent Roxarsone, and Methylarsenite*, Applied and Environmental Microbiology, 84 (2018), pp. e01842–18.

[116] P. M. Shih, D. Wu, A. Latifi, S. D. Axen, D. P. Fewer, E. Talla, A. Calteau, F. Cai, N. T. de Marsac, R. Rippka, M. Herdman, K. Sivonen, T. Coursin, T. Laurent, L. Goodwin, M. Nolan, K. W. Davenport, C. S. Han, E. M. Rubin, J. A. Eisen, T. Woyke, M. Gugger, and C. A. Kerfeld, *Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing*, Proceedings of the National Academy of Sciences of the United States of America, 110 (2013), pp. 1053–1058.

[117] J. M. Skerker, M. S. Prasol, B. S. Perchuk, E. G. Biondi, and M. T. Laub, *Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis*, PLOS Biology, 3 (2005), p. e334.

[118] S. C. Soares, V. A. C. Abreu, R. T. J. Ramos, L. Cerdeira, A. Silva, J. Baumbach, E. Trost, A. Tauch, R. Hirata, A. L. Mattos-Guaraldi, A. Miyoshi, and V. Azevedo, *PIPS: Pathogenicity island prediction software*, PLoS ONE, 7 (2012), p. e30848.

[119] L. Søgaard-Andersen, *Discovery of a Diverse Set of Bacteria That Build Their Cell Walls without the Canonical Peptidoglycan Polymerase aPBP*, mBio, 12 (2021), pp. e01342–21.

[120] F. Solinas, A. M. Marconi, M. Ruzzi, and E. Zennaro, *Characterization and sequence of a novel insertion sequence, IS*162*, from* Pseudomonas fluorescens, Gene, 155 (1995), pp. 77–82.

[121] S. Spring, B. Bunk, C. Spröer, P. Schumann, M. Rohde, B. J. Tindall, and H.-P. Klenk, *Characterization of the first cultured representative of* Verrucomicrobia *subdivision 5 indicates the proposal of a novel phylum*, The ISME Journal, 10 (2016), pp. 2801–2816.

[122] A. Stern and R. Sorek, *The phage-host arms race: shaping the evolution of microbes*, Bioessays, 33 (2011), pp. 43–51.

[123] C. A. Suttle, *Viruses in the sea*, Nature, 437 (2005), pp. 356–361.

[124] M. Syvanen and C. Kado, *Horizontal Gene Transfer*, Academic Press, Cambridge, MA, USA, Dec 2001.

[125] R. Tam and M. H. Saier, Jr., *Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria*, Microbiological Reviews, 57 (1993), pp. 320–346.

[126] E. Tatum and J. Lederberg, *Gene recombination in the bacterium escherichia coli*, Journal of bacteriology, 53 (1947), p. 673.

[127] J. Tian, L. Wang, P. Liu, Y. Geng, G. Zhu, R. Zheng, Z. Liu, Y. Zhao, J. Yang, and F. Peng, Deinococcus psychrotolerans *sp. nov., isolated from soil on the South Shetland Islands, Antarctica*, International Journal of Systematic and Evolutionary Microbiology, 69 (2019), pp. 3696–3701.

[128] Q. Tu and D. Ding, *Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis*, FEMS Microbiology Letters, 221 (2003), pp. 269–275.

[129] G. S. Vernikos and J. Parkhill, *Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the salmonella pathogenicity islands*, Bioinformatics, 22 (2006), pp. 2196–2203.

[130] G. S. Vernikos and J. Parkhill, *Resolving the structural features of genomic islands: A machine learning approach*, Genome Research, 18 (2008), pp. 331–342.

[131] C. J. H. von Wintersdorff, J. Penders, J. M. van Niekerk, N. D. Mills, S. Majumder, L. B. van Alphen, P. H. M. Savelkoul, and P. F. G. Wolffs, *Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer*, Frontiers in Microbiology, 7 (2016).

[132] S. Waack, O. Keller, R. Asper, T. Brodag, C. Damm, W. F. Fricke, K. Surovcik, P. Meinicke, and R. Merkl, *Score-based prediction of genomic islands in prokaryotic genomes using hidden markov models*, BMC Bioinformatics, 7 (2006), p. 142.

[133] L.-F. Wang and M. Yu, *Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics*, Curr. Drug Targets, 5 (2004), pp. 1–15.

[134] Z. Wang, L. Duan, F. Liu, Y. Hu, C. Leng, Y. Kan, L. Yao, and H. Shi, *First report of* Enterobacter hormaechei *with respiratory disease in calves*, BMC Veterinary Research, 16 (2020), pp. 1–4.

[135] W. Wei and F.-B. Guo, *Prediction of genomic islands in seven human pathogens using the z-island method*, Genetics and Molecular Research, 10 (2011), pp. 2307–2315.

[136] C. R. Woese and G. E. Fox, *Phylogenetic structure of the prokaryotic domain: the primary kingdoms*, Proceedings of the National Academy of Sciences, 74 (1977), pp. 5088–5090.

[137] S. N. Wright, J. S. Gerry, M. T. Busowski, A. Y. Klochko, S. G. McNulty, S. A. Brown, B. E. Sieger, P. K. Michaels, and M. R. Wallace, Gordonia bronchialis *sternal wound infection in 3 patients following open heart surgery: intraoperative transmission from a healthcare worker*, Infection Control and Hospital Epidemiology, 33 (2012), pp. 1238–1241.

[138] L. Ye, C. Hou, and S. Liu, *The Role of Metabolizing Enzymes and Transporters in Antiretroviral Therapy*, Current Topics in Medicinal Chemistry, 17 (2017), pp. 340–360.

[139] C. C. Yeo and C. L. Poh, *Characterization of IS*1474*, an insertion sequence of the IS*21 *family isolated from* Pseudomonas alcaligenes *NCIB 9867*, FEMS Microbiology Letters, 149 (1997), pp. 257–263.

[140] S. H. Yoon, Y.-K. Park, and J. F. Kim, *PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands*, Nucleic Acids Research, 43 (2015), pp. D624–D630.

[141] S. H. Yoon, Y.-K. Park, S. Lee, D. Choi, T. K. Oh, C.-G. Hur, and J. F. Kim, *Towards pathogenomics: a web-based resource for pathogenicity islands*, Nucleic Acids Research, 35 (2007), pp. D395–D400.

[142] J. P. Zehr, S. R. Bench, B. J. Carter, I. Hewson, F. Niazi, T. Shi, H. J. Tripp, and J. P. Affourtit, *Globally Distributed Uncultivated Oceanic N_2-Fixing Cyanobacteria Lack Oxygenic Photosystem II*, Science, 322 (2008), pp. 1110–1112.

[143] L. Zhang, D. Xu, Y. Huang, X. Zhu, M. Rui, T. Wan, X. Zheng, Y. Shen, X. Chen, K. Ma, and Y. Gong, *Structural and functional characterization of deep-sea thermophilic bacteriophage GVE2 HNH endonuclease*, Scientific Reports, 7 (2017), pp. 1–13.

[144] R. Zhang and C.-T. Zhang, *A systematic method to identify genomic islands and its applications in analyzing the genomes of corynebacterium glutamicum and vibrio vulnificus CMCP6 chromosome i*, Bioinformatics, 20 (2004), pp. 612–622.

[145] R. Zhang and C.-T. Zhang, *Genomic islands in the corynebacterium efficiens genome*, Applied and Environmental Microbiology, 71 (2005), pp. 3126–3130.

[146] R. Zhang and C.-T. Zhang, *Accurate localization of the integration sites of two genomic islands at single-nucleotide resolution in the genome ofBacillus cereusATCC 10987*, Comparative and Functional Genomics, 2008 (2008), pp. 1–6.

[147] X. Zhang, C. Peng, G. Zhang, and F. Gao, *Comparative analysis of essential genes in prokaryotic genomic islands*, Scientific Reports, 5 (2015), pp. 1–9.

[148] N. D. Zinder and J. Lederberg, *Genetic exchange in salmonella*, Journal of bacteriology, 64 (1952), p. 679.

[149] J. Zuegg, C. Muldoon, G. Adamson, D. McKeveney, G. Le Thanh, R. Premraj, B. Becker, M. Cheng, A. G. Elliott, J. X. Huang, M. S. Butler, M. Bajaj, J. Seifert, L. Singh, N. F. Galley, D. I. Roper, A. J. Lloyd, C. G. Dowson, T.-J. Cheng, W.-C. Cheng, D. Demon, E. Meyer, W. Meutermans, and M. A. Cooper, *Carbohydrate scaffolds as glycosyltransferase inhibitors with* in vivo *antibacterial activity*, Nature Communications, 6 (2015), pp. 1–11.

Appendices

Appendix A

A.1 Algorithms

Algorithm 1: Patterns

Input: GIs, GIs_Proteins
Output: Patterns Files

```
Apriori_Sets ← Apriori(GIs Proteins)
Patterns=[]
for S in Apriori_Sets do
    Set_Patterns=itertools.permutations(S)
    for SP in list(Set_Patterns) do
        Patterns.append(SP)
    end
end
Interesting_Patterns=[]
for P in Patterns do
    if 3 ≤ length(P) ≤ 5 then
        Pattern_GIs=[]
        for GI in GIs do
            if P in GI then
                Pattern_GIs.append(GI)
            end
        end
        (Superkingdom,Phylum,Class,Order,Family,Genus,Species) ← Taxonomy(Pattern_GIs)
        if len(Pattern_GIs) > 10 then
            if Most_Frq_Order = 7 then
                Interesting_Patterns.append ←
                  (P,Superkingdom,Phylum,Class,Order,Family,Genus,Species, len(Pattern_GIs))
            end
        end
    end
end
return Interesting_Patterns
```

Algorithm 2: Presence of Phages in the Patterns

```
Input: Patterns Files
Output: Patterns_Phages Proteins Files
for P in Patterns do
    Pattern_GIs ← Retrieve_GIs(P)                    ▷ Retrieve the GIs that have the P pattern
    for GI in Pattern_GIs do
        GI_Proteins ← Pattern_GI_Proteins(GI,P)              ▷ Retrieve the proteins of the GI
        InputFile_GI_Proteins=open(Input_File.txt,w)
        InputFile_GI_Proteins.write(GI_Proteins)
        OutputFile_GI_Viral_Proteins=open(P+'_'+GI+'.txt',w)
        blastp -db Viral_DataBase -query InputFile_GI_Proteins -out
          OutputFile_GI_Viral_Proteins -evalue 10e-10
    end
end
```

Algorithm 3: Patterns — Phage Information

```
Input: Patterns_PhagesProteins Files
Output: Patterns_PhagesProteinsInfo Files
for P in Patterns do
    PatternFile=open(P+'.txt')
    Pattern_GIs ← Retrieve_GIs(P)
    for GI in Pattern_GIs do
        File=read(P+'_'+GI+'.txt')
        index = −1
        BP =''
        for Line in File do
            (BacteriaProtein, PhageProteinAccessoin, Evalue) = Get_Line_Data(Line)
            (OrganismName , OrganismAccession)=EFetch(PhageProteinAccessoin)
            if index==-1 or BacteriaProtein != BP then
                index=index+1
                Organism_Name.append([])
                Organism_Acc.append([])
                Protein_Acc.append([])
                Protein_Evalue.append([])
            end
            Organism_Name[index].append(OrganismName)
            Organism_Acc[index].append(OrganismAccession)
            Protein_Acc[index].append(PhageAccessoin)
            Protein_Evalue[index].append(Evalue)
            BP=BacteriaProteiny
        end
        if index ==length(P) then
            GI_Phage_Intersection=Intersection(Organism_Acc)
            for Phage in GI_Phage_Intersection do
                PatternFile.write(GI, Phage, Max(E value))
            end
        end
    end
end
```

Algorithm 4: Extracting HGT Connections

Input: Patterns_PhagesProteinsInfo Files

Output: Phage_Bacteria_Connections File

```
for P in Patterns do
    PatternFile=read(P+'.txt')
    PatternPhageInfo=write('PhagesInGIs'+P+'.txt')
    Phages=[], Seen=[], P_Phages=[], P_GIs=[], P_Evalue=[]
    for Line in PatternFile do
        (Phage, GI, Evalue) ← Get_Data(Line)
        P_Phages.append(Phage),  P_GIs.append(GI),  P_Evalue.append(Evalue)
        if Phage not in Seen then
            Phages.append(Phage),  Seen.append(Phage)
        end
    end
    for Phage in Phages do
        P_Index=0, Bacteria_GIs=[], Evalues=[]
        for PP in P_Phages do
            P_Index+=1
            if Phage == PP then
                Bacteria_GIs.append(P_GIs[P_Index])
                Evalues.append(P_Evalue[P_Index])
            end
        end
        PatternPhageInfo.write(P, Phage, Bacteria_GIs, Evalues)
    end
end
```

Algorithm 5: Filtering HGT Connections — Taxonomy

Input: Phage_Bacteria_connections File

Output: Phage_Bacteria_connections_TaxFilter File

```
for P in Interesting_Patterns do
    PatternFile=read('PhagesInGIs'+P+'.txt')
    GIsLineageFile=write('GIsLineage'+P+'.txt')
    GIsEvaluesFile=write('GIsLineageEvalues'+P+'.txt')
    for Line in PatternFile do
        Bacteria_List ← Get_Bacteria(Line)
        Lineage=[]
        for B in Bacteria_List do
            Lineage.append(Get_Lineage(B))
        end
        Common_Lineage=Intersection(Lineage)
        if Length (Common_Lineage) <5 then
            GIsLineageFile.write(Line, Common_Lineage)
            Bacteria_GIs=[],    Bacteria_Evalues=[]
            for B in Bacteria_List do
                if Evalue <10e-100 then
                    Bacteria.append(B)
                end
            end
            Phage ← GetPhage(Line)
            GIsEvaluesFile.write(Phage, Bacteria_GIs, Bacteria_Evalues)
        end
    end
end
```

Algorithm 6: Connections

Input: Phage_Bacteria_connections_TaxFilter File

Output: Novel_HGT_connections File

```
PatternFile=read('GIsLineageEvalues'+P+'.txt')
Phages_vs_Bacteria=write(PhagesVSBacteria'+P+'.txt')
Bacteria_vs_Bacteria=write(BacteriaVSBacteria'+P+'.txt')
HGT=write(HGT.txt)
for Line in PatternFile do
    (Phage, GIs, Bacteria_Name) ← Get_Data(Line)
    (Phage_Proteins) ← Get_Proteins(Phage)
    (Bacteria_Proteins) ← Get_Proteins(GIs)
    Flage1=0, Flage2=0
    PhageVSBacteria=BLAST(Phage_Proteins , Bacteria_Proteins)
    for FB in PhageVSBacteria do
        if (FB.ScientificName == Bacteria_Name) and (FB.Identity ≥ 85) and
          (TextSimilarity(FB.ScientificName,Bacteria_Name) <0.4) then
            Phages_vs_Bacteria.write(Phage, FB.GI)
            Flage1=1
        end
    end
    BacteriaVSBacteria=BLAST_BothDirections(Bacteria_Proteins , Bacteria_Proteins)
    for FB in PhageVSBacteria do
        if (FB.ScientificName == Bacteria_Name) and (FB.Identity ≥ 85) and
          (TextSimilarity(FB.ScientificName,Bacteria_Name) <0.4) then
            Bacteria_vs_Bacteria.write(FB.GI, FB.GI)
            Flage2=1
        end
    end
    if Flage1==1 and Flage2==1 and ((Phage, GIs) ∩ (GIs, GIs)) ≠ ∅ then
        HGT.write((Phage,GIs),(GIs,GIs))
    end
end
```

Algorithm 7: Connections — Extract Species Subsequences

Input: Novel_HGT_connections File
Output: Blast Results, Clustal

```
for connection in connectionsFile do
    for i in connectionsBacteria do
        Coordinates ← BLAST(Phage,Bacteria)
    end
    //Get Phage Coordinates
    if Phage_coordinates_Equal: then
        (Start, End) ← Get_any_phage_coordinates()
    end
    else if Phage_Starts_Equal then
        End ← Get_Phage_coordinates_End()
    end
    else if Phage_Ends_Equal then
        Start ← Get_Phage_coordinates_Start()
    end
    else
        //Phage_coordinates_not_Equal
        (Start, End) ← Compute_phage_coordinates()
    end
    Phage subsequence ← Extract_Phage_subsequence (Start, End)
    PhageFile ← write(Phage subsequence)
    ClustalFile ← write(Phage subsequence)
    //Get Bacteria Coordinates
    if length(Bacteria)==1 then
        //Onebacteria
        Clustal ← write(Bacteria Subsequence)
    end
    else
        //More than one bacteria
        for i in Bacteria do
            Difference = abs((PhageS − PhageE + 1) − (abs(Bacterium_E − Bacterium_S) + 1))
            X = Difference/2;   A = int(X);   Y = X − A
            if Bacteria_Phage_Length_Difference == 0 then
                Clustal ← write(Bacteria Subsequence)
            end
            else if Bacteria_Phage_Length_Difference > 1200 then
                New_Coordinates ← BLAST(phage, bacteria(i))
                Clustal ← write(Bacteria Subsequence (New_Coordinates))
            end
            else if Bacteria_Phage_Length_Difference > 0 then
                Clustal ← write(Bacteria_Subsequence_Extend)
            end
            else
                Clustal ← write(Bacteria_Subsequence_Trim)
            end
        end
    end
end
```

Algorithm 8: The Forming of Genomic Islands

Input: Genomic Islands

Output: The updated version of the genomic islands

```
For each Genome G:
    Sort_GIs_Coordinates() // Sort the GIs in Genome G according to the start point
    GI_Overlap = 0
    For each GI:
        GI = Get_GI_Name
        S = Get_GI_Start_Coordinate
        E = Get_GI_End_Coordinate
        IF GI_Overlap == 0 :
            GI_Overlap = 1,    S_Overlap = S,    E_Overlap = E
        ELSE:
            IF S >= S_Overlap and S <= E_Overlap:
                IF S < S_Overlap:
                    S_Overlap = S
                IF E > E_Overlap:
                    E_Overlap = E
            ELSE:
                Overlap_GIs.write(GI+ ' ' +S_Overlap+ ' ' +E_Overlap)
                GI_Overlap = 1,    S_Overlap = S,    E_Overlap = E
    IF GI_Overlap == 1:
        Overlap_GIs.write(GI+ ' ' +S_Overlap+ ' ' +E_Overlap)
```

Algorithm 9: The Location of the GIs in Relation to the oriC

Input: Genomic_Islands
Output: The distance from the GIs to the oriC

```
oriC=open('oriC.txt','w')
Species=open('Species.txt','w')
MostFreqSpecies=open('MostFeqSpecies.txt','w')
Species_Circular_Seen=[]; Species_Linear_Seen=[]
MostFreq_Species=['Escherichia coli', 'Salmonella enterica', 'Klebsiella pneumoniae', 'Bordetella
pertussis', 'Pseudomonas aeruginosa']
Middle=0; Oric=0; Species_Middle=0; Species_Oric=0
MostFreq_Middle=0; MostFreq_Oric=0;
For each Genome:
  For each GI in Genomic_Islands:
    G = Get_G(GI)
    S = Get_Start_GI(G,GI)/ Length (G)
    Structure=Get_Structure(GI)
    Species=Get_Species(G)
    IF Genome_Structure == "circular" :
      oriC.write(str(S)+'circular')
      IF 0.25 < Start <= 0.75: Middle = Middle + 1
      ELSE: oriC = oriC + 1
      IF (Species not in Species_Circular_Seen) or (not read all GIs) :
        Species.write(str(S)+'circular')
        Species_Circular_Seen.append(Species)
        IF 0.25 < Start <= 0.75: Species_Middle = Species_Middle + 1
        ELSE: Species_oriC = Species_oriC + 1
      IF Species in MostFreq_Species :
        MostFreqSpecies.write(str(S)+Species+'circular')
        IF 0.25 < Start <= 0.75: MostFreq_Middle.add(Species, 1)
        ELSE: MostFreq_oriC.add(Species, 1)
    ELIF Genome_Structure == "linear":
      oriC.write(str(S)+'linear')
      IF (Species not in Species_Linear_Seen) or (not read all GIs) :
        Species.write(str(S)+'linear')
        Species_Linear_Seen.append(Species)
      IF Species in MostFreq_Species :
        MostFreqSpecies.write(str(S)+Species+'linear ')
```

Algorithm 10: Compute the Distance between the GIs

Input: Genomic Islands, Genomes
Output: Distance between the GIs

```
Distance_GIs=open('Distance_GIs.txt','w')
Distance_Species=open('Distance_Species.txt','w')
Distance_MostFreqSpecies=open('Distance_MostFreqSpecies.txt','w')
Species_Circular_Seen=[]; Species_Linear_Seen=[]
MostFreq_Species=['Escherichia coli', 'Salmonella enterica', 'Klebsiella pneumoniae', 'Bordetella
pertussis', 'Pseudomonas aeruginosa']
For G in each Genomes:
   GIs=[]; S_Cr=[]; E_Cr=[];
   Species=Get_Species(G)
   For each GI in G:
      GI=Get_GI(G)
      S = Get_Start_GI(G,GI)/ Length (G)
      E = Get_End_GI(G,GI)/ Length (G)
      GIs.append(GI),    S_Cr.append(S),    E_Cr.append(E)
   //Compute the distances
   count = length(GIs)
   IF Genome_Structure == "circular":
      For i in range (0,count):
         IF i == (count − 1):
            Distance=round(abs( (S_Cr[0]/Genome_Size) + (1-(E_Cr[i]/Genome_Size)) ),5)
         ELSE:
            Distance=round(abs( (S_Cr[i+1]-E_Cr[i])/Genome_Size ),5)
         Distance_GIs.write(str(Distance),'circular')
         IF (G not in Species_Circular_Seen) or (not read all GIs) :
            Distance_Species.write(str(Distance),' circular')
         IF Species in MostFreq_Species :
            Distance_MostFreqSpecies.write(str(S)+Species+'circular')
   ELIF Genome_Structure == "linear":
      For i in range (0,count-1):
         Distance=round(abs( (S_Cr[i+1]-E_Cr[i])/Genome_Size ),5)
         Distance_GIs.write(str(Distance)+' linear')
         IF (G not in Species_Linear_Seen) or (not read all GIs) :
            Distance_Species.write(str(Distance),'linear')
         IF Species in MostFreq_Species :
            Distance_MostFreqSpecies.write(str(S)+Species+'linear')
   GIs.clear(); S_Cr.clear(); E_Cr.clear()
```

Algorithm 11: Distribution of the GIs

Input: Genomic Islands
Output: Distribution of GIs

```
GIs_Distribution=open('GIs_Distribution.txt', 'w')
n=10
   P=[]; a=1/n; sum=0
For i in range (0,n):
   sum+=a
   P.append(round(sum,1))
index=-1; Parts_list_Circular=[]; Parts_list_Linear=[];
For PL in range(0,n):
   Parts_list_Circular.append([]); Parts_list_Linear.append([]);
For i in GIs:
   index+=1
   Genome_IDX=Genome.index(i.split('_B_')[0])
   Genome_Size=int(size[Genome_IDX])
   Genome_Structure=Structure[Genome_IDX]
   Start=S[index]/Genome_Size; End=E[index]/Genome_Size
   Length=abs(End-Start); Prev_j= 0; P_index=-1
   For j in P:
      P_index+=1;
      IF Prev_j < Start <= j:
         IF Prev_j < End <= j:
            IF Genome_Structure == "circular": Parts_list_Circular[P_index].append(i)
            ELIF Genome_Structure == "linear": Parts_list_Linear[P_index].append(i)
         ELSE:
            IF j − Start > 0:
               IF Genome_Structure == "circular" : Parts_list_Circular[P_index].append(i)
               ELIF Genome_Structure == "linear" : Parts_list_Linear[P_index].append(i)
            For k in range (P_index,len(P)-1):
               IF P[k] < End <= P[k + 1]:
                  IF Genome_Structure == "circular": Parts_list_Circular[k+1].append(i)
                  ELIF Genome_Structure == "linear": Parts_list_Linear[k+1].append(i)
                  break
               ELSE:
                  IF Genome_Structure == "circular": Parts_list_Circular[k+1].append(i)
                  ELIF Genome_Structure == "linear": Parts_list_Linear[k+1].append(i)
      Prev_j=j
For i in Parts_list_Circular:
   GIs_Distribution.write(str(len(i))+'circular')
For i in Parts_list_Linear:
   GIs_Distribution.write(str(len(i))+'linear')
```

A.2 Tables

Table A.1: Connections (Phages BLAST Analysis)

CI1	CI2	Phage1 Acc.	Phage2 Acc.	Ident.	E Value	Ph1 Len	Ph2 Len	Overlap	Cov. 1	Cov. 2
54	58	NC_027339.1	NC_019708.1	94.737	2.43E-160	38389	37595	361	1%	1%
54	92	NC_027339.1	NC_004313.1	83.345	0	38389	40149	9649	25%	24%
54	94	NC_027339.1	NC_021857.1	98.164	0	38389	41475	12093	32%	29%
54	96	NC_027339.1	NC_022749.1	98.186	0	38389	39758	6008	16%	15%
58	73	NC_019708.1	MK422452.1	79.772	0	37595	34141	3421	9%	10%
58	94	NC_019708.1	NC_021857.1	88.354	9.77E-135	37595	41475	395	1%	1%
58	96	NC_019708.1	NC_022749.1	87.342	4.39E-128	37595	39758	395	1%	1%
58	385	NC_019708.1	MK416021.1	75.261	0	37595	38370	1722	5%	4%
58	571	NC_019708.1	MF695815.1	75.134	0	37595	39241	1673	4%	4%
64	200	LR881103.1	MF417956.1	90.841	0	16548	16570	3188	19%	19%
64	201	LR881103.1	MF417957.1	90.841	0	16548	16537	3188	19%	19%
64	571	LR881103.1	MF695815.1	79.545	1.41E-159	16548	39241	792	5%	2%
73	571	MK422452.1	MF695815.1	92.448	4.87E-157	34141	39241	384	1%	1%
73	573	MK422452.1	MK416014.1	92.664	0	34141	29301	1445	4%	5%
92	94	NC_004313.1	NC_021857.1	83.591	0	40149	41475	9629	24%	23%
92	96	NC_004313.1	NC_022749.1	82.286	0	40149	39758	11031	27%	28%
94	96	NC_021857.1	NC_022749.1	97.284	0	41475	39758	6739	16%	17%
200	201	MF417956.1	MF417957.1	99.982	0	16570	16537	16537	99%	100%
385	571	MK416021.1	MF695815.1	92.308	1.46E-08	38370	39241	39	0%	0%
385	573	MK416021.1	MK416014.1	81.407	1.26E-82	38370	29301	398	1%	1%
571	573	MF695815.1	MK416014.1	98.891	0	39241	29301	8385	21%	29%

Table A.2: Connections (Phages and Bacteria BLAST Analysis)

CI	Species Name	Species Accession	Species Name	Species Accession	Ident.
44	Bacteriophage_sp.	MN855837.1	Lactobacillus_curvatus	NZ_CP026116.1	90.268
44	Bacteriophage_sp.	MN855837.1	Lactobacillus_sakei	NZ_AP017931.1	93.356
44	Lactobacillus_curvatus	NZ_CP026116.1	Lactobacillus_sakei	NZ_AP017931.1	93.705
54	Enterobacteria_phage_SfI	NC_027339.1	Escherichia_albertii	NZ_CP030787.1	96.715
58	Enterobacteria_phage_mEp235	NC_019708.1	Cronobacter_sakazakii	NC_009778.1	80.314
64	Escherichia_phage_Henu7	LR881103.1	Klebsiella_michiganensis	NZ_CP035214.1	94.267
64	Escherichia_phage_Henu7	LR881103.1	Enterobacter_soli	NC_015968.1	90.602
64	Escherichia_phage_Henu7	LR881103.1	Pluralibacter_gergoviae	NZ_CP020388.1	87.451
64	Klebsiella_michiganensis	NZ_CP035214.1	Enterobacter_soli	NC_015968.1	90.365
64	Klebsiella_michiganensis	NZ_CP035214.1	Pluralibacter_gergoviae	NZ_CP020388.1	82.271
64	Enterobacter_soli	NC_015968.1	Pluralibacter_gergoviae	NZ_CP020388.1	94.589
73	Klebsiella_phage_ST13-OXA48phi12.2	MK422452.1	Citrobacter_amalonaticus	NZ_LT556084.1	80.409
92	Salmonella_phage_ST64B	NC_004313.1	Escherichia_albertii	NZ_CP030787.1	84.897
94	Shigella_phage_SfII	NC_021857.1	Escherichia_albertii	NZ_CP030787.1	96.734
96	Shigella_phage_SfIV	NC_022749.1	Escherichia_albertii	NZ_CP030787.1	96.989
200	uncultured_Caudovirales_phage	MF417956.1	Klebsiella_michiganensis	NZ_CP035214.1	88.169

Continued on next page

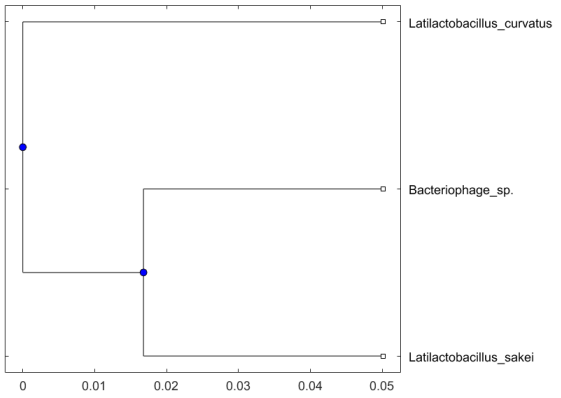
Table A.2 – continued from previous page

CI	Species Name	Species Accession	Species Name	Species Accession	Ident.
200	uncultured_Caudovirales_phage	MF417956.1	Enterobacter_soli	NC_015968.1	94.426
200	uncultured_Caudovirales_phage	MF417956.1	Pluralibacter_gergoviae	NZ_CP020388.1	90.268
200	Klebsiella_michiganensis	NZ_CP035214.1	Enterobacter_soli	NC_015968.1	90.365
200	Klebsiella_michiganensis	NZ_CP035214.1	Pluralibacter_gergoviae	NZ_CP020388.1	82.271
200	Enterobacter_soli	NC_015968.1	Pluralibacter_gergoviae	NZ_CP020388.1	94.589
201	uncultured_Caudovirales_phage	MF417957.1	Klebsiella_michiganensis	NZ_CP035214.1	88.169
201	uncultured_Caudovirales_phage	MF417957.1	Enterobacter_soli	NC_015968.1	94.426
201	uncultured_Caudovirales_phage	MF417957.1	Pluralibacter_gergoviae	NZ_CP020388.1	90.268
201	Klebsiella_michiganensis	NZ_CP035214.1	Enterobacter_soli	NC_015968.1	90.365
201	Klebsiella_michiganensis	NZ_CP035214.1	Pluralibacter_gergoviae	NZ_CP020388.1	82.271
201	Enterobacter_soli	NC_015968.1	Pluralibacter_gergoviae	NZ_CP020388.1	94.589
383	Klebsiella_phage_ST13-OXA48phi12.2	MK422452.1	Salmonella_enterica	NZ_CP037894.1	80.085
383	Klebsiella_phage_ST13-OXA48phi12.2	MK422452.1	Enterobacter_hormaechei	NZ_CP032841.1	81.947
383	Salmonella_enterica	NZ_CP037894.1	Enterobacter_hormaechei	NZ_CP032841.1	90.646
385	Klebsiella_phage_ST846-OXA48phi9.1	MK416021.1	Citrobacter_braakii	NZ_CP020448.2	77.904
385	Klebsiella_phage_ST846-OXA48phi9.1	MK416021.1	Citrobacter_tructae	NZ_CP038469.1	77.961
385	Citrobacter_braakii	NZ_CP020448.2	Citrobacter_tructae	NZ_CP038469.1	91.507
571	Klebsiella_phage_KPP5665-2	MF695815.1	Citrobacter_freundii	NZ_LS992183.1	84.107
571	Klebsiella_phage_KPP5665-2	MF695815.1	Pantoea_vagans	NZ_CP020820.1	78.460
571	Citrobacter_freundii	NZ_LS992183.1	Pantoea_vagans	NZ_CP020820.1	91.307
573	Klebsiella_phage_ST16-OXA48phi5.3	MK416014.1	Citrobacter_freundii	NZ_LS992183.1	83.455
573	Klebsiella_phage_ST16-OXA48phi5.3	MK416014.1	Pantoea_vagans	NZ_CP020820.1	78.541
573	Citrobacter_freundii	NZ_LS992183.1	Pantoea_vagans	NZ_CP020820.1	91.307
1385	Lactobacillus_phage_Sha1	NC_019489.1	Pediococcus_pentosaceus	NZ_CP021927.1	88.159

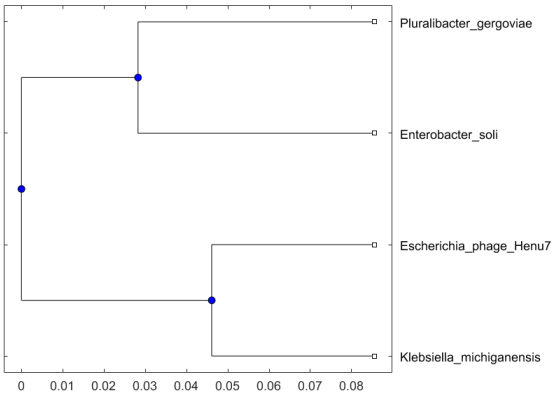
Table A.3: Connections (GIs Coordinates Information)

CI	Q Species	S Species	Q length	S length	A Len	Q Start	Q End	S Start	S End	GI Start	GI End
44	MN855837.1	NZ_CP026116.1	10442	1898076	5004	1	5001	723582	718580	698993	744389
44	MN855837.1	NZ_AP017931.1	10442	1950487	4997	1	4997	477956	482951	469294	485205
44	NZ_CP026116.1	NZ_AP017931.1	1898076	1950487	15124	1340969	1356052	129474	114392	-	-
54	NC_027339.1	NZ_CP030787.1	38389	4659718	5875	2	5876	1010412	1016286	994657	1036029
58	NC_019708.1	NC_009778.1	37595	4368373	5801	1	5757	988865	994638	956100	1008032
64	LR881103.1	NZ_CP035214.1	16548	5914592	4675	12126	16499	2378985	2374312	2371719	2390239
64	LR881103.1	NC_015968.1	16548	4812833	3107	12229	15327	4260640	4257536	4251287	4268785
64	LR881103.1	NZ_CP020388.1	16548	5408082	4327	12237	16548	403350	399027	395976	412314
64	NZ_CP035214.1	NC_015968.1	5914592	4812833	20031	2406622	2426580	4291618	4311590	-	-
64	NZ_CP035214.1	NZ_CP020388.1	5914592	5408082	28563	2428090	2456422	1176073	1147748	-	-
64	NC_015968.1	NZ_CP020388.1	4812833	5408082	14784	4296848	4311604	1192305	1177552	-	-
73	MK422452.1	NZ_LT556084.1	34141	5048670	6253	16852	23087	2178196	2184423	2174695	2191206
92	NC_004313.1	NZ_CP030787.1	40149	4659718	7813	1	7781	1010446	1018233	994657	1036029
94	NC_021857.1	NZ_CP030787.1	41475	4659718	7899	1	7898	1010336	1018233	994657	1036029
96	NC_022749.1	NZ_CP030787.1	39758	4659718	6741	1	6731	1010394	1017125	994657	1036029
200	MF417956.1	NZ_CP035214.1	16570	5914592	3271	12052	15315	2383552	2380297	2371719	2390239
200	MF417956.1	NC_015968.1	16570	4812833	3068	12248	15315	4260624	4257557	4251287	4268785
200	MF417956.1	NZ_CP020388.1	16570	5408082	3062	12248	15309	403342	400281	395976	412314
200	NZ_CP035214.1	NC_015968.1	5914592	4812833	20031	2406622	2426580	4291618	4311590	-	-
200	NZ_CP035214.1	NZ_CP020388.1	5914592	5408082	28563	2428090	2456422	1176073	1147748	-	-
200	NC_015968.1	NZ_CP020388.1	4812833	5408082	14784	4296848	4311604	1192305	1177552	-	-
201	MF417957.1	NZ_CP035214.1	16537	5914592	3271	12023	15286	2383552	2380297	2371719	2390239
201	MF417957.1	NC_015968.1	16537	4812833	3068	12219	15286	4260624	4257557	4251287	4268785
201	MF417957.1	NZ_CP020388.1	16537	5408082	3062	12219	15280	403342	400281	395976	412314
201	NZ_CP035214.1	NC_015968.1	5914592	4812833	20031	2406622	2426580	4291618	4311590	-	-
201	NZ_CP035214.1	NZ_CP020388.1	5914592	5408082	28563	2428090	2456422	1176073	1147748	-	-
201	NC_015968.1	NZ_CP020388.1	4812833	5408082	14784	4296848	4311604	1192305	1177552	-	-
383	MK422452.1	NZ_CP037894.1	34141	4820913	5885	16876	22738	2126051	2131911	2107250	2161140
383	MK422452.1	NZ_CP032841.1	34141	4739272	5650	16876	22509	2793064	2787434	2770640	2816045
383	NZ_CP037894.1	NZ_CP032841.1	4820913	4739272	20013	492423	512365	4258623	4238667	-	-
385	MK416021.1	NZ_CP020448.2	38370	5244290	6051	7579	13591	1112570	1106562	1081621	1135661
385	MK416021.1	NZ_CP038469.1	38370	4840504	6053	7579	13591	1215983	1209974	1164030	1236449
385	NZ_CP020448.2	NZ_CP038469.1	5244290	4840504	50359	3935975	3986257	3801923	3852179	-	-
571	MF695815.1	NZ_LS992183.1	39241	5009078	5776	1	5766	4549175	4554940	4546666	4571702
571	MF695815.1	NZ_CP020820.1	39241	4023751	5752	1	5714	1257421	1263131	1234925	1263088
571	NZ_LS992183.1	NZ_CP020820.1	5009078	4023751	8731	490562	499274	457932	466652	-	-
573	MK416014.1	NZ_LS992183.1	29301	5009078	6419	16767	23166	4548538	4554940	4546666	4571702
573	MK416014.1	NZ_CP020820.1	29301	4023751	5853	17304	23114	1257324	1263131	1234925	1263088
573	NZ_LS992183.1	NZ_CP020820.1	5009078	4023751	8731	490562	499274	457932	466652	-	-
1385	NC_019489.1	NZ_CP021927.1	41726	1757573	5574	5658	11201	809268	814806	795677	836587

A.3 Figures

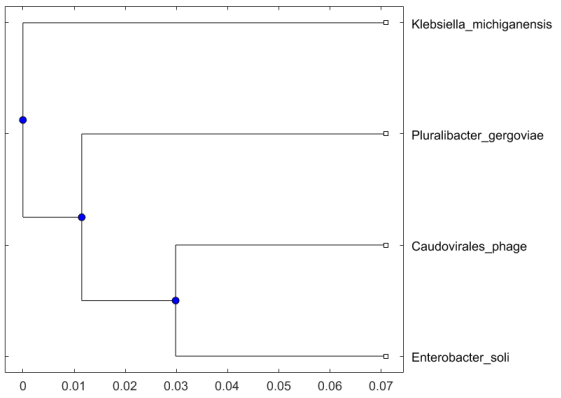


(a) Connection 44

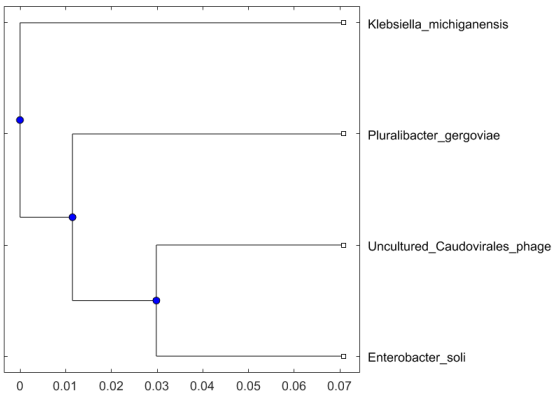


(b) Connection 64

Figure A.1: The phylogenetic trees of connection 44 and connection 64

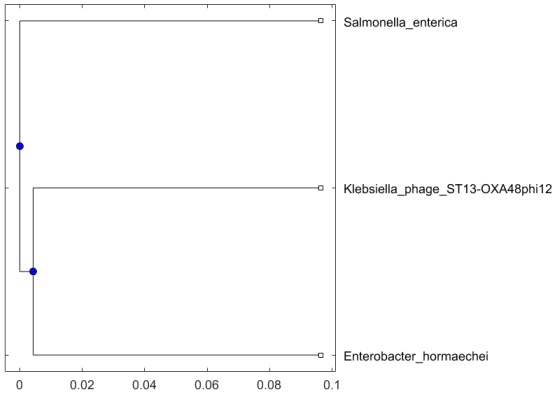


(a) Connection 200

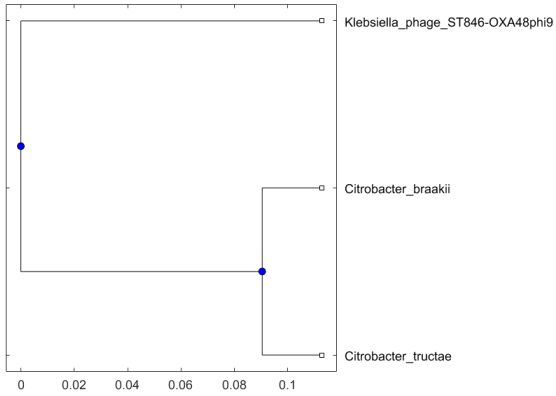


(b) Connection 201

Figure A.2: The phylogenetic trees of connection 200 and connection 201

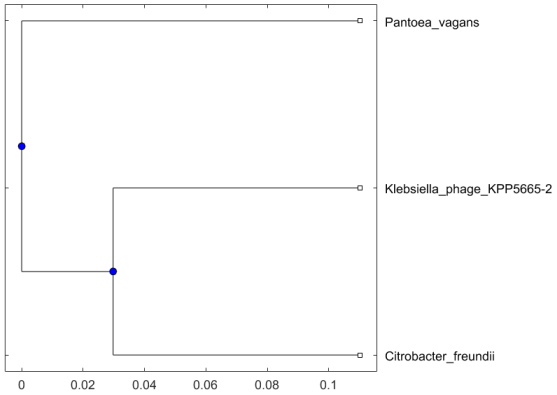


(a) Connection 383

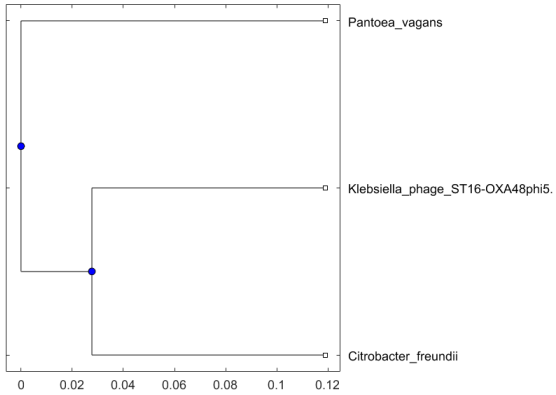


(b) Connection 385

Figure A.3: The phylogenetic trees of connection 383 and connection 385



(b) HGT connection 571



(b) HGT connection 573

Figure A.4: The phylogenetic trees of connection 571 and connection 573