Development of novel computational tools based on analysis of DNA compositional biases to identify and study the distribution of mobile genomic elements among bacteria

by

Keoagile Ignatius Oliver Bezuidt

Submitted in partial fulfillment of requirements for the degree Magister Scientiae

in the Faculty of Natural and Agricultural Sciences

Bioinformatics and Computational Biology Unit

Department of Biochemistry

University of Pretoria

Pretoria

September 2009

Declaration

I, Keoagile Ignatius Oliver Bezuidt, declare that the thesis/dissertation, which I hereby submit for the degree *Magister Scientiae* at the University of Pretoria has not been previously submitted by me for degree purposes at any other University and I take note that, if the thesis/dissertation is approved, I have to sumbit the additional copies, as stipulated by the relevant regulations, at least six weeks before the following graduation takes place and if I do not comply with the stipulations, the degree will not be conferred upon me.

SIGNATURE..... Date.....

Acknowledgements

- My supervisors Dr. Oleg Reva and Prof. Fourie Joubert for all their support and guidance during the course of my Msc.
- My parents and brothers for their support through my studies.
- All my fellow friends at the Bioinformatics and Computational Biology Unit for all their help.
- The National Bioinformatics Network for awarding me the bursary.

Contents

List of abbreviations

ACLAME	A Classification Of Mobile Genetic Elements
BLAST	Basic Local Alignment Search Tool
BLAT	Blast-Like Alignment Tool
CAI	Codon Adaptation Index
CDS	Coding Sequence
CGR	Chaos Game Representation
CSS	Cascading Style Sheet
D	Distance
GEI-DB	Genomic Islands Database
GI	Genomic Islands
GUI	Graphical user interface
HGT	Horizontal Gene Transfer
HTML	Hyper Text Markup Language
HMM	Hidden Markov Model
MCL	Markov Clustering Algorithm
MGE	Mobile Genomic Elements
NCBI	National Center For Biotechnology Information
NRDB-NCBI	Non-redundant Protein Database - National Center For Biotechnology Information
ORF	Open Reading Frame
OU	Oligonuclotide Usage
OUV	Oligonucleotide Usage Variance
PAI	Pathogenicity Islands
PAI-DB	Pathogenicity Islands Database
PS	Pattern Skew

RV	Relative Variance
RSCU	Relative Synonymous Codon Usage
IS elements	Insertion sequence elements
SPE	Streptococcal Pyrogenic Exotoxins
SPI	*Salmonella* pathogenicity island
SWGIS	SeqWord Gene Island Sniffer
SWGB	SeqWord Genome Browser

List of Figures

List of Tables

Chapter 1

Introduction

The advent of comparative genomics of hundreds of sequencing projects has revealed the real dimensions about the contribution of horizontal gene transfer as one of the dominant factors in the evolution of prokaryotic gene order. Further developments in sequencing are set to massively increase the availability of genomic sequences of environmental and medical importance. To date, over 200 fully sequenced prokaryotic genomes are available at the National Center for Biotechnology Information (NCBI) (Wheeler *et al.*, 2002). Their analysis has revealed that mobile genomic elements are prominent in bacterial evolution and adaptation to all sorts of environmental pressures mediated by the transfer of genetic material among fixed lineages and even across species borders.

Two strains of bacteria of the same species can differ by as much as 30% of the accessory parts of their genomes (Sueoka, 1962). These differences mostly result from mechanisms such as: insertions, deletions, transpositions, duplications, recombinations and rearrangements of residues of mobile DNA sequences. Various comparative and statistically driven *in silico* methods have been developed trying to decode the rearranged genomic structures as well as the characteristics of gene flow among differ-

ent species. Sequence data were found to display wide variations in their nucleotide compositions across bacterial species, as a result of an evolutionary factor that infers genome plasticity, known as horizontal gene transfer (Hacker *et al.*, 2003*b*).

1.1 Horizontal gene transfer

It is increasingly becoming apparent that genetic material within single and multi-celled organisms have been acquired by horizontal gene transfer since the early stages of life. The exchange of genetic material was found to have occurred in different domains of life: Archaea, Bacteria, and Eukarya (Choi and Kim, 2007). Horizontal gene transfer, defined as a mechanism that promotes the transfer of foreign genomic segments between lineages was found to be relatively common in prokaryotes and less common in higher-order organisms. This mechanism effectively contributes to the evolution and diversity of bacterial species by the transfer of novel genomic segments to parts of their genomes, but not all genomic segments undergo horizontal transfer. The preferential transfer of genes in species is strongly correlated with gene function (Jain *et al.*, 1999). Informational genes, defined as the core and most conserved segments in a genome are present in almost all organisms. Such genes encode rRNA operons and conserved proteins, they therefore are less likely transferred, as genomes naïve to their functions are rare (Ochman, 2001; Dutta and Pan, 2002). Operational genes, also defined as the accessory parts in a genome are most likely horizontally transferred as compared to informational genes (Rivera *et al.*, 1998; Jain *et al.*, 1999). The transfer of operational genes is a continual process and is far more important in prokaryotic diversity of different sources (Jain *et al.*, 1999; Ochman, 2001). For horizontal gene transfer to become a success, the acquisition of foreign DNA segments must be counterbalanced by DNA loss (Lawrence, 1999). Acquired DNA providing functions that are beneficial to the host may be maintained, while DNA providing less beneficial functions may be lost (Lawrence, 1999).

Mobile genetic elements possess genes that contribute not only to bacterial speciation and adaptation to different niches, but also carry with them factors that contribute to the bacteria's fitness traits, secondary metabolism, antibiotic resistance and symbiotic interactions (Hsiao *et al.*, 2003*a*; Dobrindt *et al.*, 2004; Mantri and Williams, 2004) that are of medical and agricultural importance. Collectively, the latter factors form part of a "gene organization" known as the flexible gene pools. The flexible gene pools are named according to the types of functions they encode, those that encode virulence features are designated pathogenicity islands (PAI).

PAI were first identified in human pathogenic strains of *E. coli*, the acquisition of genes of their sort have been shown to possess the ability to confer a virulence phenotype upon a normally avirulent strain (Ochman, 2001). PAIs are highly variable mobile DNA segments present only in one or more strains of a given species. These segments can transfer between environmental microorganisms, across species and even genus boundaries and influence virulence. Genomic elements similar to pathogenicity islands by general composition and organization are subsequently identified in non-pathogenic bacterial species, and termed genomic islands (Hacker and Carniel, 2001; Dobrindt *et al.*, 2004). Genomic islands are multigene chromosomal subunits that confer bacterial multifunctional traits and are evident of horizontal transfer.

The transfer of such subunits occurss through three mechanisms (Figure 1.1): (a) transformation, (b) conjugation and (c) transduction. These mechanisms mediate the movement and transfer of DNA segments intercellularly. Conjugation and transduction are the common players in genetic transfer, they require mobile elements such as plasmids and bacteriophages to transfer genetic elements along with the sequence features of their donor to recipient cells (Hacker and Carniel, 2001). Upon transfer, these genetic elements get established into the recipient cell either as self replicating elements such as plasmids or by getting intregrated into the chromosome (Dutta and Pan, 2002; **?**) either by homologous or illegitimate recombination techniques (Beiko

et al., 2005). Transformation, unlike conjugation and transduction does not require any form of a vector to transport genomic elements between bacteria, it is a mechanism that is mediated by the uptake of a naked DNA in the environment. The uptake usually takes place upon the release of DNA from decomposing and disrupted cell, or viral particles, or even excretions from living cells (Thomas and Nielsen, 2005).

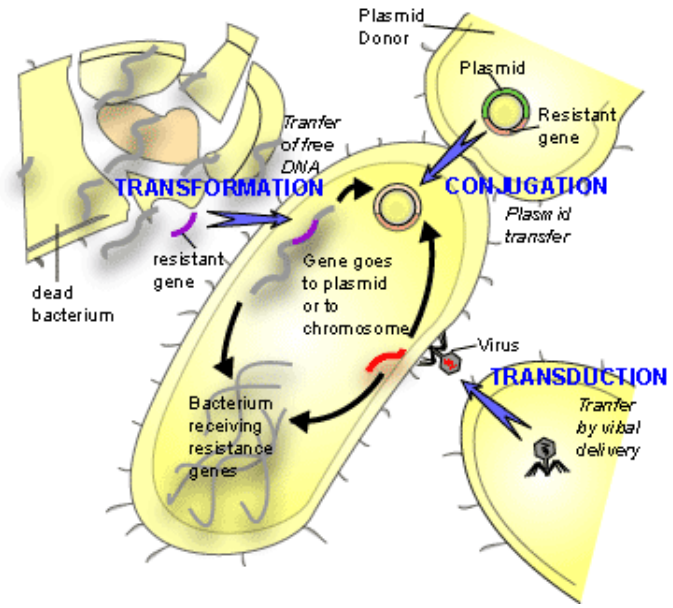


Figure 1.1: The above figure denotes three mechanisms involved in the transfer of genetic material among bacterial species through an event of horizontal transfer. The first mechanism on the figure is: (a) Transformation: the transfer of genetic material that is mediated by an uptake of free DNA, followed by (b) Conjugation: transfer of genetic material that is mediated by plasmids, the mechanism requires cell to cell contact. (c) Transduction: the transfer of genetic material that is mediated by bacteriophages (viruses that infect bacteria). Adapted from `www.textbookofbacteriology.net`.

DNA composition comparisons between lineages have uncovered that genes acquired by the above mechanisms display features that are distinct from those of their recipient genomes (Hacker and Carniel, 2001; van Passel *et al.*, 2005). Genes acquired by horizontal transfer can often display atypical sequence characteristics and a restricted phylogenetic distribution among related strains, thereby producing a scattered phy-

logenetic distribution (Ochman *et al.*, 2000; Dutta and Pan, 2002). Bacterial species are variable in their overall GC content but the genes in genomes of particular species are fairly uniform with respect to their base composition patterns of codon usage and frequencies of oligonucleotides (Sueoka, 1962; Ochman *et al.*, 2000; Hsiao *et al.*, 2003*a*). The phylogenetic aspect of similarity in base composition among closely related species arises from their common origin (Sueoka, 1962). Similarity is also influenced by species specific mutational pressures that act upon their genes to promote the maintenance of composition stability. The similarity of these compositions is conserved within and among lineages, as native/core genes in a given organism exhibit homogeneous G+C content and codon usage, while foreign genes display atypical characteristic features that resemble horizontal transfer. Exogenous genes display characteristic features such as: low GC content, unusual codon usage, atypical amino acid usage, tRNA site integration, insertion sequences and direct repeats at the flanks indicating acquisition from a foreign source (van Passel, Bart, Thygesen, Luyf, van Kampen and van der Ende, 2005). Chromosomal regions invaded by genomic islands exhibit typical flanks of direct repeats, at most this features are observed at the 3' end of tRNA as they are mostly favored by genomic islands as chromosomal integration sites (Figure 1.2) (Dutta and Pan, 2002). These regions also contain transposase or integrase genes that are required for chromosomal excision and integration (Auchtung *et al.*, 2005; Klockgether *et al.*, 2007), respectively.

1.2 Types of genomic islands and their encoded functions

1.2.1 PAI-pathogenicity islands

Pathogenicity islands were first identified in uropathogenic *E. coli* strains as distinct chromosomal regions in possession of genes encoding virulence factors. They

showed to differ from the core genome based on their atypical compositional features, such as low or high GC content, frequent associations with tRNA genes, and repeat sequences at the flanks. Apart from their distinct atypical compositional features, PAIs possess virulence features such as adherence factors, iron uptake systems, Toxins, Types I - VI secretion systems, antiphagocytotic determinants (Spanier and Cleary, 1980) and also mobility factors such as integrases, transposases, and phage genes. Upon the transfer of pathogenic bacteria into the host, the latter factors get encoded to enable bacteria to undergo several host-cell infection cycles. Such factors enable bacteria to adhere to host surfaces, get protection against immune cells, and produce toxins. PAIs are often located within or adjacent to chromosomal tRNA regions, widely known to serve as insertion hot spots of certain bacteriophages and plasmids. Both plasmids and bacteriophages play the most crucial role in mobilizing virulent cassettes across species boundaries, thus promoting microbial evolution and the development of novel pathogenic strains. Bacteriophages are the most abundant microorganisms and their diversity exceeds that of prokaryotes with >10 folds (Lima-Mendez *et al.*, 2008*a*; Williamson *et al.*, 2008), they are therefore considered to be ones that mainly convert bacteria into pathogens through a process of lysogenic conversion. Plasmids also carry virulent genes. Thus, bacteriophages take part in the evolution of microbial pathogenicity as much as virulence plasmids. Analysis conducted on virulence encoding bacterial sequences revealed that most of these genes are associated with temperate phages, as the majority of toxin genes were found to be phage encoded. Experimental measures conducted on *Streptococcus pyogenes CS112* indicate that the streptococcal pyrogenic exotoxins A and C (SPEs) which they carry are phage associated, as they were found to be located adjacent to phage insertion sites (Betley and Mekalanos, 1985; Johnson *et al.*, 1986; Goshorn and Schlievert, 1989). More analysis conducted on *E. coli 0157:H7* also revealed that the shiga toxins (stx1 and stx2) that they harbour are also phage related (O'Brien *et al.*, 1989).

1.2.2 Symbiosis Islands

Bacteria have different modes of establishing beneficial relationships with their host organisms, particularly multi-cellular organisms. These relationships mainly result from transmission of pathogenicity and symbiosis islands among bacterial species through horizontal transfer. Symbiosis and pathogenicity islands share similar structural properties and they both use similar mechanisms for manipulating their multicellular hosts, but unlike pathogenicity islands, symbiosis islands do not cause infections nor cause tissue damage to their hosts instead they code for housekeeping functions and establish mutualism (Uchiumi *et al.*, 2004).

The most common cases of symbiosis occur between bacteria and plants, and the best example of this has been illustrated by studying the relationship between Mesorhizobia and legumes. *Mesorhizhobium loti* are well known as symbiotic nitrogen-fixing soil bacteria, that use leguminous plants as their hosts. Plants require nitrogen as a measure of alternating their metabolic pathways, but they cannot directly use the nitrogen from the atmosphere to initiate such processes, instead they depend on microbes such as rhizobia to fix it for them so it can be usable. Rhizobia carry chromosomally integrated nitrogen fixation islands that are often located within or near tRNA genes, and the major functional trait which they possess is the conversion of dinitrogen gas (N_2) to ammonia (NH_3) that benefit their plant hosts (Uchiumi *et al.*, 2004) and in turn rhizobia get high energy carbohydrates from the plant leaves derived from photosynthesis. Rhizhobia reside and fix nitrogen within the plants roots nodules which they form upon invasion. The latter resemble evolutionary mechanisms that shape up host-bacteria interactions, and the adaptation of bacteria to different host environments. As observed, rhizobia seem to be leading a two way life-style, for they tend to have factors that enable them to survive both in the soil and plants. Life in the soil allows them frequent tranfer and acquisition of genes to and from other bacteria in the same niche, which thus further shape up their genomic

makeup (Sullivan *et al.*, 2002; Uchiumi *et al.*, 2004).

1.2.3 Antibiotic resistance islands

Multitudes of disease-causing bacteria develop resistance to their prefferred antibiotic treatment through different mechanisms. They can either develop resistance by random mutation, transformation or transduction, but the most common method through which bacteria acquire drug resistance gene cassettes is conjugation. The acquisition of these genomic fragments enables bacteria to survive and replicate in the presence of antibiotic doses by encoding genes that render them resistant. Antibiotic resistance genes are often carried by transposable elements which are frequently located in plasmids, that thus promote their dissemination across species borders. Apart from transposons, integrases serve as the other type of mobile genetic element that carry resistance genes with them aiding the spread of resistance. Several bacteria have developed resistance to agents such as tetracyclines. Tetracyclines are a family of broad spectrum antibiotics that previously were effective in inhibiting protein synthesis in a wide range of microbes until the emergence of superbugs. Superbugs are refferred to as pathogenic bacteria that carry different kinds of resistance genes in their genome. Most of the tetracycline resistance genes are identified in resistance plasmids, making horizontal transfer the likely method of their transfer (Hartman *et al.*, 2003; Pezzella *et al.*, 2004). Several gene types that confer resistance to the latter antibiotics have been identified, designated as *tet* of classes A to G that encode tetracycline resistance (Pezzella *et al.*, 2004). Tetracycline resistance gene: tet(A) was found to be associated with transponson Tn1721 carried by plasmid pGFT1 of *Salmonella enteric subsp. enterica serovar* (Frech and Schwarz, 1998). Tet(A)-1 an alelle of tet(A) (Hartman *et al.*, 2003) was identified in both *Salmonella* spp and entero-invasive *E. coli* carried by plasmid pSSTA-1. Several other tet genes, designated tet(L), tet(H) and tet(O) were identified in *Actinobacillus* by Blanco *et al.* (2006) and were also found to be associated with plasmids p11745 and p9555.

1.2.4 Catabolic genomic islands

Bacteria have different modes of survival within and throughout the outskirts of different environments with various conditions. They have been found to possess entities that enable them to degrade xenobiotic chemicals that are harmful to living organisms. Amazingly, bacteria evolve to develop strategies to detoxify such harmful substances and utilize their by-products as sources for growth and supplemental energy. These entities occur in a variety of organisms, most especially the ones that inhabit continually polluted areas such as aquatic sediments and soil. Molecular analysis of the catabolic pathways has indicated that such organisms may have adapted to organic pollutants by expressing new functions to resist their toxic effects to use them for their benefit (Top and Springael, 2003). Such genes have frequently been found to be located within DNA fragments that are termed catabolic genomic islands, and to possess structural features of transposable elements, suggesting that they could be disseminated across environmental organisms through horizontal transfer events. Several transposons such as Tn5707, Tn4371, Tn1721 involved in catabolic pathways have been identified in *Pseudomonas*, *Ralstonia, E. coli* and *Alcaligenes* and were found to be associated with plasmids, RP4, pP51 and pENH91 (Ogawa and Miyashita, 1995; Merlin *et al.*, 1999). Most of these elements have been noted to integrate in chromosomes upon their acquisition. Their characteristic analysis also revealed that most of the catabolic genes are associated with tRNAs and that they encode phage-related integrases (Butler *et al.*, 2007).

1.3 Features and detections of GI

The identification of genomic islands falls mainly on the basis of compositional features that distinguish them from native genes in the genome (Daubin *et al.*, 2003).

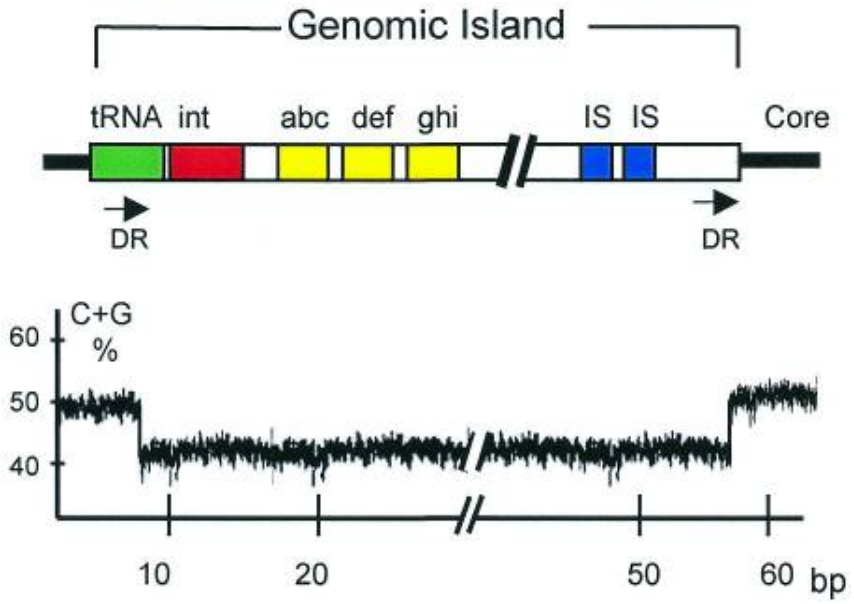


Figure 1.2: Schematic model of a genomic island of bacteria. This figure displays the characteristics of genomic islands and how they differ with the composition their host. Genomic islands often display low G+C % (guanine cytosine content of the genomic island) compared to that of the host genome. They are often inserted at the tRNA site and are also compounded by direct repeats at the flanks. Abbreviations in the diagram: DR, direct repeats; int, integrase gene; abc, def, ghi, are the genes encoding specific functions; IS, insertion sequence elements; bp, base pair Adapted from Hacker and Carniel, 2001.

Most of the previously published methods for the detection of genomic islands search for elements that possess deviant compositional features (Philippe and Douady, 2003). The deviant composition method is most plausible since it does not involve phylogenetic approaches and DNA comparison between multiple species to detect tree incongruencies and abnormal sequence similarities. At most, genes that appear atypical in their current genomic context are suspected of having been introduced from a foreign source (Lawrence and Ochman, 2002). The assumption is that directional mutation pressures within bacterial genomes impart distinctive biases to the composition of native genes in the genome in a way that recently acquired genes will appear atypical by comparison if they have evolved in a genome with different mutational biases (Lawrence and Ochman, 2002). During introgression the horizontally

transferred genes display characteristic features of the donor genome but over time they start reflecting the base composition of their new host, for newly acquired genes get affected by the same directional mutation pressure as all the other genes in the recipient genome (Lawrence and Ochman, 1997; Dutta and Pan, 2002).

1.3.1 Codon bias and AT/GC content

Horizontally acquired genes are mainly A+T-rich and displaced in a similar direction, towards the AT-rich codons (Daubin *et al.*, 2003). It has been stipulated that genes that are susceptible to horizontal transfer use similar patterns of codons, patterns that are typical to their donor genome (Ermolaeva, 2001), also in that the density of their codons GC compositions deviate from the pattern used by their recipient genome. The differences were found to be visible in each codon position, but mainly the third codon position since it is the most likely to change synonymously (Lawrence and Ochman, 1997; Daubin *et al.*, 2003). Ranjan *et al.* (2007) have shown that codons with dinucleotide patterns such as: AA, AT, AG, TA and TC at their first two positions are more abundant in GC poor genomes. Those with bases such as: GG, GC, CT, CG and CC at the first two positions are more abundant in G+C rich genomes (Rajan *et al.*, 2007). Moreover, Daubin *et al.* (2003) examined the base composition and codon usage in genes unique to genomes from several bacterial species and found that genes believed to be recently acquired have a relatively low GC content and atypical codon usage patterns when compared with surrounding genes, even in AT-rich genomes (Charkowski, 2004). Sharp and Li (1987) proposed a method to calculate the codon adaptation index (CAI) for each gene in a genome as a measure of similarity of a genes synonymous codon usage to that of a standard set of highly expressed gene's in a genome (Koski *et al.*, 2001) based on an organism's base composition. The CAI is a indicator originally designed to measure the weight of each codon from its frequency within a pool of highly expressed genes (Sharp and Li, 1987) and to measure the dominating codon bias (Carbone *et al.*, 2003). It can

also be used to check for genes with a codon usage that is distinct from native genes in a given genome. The CAI (Sharp and Li, 1987) is calculated as follows:

$\text{RSCU}_{ij} = X_{ij} \; / \; \frac{1}{n_i} \sum_{j=1}^{ni} X_{ij}$ [i]

$\text{W}_{ij} = \text{RSCU}_{ij} \; / \; \text{RSCU}_{imax} = X_{ij} / \; X_{imax}$ [ii]

where X_{ij} denotes the number of occurrences of the jth codon in the ith amino acid of a given protein, and n_i denotes the number of occurrences of alternative codons for the ith amino acid. W_{ij} denotes the frequency of use of the jth codon in a reference set of highly expressed genes (Ermolaeva, 2001), whereas RSCU_{imax} and X_{max} represent the codon most frequently used in the ith amino acid. Subsequently, the determination for CAIs of all the coding regions in the genome are computed by the following equation: $\text{CAI} = \sum_{i=1}^{64}[codon\_freqi * ln(W_i)] / \sum_{i=1}^{64} codon\_freqi$, where $codon\_freqi$ denotes the relative usage of codon i in the given codon sequence, and Wi the ratio of the occurence of codon i relative to the occurences of codons of highly expressed genes. Davids and Zhang (2008) effectively modified the CAI method and implemented it in determining the differences in gene expression levels of horizontally transferred genomic islands in accordance to core (shared by all *E. coli* strains) and non-core (present in one strain and not all others) genes of different *E. coli* strains. Their analysis illustrated that core genes, though they evolve slowly, have higher gene expression levels and an increased codon adaption index as compared to non-core and highly evolving HGT genes (Figure 1.3). This also explains why conserved genes use optimized codon usage patterns while putative or horizontally transferred genes do not.

1.3.2 Oligonucleotides

In 1995 (Karlin and Burge, 1995), the concept of detecting horizontally transferred genes by calculating oligonucleotide frequencies was introduced. This approach uses

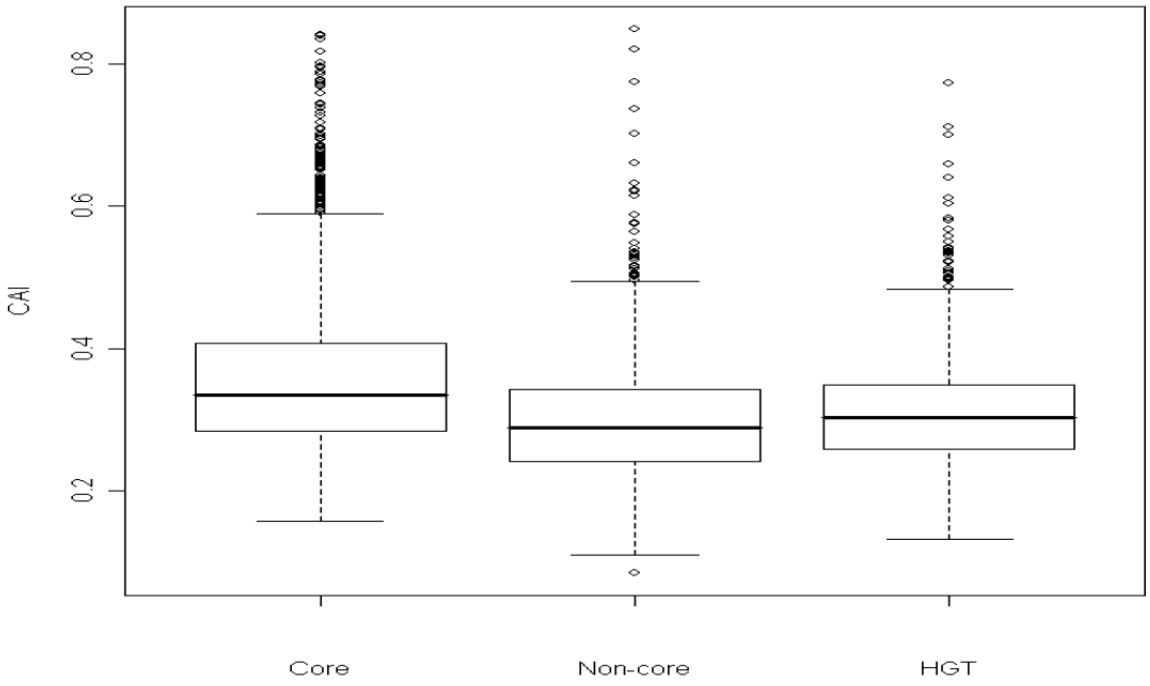


Figure 1.3: The figure illustrates distributions of conserved core and non-core genes that are shared by all *E. coli* strains. Adapted from Davids and Zhang, 2008.

extensive statistical parameters to determine genomic segments that display siginificant differences in oligonucleotide usage patterns compared to the rest of the genome (Karlin and Burge, 1995). It detects horizontal transfer events using global sequence patterns rather than using individual genes, and it also does not require DNA similarity analysis or phylogenetic distributions (Karlin *et al.*, 1997). Oligonucleotides are simply defined as chains of overlapping short words of the same or different lengths. Patterns of frequencies of oligonucleotides in genomes are not random (Reva and Tummler, 2004) and can be used to reveal different properties of DNA (Bohlin *et al.*, 2008). The occurrences of these patterns are an influence of DNA structural properties such as base stacking energy, propeller twist angle, protein deformability, bendability, position preference or repair mechanisms (Baldi and Baisnee, 2000). On the other hand they could be a result of correlations of codon usage and environmental pressures exerted on the genome. Oligonucleotide patterns were examined for

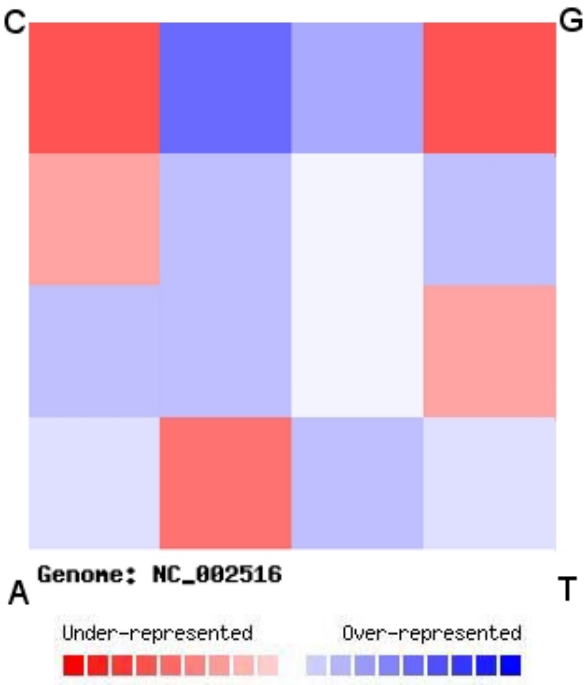


Figure 1.4: The visual representaion of the over and under-representation of dinucleotides usage patterns in *Pseudomonas aeruginosa* genome sequence. Dinucleotides that are over-represented in the genome are represented by blue squares while the under-represented ones are represented by red squares, and the dinucleotides that are neither over-represented nor under-represented are represented by the white blocks. Adapted from http://insilico.ehu.es/oligoweb/info/CGR.php?#CGR.

various organisms and were shown to be species specific (Deschavanne *et al.*, 1999). Horizontally transferred genomic elements display the oligonucleotide characteristics of their source. The screening of local variations of usages of words along genomes is expected to detect the regions of interest where HGT might be located (Deschavanne *et al.*, 1999).

A signature pattern of DNA nearest neighbour also known as dinucleotides was the most studied right after the introduction of the oligonucleotide usages concept (Karlin *et al.*, 1994; Karlin and Burge, 1995). Karlin and Burge (1995) studied distributions of dinucleotides by establishing the statistical formula: $\rho_{xy}= f_{xy}/f_x f_y$ of dinucleotide abundance values, where f_x denotes the frequency of the mononucleotide

X and f_{xy} denotes the frequency of the dinucleotide XY. By using the formulae ($\rho_{xy}= f_{xy}/f_x f_y$), it was observed that frequencies of dinucleotide compositions were uniform across the entire genome, and were regarded as a stable property of DNA of a given organism (Karlin *et al.*, 1994; Srividhya *et al.*, 2007), also that organisms that are of the same genera share a similar pattern of dinucleotides than do organisms that are not (Karlin, 1998). Karlin *et al.* (1997) also proposed the following formula: $\delta*(f,g)= \frac{1}{16}\Sigma|\rho*xy(f)-\rho*xy(g)|$, which is used to measure the average difference δ between oligonucleotide patterns of different subsets f and g in a single genome. The latter formula determines the the average absolute difference of the dinucleotide relative abundance values as follows: it calculates the dinucleotide abundance values of fragments f and g by using ($\rho_{xy}= f_{xy}/f_x f_y$) denoted as $\rho*xy(f)$ and $\rho*xy(g)$ in the relative abundance difference formula. Upon their determination, the formula $\delta*(f,g)= \frac{1}{16}\Sigma|\rho*xy(f)-\rho*xy(g)|$ is then used as a measure of genomic signature differences between two fragments of the same genome, to characterize the signature differences between native and horizontally acquired genes. This method is applicable to the detection of exogenous genomic elements as it allows discrimination among sequences coming from different organisms (Karlin, 2001). Genomic segments that exhibit significant differences in dinucleotide patterns compared to the rest of the genome are more likely to contain horizontally transferred elements (Karlin and Burge, 1995).

Deschavanne *et al.* (1999) took the Oligonucleotides concept further by using Chaos Game Representation (CGR) (Jeffrey, 1990) to represent frequencies of DNA signature patterns. The CGR concept was first introduced by Jeffrey (1990), it is a concept that was originally coined to reveal an underlying structure in the sequence of random numbers in a form of pictures known as *attractors*. The whole idea was to use CGR to visualize the underlying structures on DNA sequences from four letters 'a', 'c', 'g', 't or u' instead of just using numbers. Jeffery used CGR in the form of a

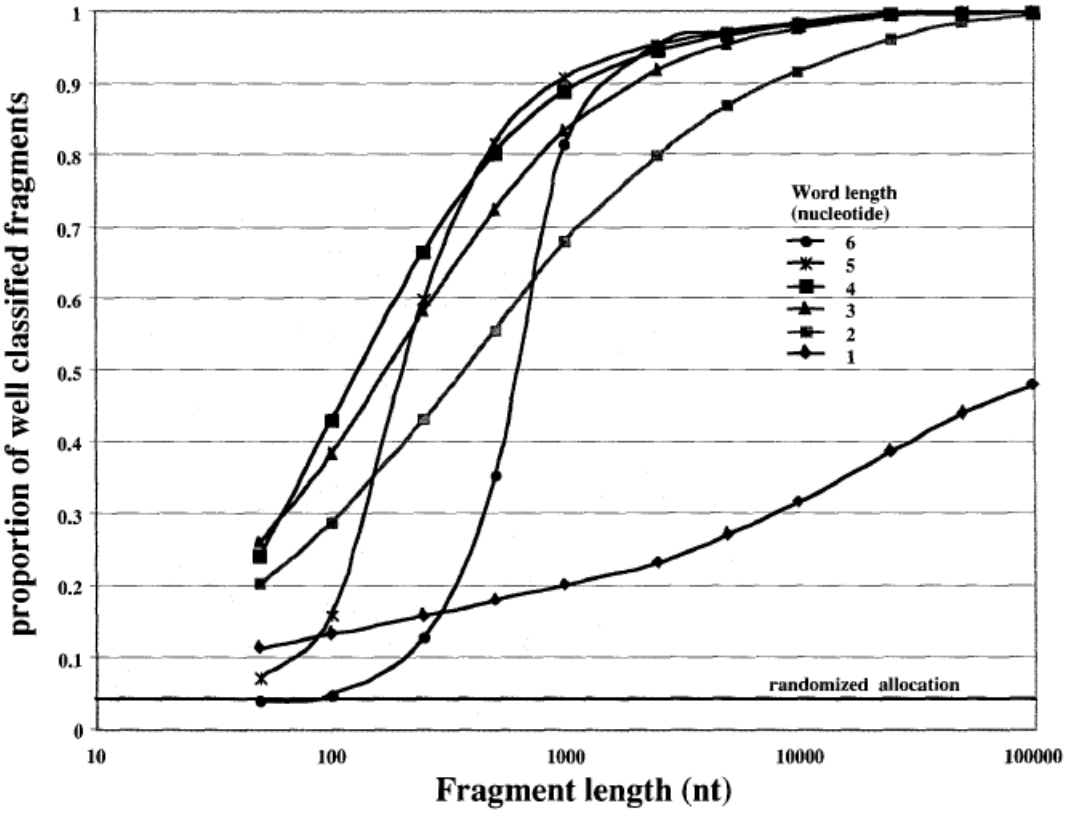


Figure 1.5: Distributions of oligomers of different lengths. Adapted from Deschavanne *et al.*, 2000.

square and labelled each corner of every vertice with bases 'a', 'c', 'g', 't'. Then he made an example on how the concept works by plotting the first 6 bases: 'gaattc' of human beta globin region, chromosome 1. The procedure was described as follows: the first base 'g' gets plotted halfway between the center and vertex 'g', second base 'a' gets plotted halfway between the previous plotted point 'g' and vertex 'a', third point 'a' gets plotted halfway between previously plotted point 'a' and vertex 'a', the last three bases 'ttc' also follow the same plotting procedure. Deschavanne bettered the CGR algorithm in a way that it could be used to represent DNA signature patterns in the form of fractal images (Figure 1.4), where every block in the image corresponds to the frequency of a specific word (Oligonucleotide) (Deschavanne *et al.*, 1999). The generated images are divided into four quadrants where each gets subse-

quently divided into four subquadrants, each containing a unique sequence pattern. Oligonucleotide frequencies are displayed by the intensity of each pixel, the darker the pixel the higher the frequency of a pattern.

1.4 Oligonucleotides as phylogenetic signals

It was thought that dinucleotide frequencies could successfully be used to detect horizontal transfer events, until Pride *et al.* (2003) and Dufraigne *et al.* (2005) found that base composition of words that are 1 and 3 bases long are poorly species specific, and do not allow a good discrimination between species since they are just an influence of codon distributions. Longer words may be more species specific even though their frequencies in genomes may appear to be more variable, and they can also be used to discriminate among closely related species (Deschavanne *et al.*, 1999). Deschavanne *et al.* (2000) classified genomic signatures of different lengths (Figure 1.5), indicating that longer signatures improve classification especially in longer DNA fragments, and it appeared that tetranucleotides are the best classifiers as compared to penta and heptanucletoides, even when classified by naive Bayesian methods using DNA fragments as short as 400 bases (Sandberg *et al.*, 2001). Following an increase in number of genome sequencing projects the latter approaches have been widely used in the classifications of metagenomic samples obtained from diverse microbial communities, for tetranucleotides have a high discriminatory power in metagenomes with low community diversity (Teeling *et al.*, 2004*b*,*a*; Willner *et al.*, 2009). Essentially, tasks that are addressed in the metagenomic data analysis are the assembling of short gene fragments and predicting their taxonomic origins which aid in gene function identification and reconstruction of microbial compositions (Diaz *et al.*, 2009). Initially, environmental gene fragments were classified by the use of highly conserved ribosomal RNAs as they were found to infer phylogenetic relationships between different microbial datasets, and could thus be used to identify source

organisms for given sets of gene fragments. Although the latter approach allows the most accurate classifications of organisms, their phylogeny is based on single genetic elements (Pride *et al.*, 2003) and can therefore only characterize a limited number of fragments (McHardy and Rigoutsos, 2007). Recently the latter approach has been complemented by developments of similarity and composition based type methods for classifications and binning of genomic fragments. The similarity based methods, such as CARMA (Krause *et al.*, 2008) and MEGAN (Huson *et al.*, 2007, 2009) perform classifications by comparing reads with reference databases of known sequences using BLAST (Altschul *et al.*, 1990), fragments are grouped into taxa based on the hits obtained from the search, although the latter approaches perform well their only downfall is that they can only classify reads into taxa if they have their closely related reference genomes available in the database. Currently, the methods that perform reliably are those that classify genomic fragments on the basis of their composition characteristics (Teeling *et al.*, 2004*b*), because gene compositions have long been shown to carry phylogenetic signals (Karlin and Burge, 1995; Deschavanne *et al.*, 1999; Pride *et al.*, 2003). The application of tetranucleotides in assigning genomic fragments to their taxonomic groups is the most favored because of the discriminatory power they offer between species, and can also be used as a tool to measure the extent of horizontal transfer that occurs between environmental bacteria (Tamames and Moya, 2008), as well as try to identify some of their potential donors (Sandberg *et al.*, 2001).

1.5 GI online resources

Several online resources have been put together now that there is enough genomic data that is publicly available. These resources pursue the fascinating goal of quantitatively analyzing the amount of genomic islands in prokaryotes and classifying them according to their horizontal transfer events. The resources that are about to

be discussed in this section use some of the above mentioned methods and features together with statistical formalization to detect horizontal gene transfer events. The resources are described in the following subsections.

1.5.1 Islandpath. Aiding genomic island identification

Islandpath (`http://www.pathogenomics.sfu.ca/islandpath/`) is a web-resource that allows the prediction of genomic island-associated features in a full genome context (Hsiao *et al.*, 2003*b*) by searching for genomic segments that possess distinct compositional features such as GC content and dinucleotide bias. Islandpath uses a single open reading frame as a basic unit for the calculation of GC percentage, which permits the analysis of gene-by-gene variance in proposed coding sequences by allowing the detection of genes with GC frequency that differs significantly from the average genome GC frequency. The resource also implements the dinucleotide abundance formulae previously developed and published by Karlin (1994) to aid with the detection of horizontaly transferred elements. Hsiao *et al.* (2003*b*) use the formulae in a way that genes of potential horizontal transfer are detected by calculating their average relative abundance differences $\delta*(f,g)$ for each ORF-cluster (6-ORF's) in the genome where f-fragment denotes ORF-cluster sequences and g-genome denotes all ORF's in the genome (Hsiao *et al.*, 2003*b*). Equation $\rho_{xy}= f_{xy}/f_x f_y$ was used for the dinucleotide relative abundances ρ_{xy}, where f_x denotes the frequency of mono-nucleotide x and f_x the frequency of dinucleotide xy in each of the ORF clusters or the genome (Hsiao *et al.*, 2003*a*). The mean $\delta*(f,g)$ was determined by averaging the results obtained from all the clusters in the genome, and regions that exhibit a mean $\delta*(f,g)$ greater than 1 standard deviation are marked as exogenous.

1.5.2 PAI-DB. Pathogenicity islands database

PAIDB (`http://www.gem.re.kr/paidb/`) contains the comprehensive information of all reported and potential PAI (pathogenicity islands) regions in prokaryotic genomes. It uses a homology-based method together with DNA composition anomalies to detect pathogenicity islands in bacterial genomes. The detection is performed using known PAI loci that were collected from GenBank and literature. This collection of known PAI loci were similarity searched against prokaryotic genomes using BLAT (Kent, 2002) and BLASTP (Altschul *et al.*, 1990) for genomic strips of PAI associated genes the identified overlapping strips were merged into single genomic regions. The merged regions were considered homogenous only if they exhibited over 80% and 25% in DNA and protein sequence similarity, respectively. The regions were considered pathogenic only if they were in possession of four or more virulence genes and also if they contained genes coding for transposases, IS element and integrases (Yoon *et al.*, 2005).

1.5.3 ACLAME. A Classification of mobile genetic elements

Aclame (`http://aclame.ulb.ac.be/`) is a comprehensive web resource that aids with the classification and annotation of proteins encoded by mobile genomic elements (MGEs) (Leplae *et al.*, 2004). The database is a collection of protein families obtained from bacteriophages and plasmids. The proteins were clustered into families according to functional parameters they have in common by using TRIBE-MCL, a graph theory based Markov clustering algorithm. The clusters were then used to search other public databases for related sequences to also allow annotations of MGE proteins whose functions are not known. Further similarity searches were performed with PSI-BLAST (Altschul *et al.*, 1997) and HMM (Enright *et al.*, 2002) against databases such as Swissprot (Boeckmann *et al.*, 2003), Scop (Conte *et al.*, 2002) and NRDB-NCBI (Benson *et al.*, 2003). The development of the database was led by

difficulties in the systematic analysis of mobile genomic islands and lack of MGE in existing databases that results in difficulties of identifying relationships between genomic elements harboured in widely different hosts.

1.5.4 HGT-DB. Horizontal gene transfer database

HGT-DB (`http://genomes.urv.cat/HGT-DB/`) is a composition-based web resource that provides pre-calculated averages and standard deviations for: GC content, codon usage, relative synonymous codon usage and amino acid content of bacterial and archaeal complete genomes. It also provides lists of putative genomic islands, correspondence analyses of codon usage and lists of extraneous genes in terms of their GC contents (Garcia-Vallve *et al.*, 2003). It uses a set of statistical approaches to determine the genes that deviate from the mean GC and/or average codon usage of the genome. Genes are marked as exogenous if their GC compositional features deviate by 1.5 standard deviation from the mean of the whole genome. Mahalanobis distance method was used to determine the correlations between variables by which different patterns can be identified and analyzed (Garcia-Vallve *et al.*, 2003). The latter approach was applied to measure the distances between codon usages of a given gene and the mean of genome to determine whether genes in a particular genome use the same codon pattern. Genes that are not in possession of the same codon usage pattern as their host genome are depicted by inclined distance values.

Problem Statement

The study of mobile genetic elements (MGE) distributions in bacterial communities as well as tracing their evolutionary origins has always been a great challenge in computational genomics. These MGE's are recognized as atypical genomic entities in prokaryotic genomes that influence the dissemination of genes that contribute to the

evolution of bacteria, thus imposing them to be of environmental and medical importance. Virulence-associated genomic elements have initially been detected in human pathogenic microorganisms. Apart from pathogenicity, MGE may confer other traits such as: metabolic versatility, adaptability and, symbiosity that allow commensals the abilities to inhabit variable niches. Current techniques for the identification of horizontal transfer events suffer from lack of precise predictions of borders of MGE inserts within genomes. Hence, the ignorance of the latter conceal from us the functional and evolutionary importance of environmental bacterial populations as more and more variants of species begin to emerge.

Specific Aims

The focus of this project is aimed at studying the rates of HGT and improving the predictions of MGE by the analysis of compositional biases in the genome-wide distribution of tetra-nucleotides. Therefore, to achieve the latter, the following strategies were put together:

- The development of an automated statistical method to study oligonucleotide usage patterns and the detection of mobile genomic elements in prokaryotic genomes.
- Using the statistical method that will be developed in the study to perform a global search for mobile genetic elements in the completely sequenced prokaryotic genomes available in the databases.
- Creating networks and classifying mobile genetic elements into homologous groups based on their shared DNA sequences.
- Classification of mobile genomic elements into protein families and common functional profiles.

- Determination of common and preferred gene entities that are harboured by mobile genetic elements.
- Development of a database-driven web-based interface for mobile genomic elements.

Chapter 2

Seqword Gene Island Sniffer design and implementation

2.1 Overview

Horizontal gene transfer is an event well described as a transfer of genetic material among related and unrelated organisms, that is mediated by mobilomes such as bacteriophages and conjugative plasmids (Canchaya *et al.*, 2003; Hacker *et al.*, 2003*a*). The latter mobilomes play major beneficial roles in bacterial communities, as they disseminate genes whose products are essential in processes such as virulence, niche adaptations, metabolic functions, antibiotic resistance and, the evolution of new species. Such entities are recognized as multitudes of genomes get sequenced, most are found to carry genes with no known function. Explorations of genomic sequences made it evident that DNA composition varies in bacterial genomes (Karlin and Burge, 1995; Pride *et al.*, 2003). These variations result from the exchange of genetic components between microorganisms by a mechanism of horizontal transfer. Genome-wide analysis techniques have uncovered that each microbial species gets acted upon by a specific mechanism that alters their compositional features. Therefore, horizontally

transferred genomic regions possess DNA compositional characteristics which are distinct from the rest of the genetic components of their recipient genomes. These regions have no detectable homologs in closely related organisms and show features of mobile genetic elements indicating their acquisition from exogenous sources (Jain *et al.*, 1999; Ochman *et al.*, 2000). Each bacterial species consists of a unique sequence pattern usage designated as a genomic signature. Genomic signatures, also refered to as a bias in frequencies of short DNA fragments (oligonucleotides) serve as a prevalent characteristic that distinguishes between different organisms while being invariant along the genome (Li and Sayood, 2005; Dufraigne *et al.*, 2005).

It has previously been shown that divergent genomic segments can be detected based on their oligonucleotide usage (OU) patterns (Karlin and Burge, 1995; Deschavanne *et al.*, 1999). Yet, Karlin (1998) proved that oligonucleotide frequencies serve as phylogenetic signals, and variations in signature patterns could be illustrated by studying distributions of words as small as dinucleotides. Genomic structural properties can be revealed by studying constraints beyond frequencies of di- and trinucleotides (Pride *et al.*, 2003). Distributions of longer words were later studied suggesting that their usages may be efficient in the classification of species since shorter words are poorly species specific. Genomic segments of tetranucleotide frequencies proved to be highly conserved (Noble *et al.*, 1998) and could therefore be used to discriminate between species. These observations motivated us to develop a novel tool that examines and detects variances in frequencies of oligonucleotides, to efficiently trace down the distributions of mobile genomic elements across genomes by the patterns of 4-base long words.

In the present study, a statistical method specialized for the analysis of distributions of genome compositions and identification of horizontally acquired genomic elements introduced previously (Reva and Tummler, 2004, 2005), and also a comprehensive genomic islands database were established. This method detects genomic fragments that are of potential horizontal origin by using OU statistics, particularly

by the analysis of tetramer distributions. The aim was that the study of organizations of sequence patterns of genomic fragments on the basis of oligonucleotide usage biases could reveal divergent genomic segments together with their phylogenetic features. The Seqword Gene Island Sniffer (SWGIS) tool was developed to automatically detect gene segments that have been acquired through a mechanism of horizontal transfer, and forms part of SeqWord (Ganesan *et al.*, 2008), an in-house genome browser tool which identifies and visualizes atypical bacterial genomic regions with the use of OU statistical parameters. The Genomic Islands Database (GEI-DB) is a collection of deviant genomic regions that have been identified in prokaryotes using oligonucleotide usage statistics parameters by SWGIS. It has previously been shown that the study of distributions of oligomers can be used to reveal divergent genomic regions and determine the differences in DNA structural conformations in prokaryotic genomes. Upon tetramers searches in 637 bacterial genomes obtained from NCBI (`http://www.ncbi.nlm.nih.gov/Genbank/`) a total of 3518 genomic islands coordinates were identified and used to populate the database. The underlying information provides phylogenetic profiles for individual genomes and the evolutionary history shared among lineages from different inhabitants as to how they get to colonize various niches and carry out different functional traits.

2.2 Seqword Gene Island Sniffer

The method evaluates variances of oligonucleotide frequencies (OUV) and calculates distances (D) between local pattern deviations of the same type and pattern skews-distances between patterns calculated for two strands of DNA. Oligonucleotide patterns are characterized as sets of overlapping words. For example a sequence of AGGCTGGAT can be expressed as tetramers of: AGGC, GGCT, GCTG, CTGG, TGGA, GGAT. Sets of genomic patterns are characterized as $Param$:$type\_N$-mer. $Param$ represents different parameters, which can either be distance between the

local and global OU patterns in the genome, denoted as D:$type\_N$-mer or variances of ou frequencies denoted as OUV : $type\_N$-mer, or PS for pattern skew which determines the distance between ou patterns calculated for both strands of DNA. The component *type* denotes normalization methods and can be represented as $Param$:$n0\_N$-mer for non-normalized or $Param$:$n1\_N$-mer for normalization with mononucleotide frequencies. N represents different word lengths, thus tetramers are characterized as: $Param$: $type\_4$-mer. For example, ouv of a tetranucleotide pattern normalized with a mononucleotide is represented as $RV : n1\_4mers$, whereas a pattern without any normalization is denoted as $RV : n0\_4mers$. The statisctical parameters that were implemented in the method will be discussed in greater detail in section 2.3 below.

Seqword Gene Island Sniffer is a computer program developed in the Python programming language, which detects foreign genomic inserts by determining intragenomic variations between local OU patterns computed for a DNA locus and tetranucleotide patterns of the whole genome. The normalizations of OU were used in two ways, first, by the frequencies of constituent words in the current genomic fragment designated as internal normalization (RV), and second, by the frequencies of the same type of shorter constituent words across the whole genome designated as global normalization (GRV). Foreign inserts are identified by an alternative oligonucleotide usage (increased D:$n0\_4mer$), with lower local RV:n1i_4mer and an increase in GRV:n1g_4mer. PS:n0_4mer comparison was used to filter out clusters of RNA genes characterized with extreme values of PS:n0_4mer (Reva and Tummler, 2005). The latter parameters are measured by the use of a sliding window approach, whereby values of genomic fragments of 8kbp with a 2kbp step are compared with the tetranucleotide usage values of patterns of the same type in the whole genome. If the program recognizes a statistically reliable increase of local distance $D : n0\_4$ values accompanied by a significant decrease of $RV : n1i\_4mer$ and an increase of $RV : n1g\_4mer$, the window shifts several steps back and repeats analysis, this

time with steps of 0.2kbp. In such a way the exact foreign inserts coordinates are identified.

2.3 Algorithm

Seqword Gene Island Sniffer takes as input prokaryotic nucleotide sequences in FASTA format and determines distributions of divergent genomic regions based on the analyses of number of occurrences of different overlapping N-Long ($4mers$) oligonucleotide patterns in a sequence of L_{seq} nucleotides. Oligonucleotide patterns are denoted as a matrix of deviations $\triangle_{[\xi_1 ... \xi_N]}$ of the observed from expected counts for all possible words of length N:

$$\triangle_{[\xi_1 ... \xi_N]} = \left(C_{[\xi_1 ... \xi_N]|obs} - C_{[\xi_1 ... \xi_N]|e}\right) \diagup C_{[\xi_1 ... \xi_N]|0} \text{ [i]}$$

$\xi_n$denotes any nucleotide A, T, G or C in the N-long word, $C_{[\xi_1 ... \xi_N]|obs}$ is the observed occurence of word $[\xi_1 ... \xi_N]$; and $C_{[\xi_1 ... \xi_N]|e}$ is the expected count while $C_{[\xi_1 ... \xi_N]|0}$ is the standard count of an equal distribution of oligonucleotides in the sequence: $\left(C_{[\xi_1 ... \xi_N]|0} = L_{seq} \times 4^{-N}\right)$. The occurrences of oligonucleotides of length N were determined using a normalization method, if OU is not normalized then $C_{[\xi_1 ... \xi_N]|e} =$

$C_{[\xi_1 ... \xi_N]|0}$ and if OU is normalized by empirical frequencies of shorter words of length n, $C_{[\xi_1 ... \xi_N]|e} = C_{[\xi_1 ... \xi_N]|n}$. The expected count of a word $C_{[\xi_1 ... \xi_N]|e}$ of a certain length designated N in a L_{seq} long sequence normalized by frequencies of n-mers (n < N) was calculated as follows :

$$C_{[\xi_1 ... \xi_N]|n} = L_{seq} \times F_{[\xi_1 ... \xi_N]} \times \prod_{i=2}^{N-n+1} \left[\frac{F_{[\xi_1 ... \xi_{i+n-1}]\xi_{i+n}}}{\sum_{A,T,G,C} F_{[\xi_i ... \xi_{i+n}]\xi}} \right] \text{ [ii]}$$

$F_{[\xi_1 ... \xi_N]}$denotes the observed frequencies of n-long word in a sequence, ξ is any nucleotide A, T, G or C.

The distance between two patterns, mainly global and local patterns were calculated as the sum of the absolute distances between values of deviations of observed from expected counts of identical tetramers (w, in a total 4^Ndifferent words). Distance measures were conducted after the ordering of words (occurrences of tetramers) by their $\triangle_{[\xi_1 \ldots \xi_N]}$values in patterns i and j for direct and reverse strands as follows:

$$\mathrm{D}(\%) = 100 \times \frac{\sum_{w}^{4^N} |rank_{w,i} - rank_{w,j}| - D_{min}}{D_{max} - D_{min}} \text{ [iii]}$$

Pattern skew is a particular case of D (distance) where patterns i and j are calculated for the same DNA but for direct (plus) and reverse (minus) strands, respectively. It was calculated as follows:

$D_{max} = 4^N(4^N - 1)/2$ [iv]

D_{max} is the maximal distance between that is theoretically possible between two patterns of N long words. D_{min} is the minimal distance between two patterns. The minimal distance calculated for two independent sequences is zero, but has a positive values for the two complementary strands of the DNA sequence, for the ou patterns designated for both strands of the same DNA molecule cannot be identical. Also, the minimal theoretical distance between two patterns of opposite strands is realized if the words and their reverse complements are distributed with similar frequencies and it is:

$D_{min} = 4^N$ if N is an odd number [v (a)]

but

$D_{min} = 4^N – 2^N =$ if N is an even number [v (b)]

because palindromes, which occur in both strands with the same frequency, only exist in words with an even number of nucleotides and the total number of all possible palindromes is 2^N.

The relative variance (RV) of an ou pattern was calculated as follows:

$$\text{RV} = \frac{\sum_{w}^{4^N} \Delta_w^2}{[4^N - 1]\sigma_0^2} \text{ [v]}$$

where N is word length; $\Delta_w^2$is square of a word w count deviation; and σ_0^2 is the expected variance of the word distribution in a randomly generated sequence that depends on the sequence and word length:

$$\sigma_0^2 = 0.14 + \frac{4^N}{L_{seq}} \text{ [vi]}$$

L_{seq} is sequence length, and N is word length. Normalization of OU pattern variance σ_0 makes the variances comparable regardless of the word length of OU patterns and sequence length.

2.4 SWGIS user-interface

Figure 2.1: SeqWord Gene Island Sniffer command line interface with default settings.

The SWGIS provides an easy-to-use command line interface. Parameters for the runs can all be set depending on the user's desired resolutions. Apart from MGE detection,

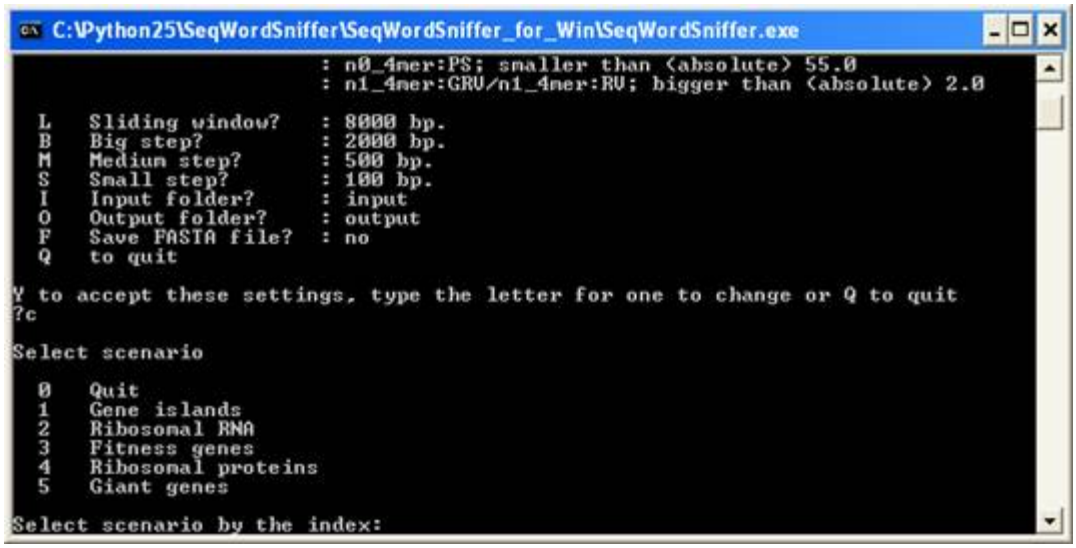


Figure 2.2: SeqWord Gene Island Sniffer command line interface with options of different sequence analysis scenaria.

the tool also allows searches for fitness genes, giant genes, or genes for ribosomal RNA and proteins. SWGIS requires no installation, it takes as input files that are in either FASTA ('FNA','FAS','FST','FASTA') or GenBank ('GBK','GBFF') formats. Figure 2.1 shows the default settings of the tool in the form of a command prompt, that were previously set to identify horizontally acquired genomic islands. The tool also offers shorthand command parameters that allow users to setup the settings for individual runs that are performed during the analysis. For instance, the <C> + <Return> keys allow one to change the scenario of identification, followed by selecting either of the keywords (0-5), each defining a type of scenario of interest as shown in Figure 2.2. There are several other shorthand keys that are to be used, that allow the user to change the settings of the tool. Each run setting starts with its corresponding letter, such as <T> for a keyword: Tasks to perform, (<T> + <Return>) allowing the user to change the settings for particular tasks and also <L> for the keyword: Sliding window (<L> + <Return>) that allows to change the lengths of sliding windows. The <Q> + <Return> keys lets the program return to the main menu upon the editing of certain resolutions, and the <Y> + <Return> keys are for

running the program. The tool processes multiple sequences files in a single run, upon the completion of the runs, the results are saved in forms of text files. The files contain information of all the identified genomic islands together with annotations and coordinates of the genes that they possess, as in the following example:

<GI> 1 <COORDINATES> 583441-620599

[583441:586308:dir]

hypothetical [586849:587388:dir]

[587952:588434:dir]

[588652:590184:dir]

[590325:598688:dir]

[599481:600833:rev]

[600973:601719:dir]

[601854:602831:dir]

[602964:605069:dir]

[605069:605704:dir]

[605726:606337:dir]

[606359:606967:dir]

[607847:614539:dir]

[614543:615793:dir]

[615803:616642:dir]

[617188:618138:rev]

[618386:618586:dir]

[618721:619155:dir]

[619250:619579:rev] <END>

<GI> 2 <COORDINATES> 3487108-3508599

[3487108:3487680:rev]

[3488641:3489390:rev]

[3489577:3490188:dir]

[3490741:3491295:rev]

[3491355:3491678:rev]

[3491722:3492231:dir]

[3492528:3493094:rev]

[3493390:3494199:rev]

[3494291:3495052:rev]

[3495061:3497235:rev]

[3497425:3498099:rev]

internal repeat sequences detected; contains peptidase family M23/M37 as detected by pfam-hmmr [3498112:3507534:rev] <END>

In this example 2 gene islands were identified in the given genome. Each block starts with the island ID and its coordinates in the genome: <GI> 1 <COORDINATES> 583441-620599. If a GenBank file was processed, the annotation and coordinates [left (start coordinate) : right (end coordinate) : strand (direct or reverse)] of all genes inside the genomic fragment follow. The end of the block is marked by <END>.

2.5 Database structure and description

The GEI-DB back-end was implemented in MySQL relational database that is available at http://anjie.bi.up.ac.za/geidb/geidb-home.php. MySQL was chosen because of its efficiency in handling different types of data, speed, ease of use, its ability to run on many platforms, and the libraries it contains, that are accessible by all the major programming languages. The genomic islands data stored within GEI-DB was organized in separate tables to allow speedy interactions and data retrieval, the tables are shown schematically in Figure 2.3. The GEI-DB is composed of six tables, such as: Genomes, Genomic islands (GI), GI-gene names, Blast results, MCL

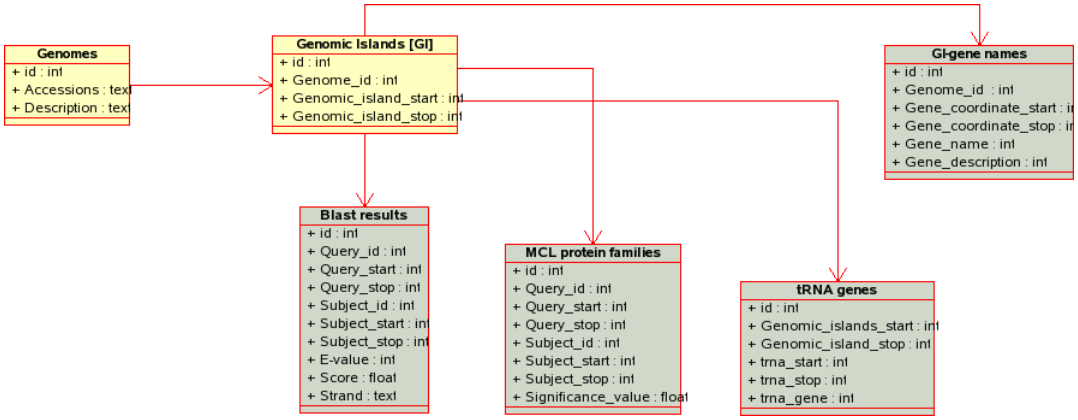


Figure 2.3: Genomic islands database schema.

protein families, tRNA genes. The Genomes table contains information on names of organisms used in the study as well as their Genbank accession numbers. Genomic islands tables contains genomic islands in the form of start and stop coordinates, and their reference genome ids. GI-gene names describes the names of the genes that are located within the identified genomic islands, specified gene locations (start and stop coordinates) are also provided. The BLAST results table provides the information on regions of similarity among the genomic islands that were identified by an all vs. all BLAST search. The MCL protein families (Lima-Mendez *et al.*, 2008*b*) table provides associations in protein functions that the genomic islands have in common. The tRNA genes table describes the names of tRNA genes that genomic islands use for attachment and integration into chromosomes.

2.6 The GEI-DB

The gene information provided in the database was retrieved through various analytical procedures. The statistical method that was used to identify GI has been described in detail in section 2.2 designated Seqword Gene Island Sniffer (SWGIS).

SWGIS searches each prokaryotic genome for foreign inserts based on the comparisons of tetranucleotide usage patterns, where the frequencies of particular tetramers are compared with expected occurrences of the same tetramers throughout the whole genome. The genomic regions with tetramer frequencies that deviate from the genome signature genomic islands (GI). Upon their detection, these regions are further analysed using tools such as BLAST, MCL algorithm (Lima-Mendez *et al.*, 2008*b*), and tRNAscan (Lowe and Eddy, 1997). BLAST is widely known as a rigorous statistical driven tool that performs sequence similarity searches to study and reveal evolutionary relationships between organisms. It was used to perform similarity searches of the identified genomic islands against one another on the nucleotide level. The search was performed to identify conserved genomic regions that are shared by genomic islands from different bacterial lineages and to classify them into groups according to the evolutionary information they have in common. The genomic islands were further grouped into clusters of proteins that share similar functional characteristics using a Markov clustering algorithm (TRIBE-MCL). TRIBE-MCL is a method mainly used to classify proteins into clusters of families represented as square matrices, on the basis of their precomputed sequence similarity information (Enright *et al.*, 2002). Moreover, each genomic island was scanned for tRNA genes using the tRNAscan tool (Lowe and Eddy, 1997) , for tRNA genes are known to be widely favored by mobile genomic elements as chromosomal insertion sites.

Currently, the GEI-DB contains a set of 3518 precalculated genomic islands (GI) identified in 637 prokaryotic genomes. The results that were obtained from the latter analysis were populated in GEI-DB, in accordance to the schema displayed in Figure 2.3 and are easily retrieved and viewed in the form of a web-interface. The web-interface was implemented in PHP, JAVA script, hypertext markup language (HTML) and cascading style sheets (CSS) (Figure 2.4). PHP was chosen for its ease to access and manipulate MySQL databases, also that it is best suited for web development. It can be used with web-servers such as Apache, Netscape, Microsoft

IIS, and can also run on platforms such as Solaris, Linux, FreeBSD, Mac OS X and Windows. The PHP code was incorporated with a MySQL database to enable fast connections and queries to end-users, also to allow viewing of results in the form of a text-based web document. The GEI-DB system architecture as illustrated in Figure 2.4 allows a better understanding on how the data is stored, retrieved and viewed. The figure shows that all the contents of the GEI-DB are stored in a MySQL database, therefore sequence retrieval from the database allows the apache web-sever to process the PHP commands and send the output to the web-browser in the form of a dynamic page (php generated HTML page), which therefore allows viewing of the information stored in the back-end such as: genomes, genomic islands, GI-gene names, BLAST results, MCL protein families and tRNA genes.

2.6.1 Web search interface

The user-friendly GEI-DB pages are organized in a way that easily allows users access to both the relational database and the underlying data that belongs to individual genomes. The main page (Figure 2.5) provides names and Genbank accession numbers of the deposited organisms and also offers a simple selection for individual organisms which further allows the exploration of their underlying information. Upon selection of an organism of choice, users get presented with a page that includes genomic islands profiles that belongs to their associated genome. Each genomic island has additional hyperlink parameters that allow the viewing of either the names of genes that each harbours (Figure 2.6-B), their phylogenetic inferences obtained by BLAST (Figure 2.6-C), classifications of protein families determined by MCL (Figure 2.6-D) or the tRNA genes that each carry (Figure 2.6-E).

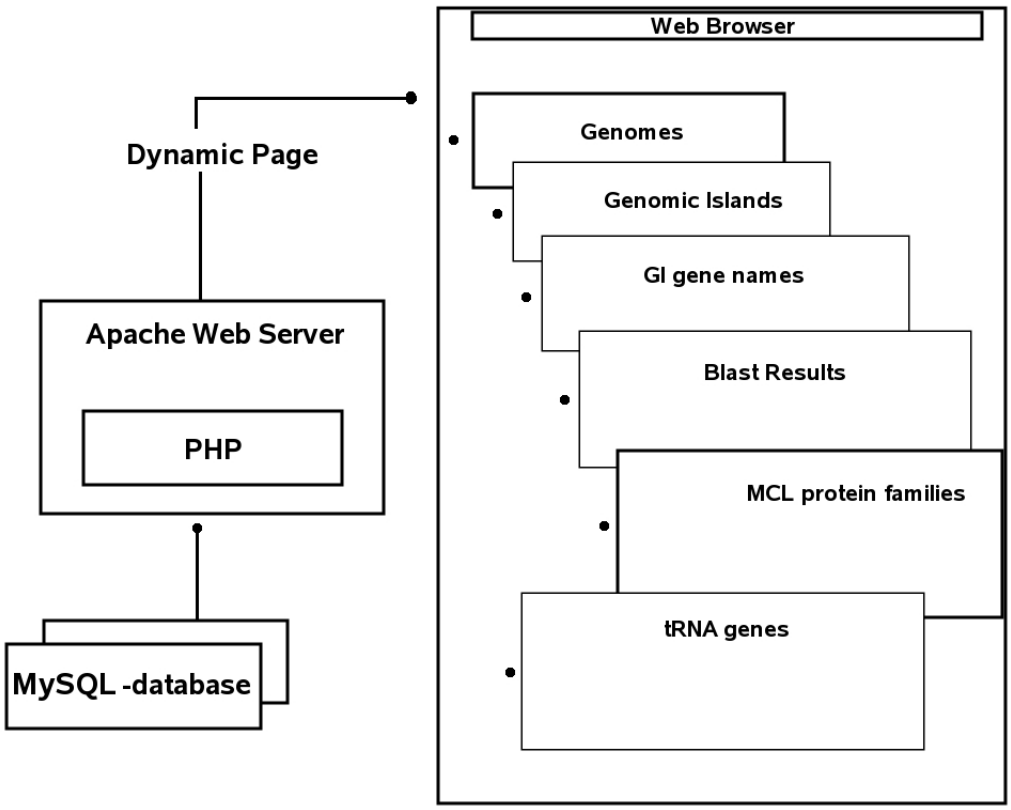


Figure 2.4: Genomic islands database architecture.

2.6.2 Principles underlying the GEI-DB

The GEI-DB was designed to provide insights in the central dogma underlying the functional characteristics of prokaryotic genomes. Multitudes of prokaryotes have been found to inhabit niches that were initially thought to be too extreme to support life, while others evolve to cause biodegradations and produce substances that are harmful to the entire biosphere. Intensive analysis has been conducted on such organisms, and it revealed that most of them carry genes which alter their biochemical pathways. Such genes were found to be acquired from other obligatory organisms that thus enable bacteria to maintain the stability of their gene expression levels and render them fit. Remnants of such genes were found to be in association with small genomic elements that are able to move from one genomic region to another or from

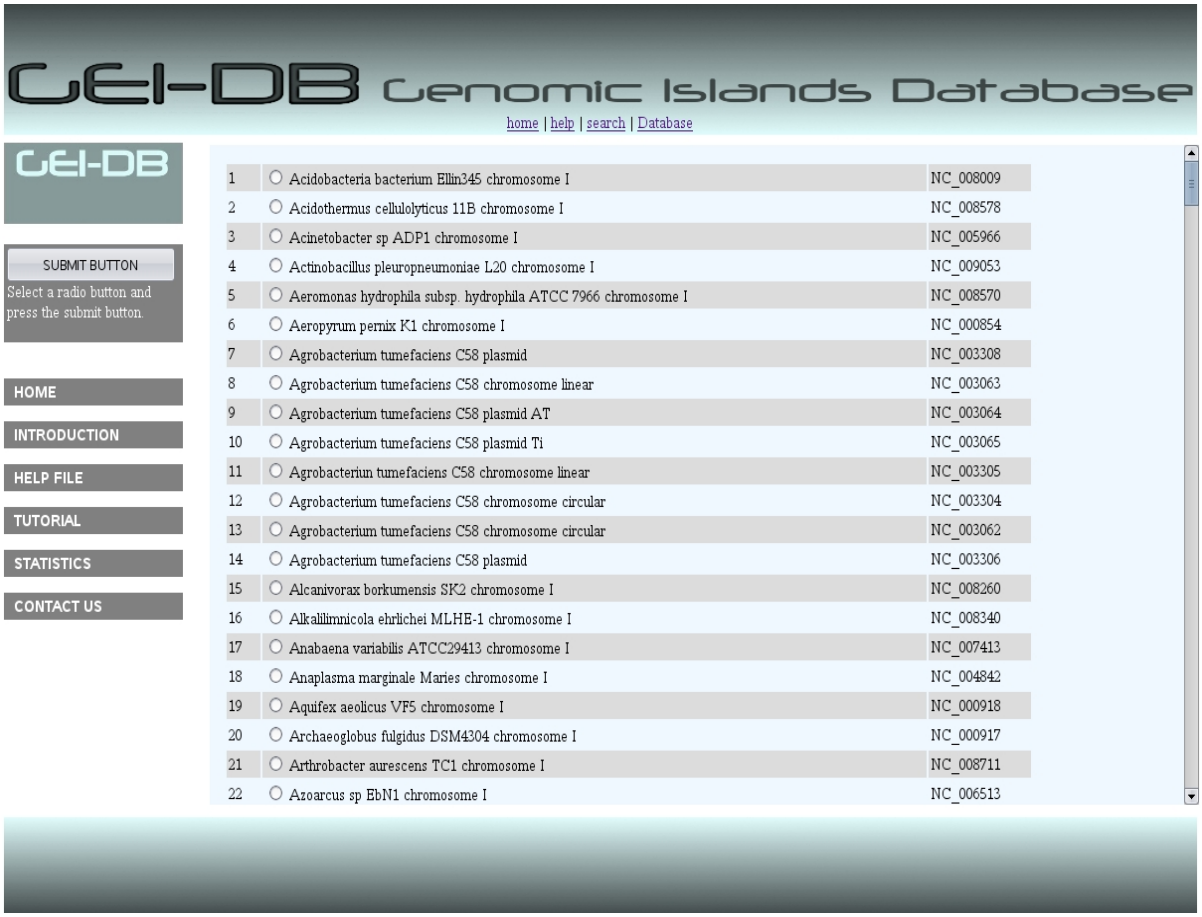


Figure 2.5: GEI-DB font page layout. The page displays the sites optional buttons and names of organism that were used in the study, together with their accession numbers.

one organism to the next regardless of their evolutionary backgrounds. GEI-DB offers information based on such genes-designated genomic islands, that were identified in prokaryotic genomes. Below are components of the GEI-DB that briefly describe the information that is constituted by the databases that can potentially be used to understand GIs evolutionary properties that shape up the microbial communities.

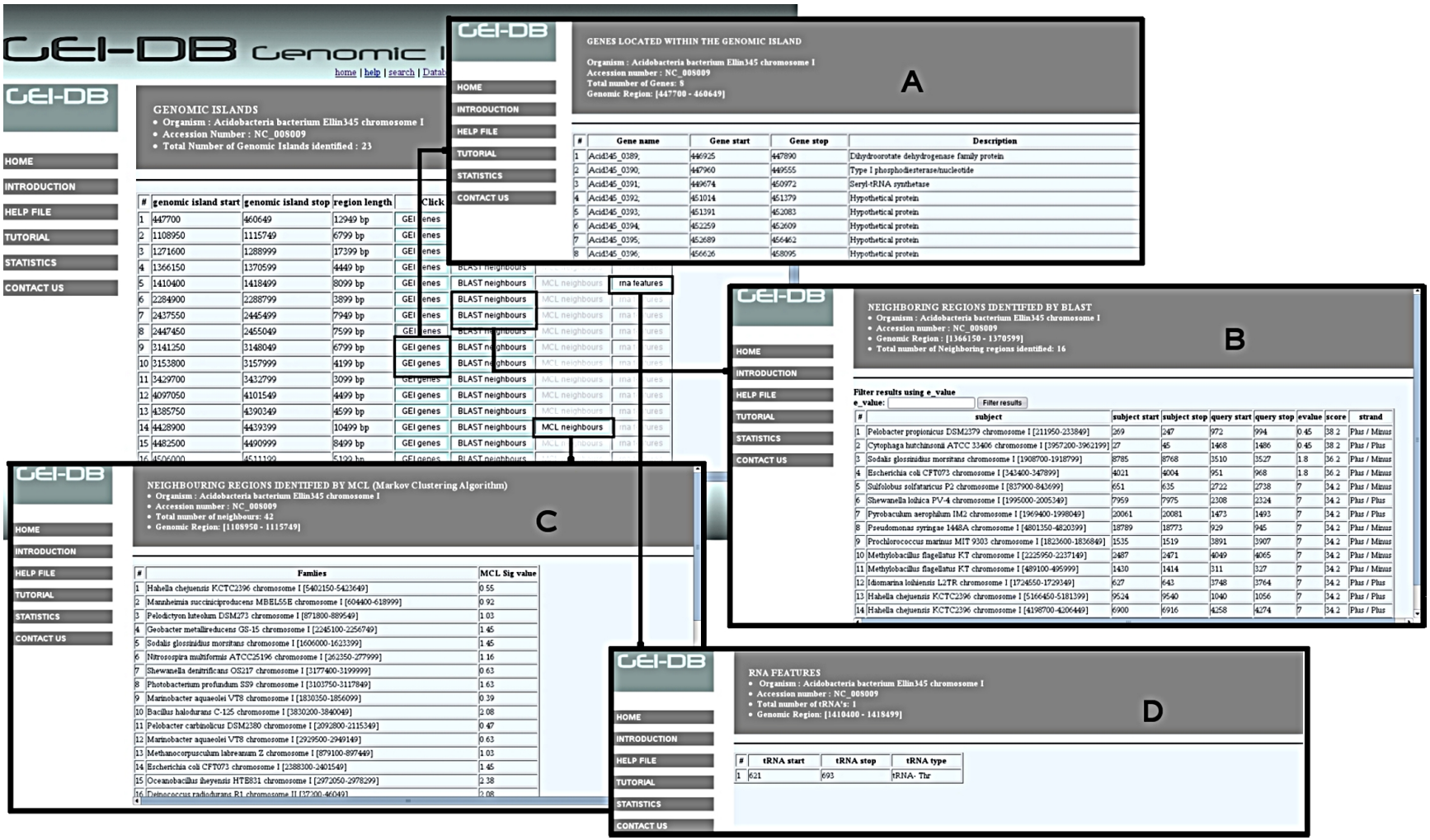


Figure 2.6: GEI-DB additional pages.

- **Genomic Islands [Figure 2.6 -A]**

The Genomic Islands page presents the user with an array of precalculated genomic islands that belong to a particular organism. The page also offers brief statistics on the displayed genomic islands such as the total number of genomic islands and the name of the organism they were identified in, followed by a list of GI coordinates and their lengths. Besides each GI coordinates are further links that lead to acquisitions of their characteristic features and phylogenetic profiles. GIs that possess further information have their associated linkage parameters encircled in green, while those that do not entail any further informational records have their association linkage parameters disabled (in grey).

- **GI-gene names [Figure 2.6 -B]**

The GI-gene names page simply provides the names of genes that are located within GI regions. It also provides the coordinates corresponding to the genes that span a particular genomic island, and additionally the functions that each gene carries are provided.

- **BLAST results [Figure 2.6 -C]**

The BLAST results page provides a list of potential local homologous regions that belong to GI that were identified in organisms from different evolutionary backgrounds. The list of homologous genomic regions were acquired upon a comparison of all the genomic islands against one another. The page illustrates the name of the GI that was used as a query, also provided are the confidence scores and coordinates of genomic regions that share similarity. The similarity searches were conducted in order to group genomic islands according to the the compositional features that they have in common. The confidence (statistical) scores obtained by BLAST allow the user to evaluate the inferences of homology between individual genomic islands obtained from the search. The information also provides the user with information on

sequence conservation and similarities which were obtained from different bacterial lineages.

- **MCL results [Figure 2.6 -D]**

The MCL results page offers classifications of functional related genomic islands in the form of protein families. Functional related proteins are grouped by a measure of sequence similarity obtained by comparing all genomic islands [CDS] with one another. Thus the genomic islands that show similarity in sequence and functional characteristics are grouped together into families. The page displays the names of genomic islands that constitute a family, with significance scores that correspond to the associations of the nodes that make up linkages among genes that share highly similar functional characteristics.

- **tRNA genes [Figure 2.6 -E]**

The tRNA page offers the names of tRNA genes that certain genomic islands are associated with. tRNAs are widely known to be targeted by genomic islands as preferential chromosomal insertion sites. Most genomic islands have been shown to harbour tRNA genes or just parts of them.

2.7 Conclusion

The GEI-DB is made up of various functionalities and mobile genomic elements entities that could further be used for more analysis by the scientific community. It provides a collection of mobile elements from organims of various backgrounds, comprising prokaryotes that are of medical and environmental importance. The database also allows browsing of genomic elements that were classified according to DNA and protein compositional features that they have in common and also

gene entities that are associated with these elements. The amount of data that is available in the database allows one to be able to characterize the types of genes that are favorably acquired and transferred during horizontal events, which also allow the studying of evolutionary profiles that mobile elements have in common.

Chapter 3

SWGIS and GEI-DB utility

3.1 SWGIS and MGE analysis

Seqword Gene Island Sniffer was used to perform a global search of foreign genomic inserts throughout a total of 637 bacterial chromosomes which were downloaded from NCBI (`http://www.ncbi.nlm.nih.gov/Genbank/`), and managed to identify 3518 of such putative inserts designated as genomic islands. These genomic islands are loci which were determined by Seqword Gene Island Sniffer as divergent regions that exhibited variations in their tetranucleotide usage patterns as compared to all the other fragments in their host genomes. Their divergences are indicated by an increase in distance D of local $n0\_4mer$ patterns from those calculated for the whole chromosome, and the divergences between n1_4mer:RV and n1_4mer:GRV values. The variations of genomic components were determined as illustrated in Figure 3.1 where variances of tetranucleotides in distributions in a genome are presented in a form of x (n1_4mer:GRV) and y (n1_4mer:RV) coordinates, where the distances of OU patterns of genomic fragments from the whole genome pattern are represented by colors of dots.

Fragments that correspond to the core genome are clustered together, as depicted by

blue dots (Figure 3.1), indicating that core genes in a given genome have homogenous compositional characteristics, and are therefore illustrated by a low D and similar RV and GRV values. Fragments with compositional characteristics that are distinct from the core genome are depicted by red and brown dots, these colors indicate an increase in D, and the remarkable difference between RV and GRV patterns values. A gradient from red to green dots represent the amelioration of horizontally acquired genomic fragments. It is believed that recently acquired genomic regions resemble characteristc features of their donor genomes, but overtime they become subject to mutational pressures affecting all the genes in their recipient genome and ameliorate to resemble OU features of their new genome (Lawrence and Ochman, 1997).

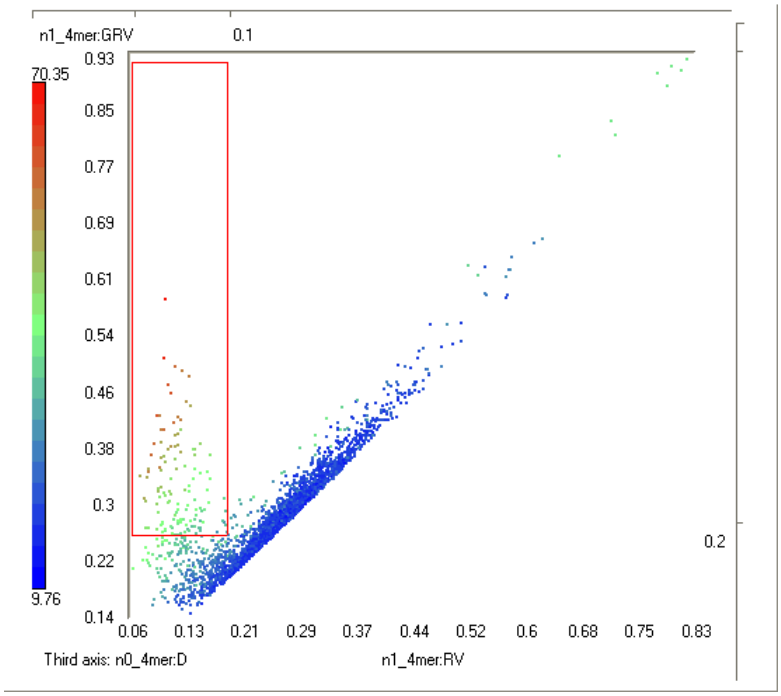


Figure 3.1: The Figure was obtained from SeqWord (in-house online tool that uses the same concept of Oligonucleotide statistics to identify atypical genomic regions) (Ganesan et al 2008). n0_4mer:*D* color coated bar situated on the left pane of the figure measures differences in variances between values of local [n1_4mer:RV] and global [n1_4mer:GRV] oligonucleotide patterns of the same type in a given genome. Colors of dots correspond to the values displayed on the n0_4mer:*D* bar on the left. Blue dots which are clustered linearly entail core genes in genome that possess similar n1_4mer:RV and n1_4mer:GRV compostional characteristics. Red dots entail an increase in distance between n1_4mer:GRV and n1_4mer:GRV genomic fragments. Green dots are horizontally acquired genomic fragments that are starting to resemble compositional characteristic features of the host genome by amelioration.

3.1.1 Divergent genomic regions and virulence determinants of *Salmonella enterica subsp. enterica serovar Typhi Ty2* (NC_004631)

Many horizontally transferred elements are associated with virulence, thus the ability of Seqword Gene Island Sniffer to detect pathogenicity islands is of practical importance. Pathogenicity islands (PAIs) are subsets of genomic islands and are associated

with atypical GC composition, mobile genes, and proximal transport RNAs, collectively known as genomic regions that depict horizontal origin (Hsiao *et al.*, 2003*a*). These horizontally acquired genomic regions harbor genes with compositional features that are distinct from all the other genes in their resident genomes. They contain genes that encode virulence determinants that regulate fitness and survival of pathogens within host cells, and also carry determinants that encode adhesins and invasins that allow pathogens access to host sites unreachable by commensal microorganisms. Acquisition of these foreign determinants contribute towards the evolution and creation of pathogenic variants and are believed to be carried around by lysogenic bacteriophages and plasmids. PAIs affect a wide range of gram-negative and gram-positive bacteria, but are mostly observed in enteropathogenic microorganisms.

In this study, a genome wide analysis of *Salmonella enterica subsp. serovar typhi Ty2* with SWGIS was performed. *Salmonella* is a facultative gram-negative intracellular pathogen that causes diseases such as gastroenteritis and typhoid fever in mammals (Marcus *et al.*, 2000). It is estimated that there are about 15 million infections and over 500 000 deaths each year resulting from typhoid fever and gastroenteritis diseases. The diseases are caused by the ingestion of contaminated water or food. Upon ingestion, the microbe adheres and invades cell linings of the intestinal epithelium and thus initiates virulence. *Salmonella* has the ability of surviving the host immune response and it can also invade phagocytic and non-phagocytic cells of the host and replicate intracellularly. This facultative pathogen resides within the phagosomal vacuole. It carries genes that protects it from environmental changes such as low pH and toxic substances that are released by the phagocyte. The microbe also harbors determinants that are responsible for its virulence, genes which are only present in virulent *Salmonella* strains but absent from their benign relatives. *Salmonella* is made up of large genomic regions of sizes that vary between 1 to 40 kb known as *Salmonella* pathogenicity islands (SPI). SPIs are defined as genomic segments within the *Salmonella* chromosome that encode genes required for direct interaction with

the host and virulence. *Salmonella* has two important virulence characteristics that are linked to its functionality, designated as *Salmonella* pathogenicity islands (SPI) types SPI-1 and SPI-2 (Hansen-Wester and Hensel, 2001). SPI-1 encodes genes that are required for host invasion while SPI-2 encodes those that are required for systemic infections. These genomic segments have similar characteristics associated with the other pathogenicity islands and genomic islands from other organisms. They exhibit atypical features such as low GC content and are mostly inserted in conserved chromosomal tRNA regions indicating their acquisition by phages, because tRNAs are widely known as phage integration sites. tRNA regions are favored as insertion sites by most mobile genetic elements because they are transcriptionally active, and ensure immediate expression and effective translational efficiency of acquired determinants (Dobrindt *et al.*, 2002; Hacker *et al.*, 2003*a*). Seqword Gene Island Sniffer managed to detect an SPI-2 [1328750 - 1356649] which will be discussed below, a pathogenicity island that has previously been identified by other existing methods (Yoon *et al.*, 2005), as a support for the validity of our method.

Table 3.1: *Salmonella enterica Ty2 chromosome I* Genomic Islands that are identified by SWGIS.

Organism : *Salmonella enterica Ty2 chromosome I*
Accession Number : NC_004631
Total Number of Genomic Islands identified : 47

id	**start**	**stop**	**length**		**id**	**start**	**stop**	**length**
1	31350	40949	9599 bp		25	2625700	2633699	7999 bp
2	535650	544949	9299 bp		26	2633550	2640299	6749 bp
3	789650	794299	4649 bp		27	2762350	2769799	7449 bp
4	803700	815149	11449 bp		28	2847050	2866799	19749 bp
5	865850	887849	21999 bp		29	2877500	2884999	7499 bp
6	952500	978749	26249 bp		30	2937500	2943599	6099 bp
7	1093300	1108849	15549 bp		31	3034200	3043549	9349 bp
8	1186900	1204199	17299 bp		32	3115950	3127799	11849 bp
9	1273600	1281999	8399 bp		33	3506200	3511899	5699 bp
10	1328750	1356649	27899 bp		34	3848550	3854449	5899 bp
11	1506050	1514549	8499 bp		35	3875500	3884399	8899 bp
12	1522550	1528149	5599 bp		36	3914150	3925849	11699 bp
13	1588250	1596249	7999 bp		37	4065700	4071799	6099 bp
14	1611500	1617499	5999 bp		38	4304900	4314999	10099 bp
15	1629000	1637799	8799 bp		39	4323700	4332449	8749 bp
16	1656500	1661599	5099 bp		40	4385150	4393599	8449 bp
17	1845000	1853899	8899 bp		41	4426250	4432899	6649 bp
18	1887650	1896699	9049 bp		42	4441900	4447849	5949 bp
19	1926700	1932249	5549 bp		43	4478900	4483699	4799 bp
20	2209850	2222349	12499 bp		44	4501350	4517799	16449 bp
21	2370400	2377849	7449 bp		45	4665850	4671349	5499 bp
22	2558650	2563099	4449 bp		46	4676100	4685499	9399 bp
23	2588500	2598149	9649 bp		47	4696650	4707749	11099 bp
24	2618900	2627249	8349 bp					

Screening the entire *Salmonella* genome detected 47 atypical genomic regions (Table 3.1). This indicates that *Salmonella enterica Ty2* genome contains many gene clusters that are of horizontal origin. Most of the identified elements in the list (Table 3.1) are pathogenic, they carry virulence genes that encode enzymes required for the biosynthesis of nutrients that are scarce within host tissues and determinants that are required for defense against microbiocidal mechanisms (Groisman and Aspedon,

1997). Genomic islands [1328750 - 1356649] and [1186900 - 1204199] (hereafter GI id 10 and GI id 8, of Table 3.1) listed in Tables 2 and 3 will be discussed in detail because they span larger chromosomal regions and carry genes that are primarily required by pathogens to survive and replicate within hosts. Determinants found on GI id 10 and id 8 were annotated in Tables 3.2 and 3.3, respectively, using SWGIS. The genes contained within these genomic regions support the fact that horizontal transfer plays a major role in allowing microoganisms to adapt to a variety of environmental conditions.

3.1.2 Pathogenicity island [1328750-1356649]

Table 3.2 lists 34 genes that are present within *Salmonella enterica subsp. serovar typhi Ty2* GI id 10 as obtained from the GEI-DB graphical user-interface (GUI). The latter genomic region was designated as *Salmonella* pathogenicity island type 2 - SPI2. The region is made up of genes that enable microbes to cope with environmental changes within the phagolysosome (Coombes *et al.*, 2003). SPI-2 is mainly made up of at least four operons that encode molecular chaperones, two component regulatory systems, secretion system effectors and the type III secretion apparatus proteins (Coombes *et al.*, 2007) of the type III secretion system (TTSS). TTSS is a functional molecular mechanism used by pathogenic microorganisms to deliver effector proteins to the host to promote cell invasion and disruption of its cytoskeleton and signalling pathways (Marcus *et al.*, 2000; Dai and Zhou, 2004). The system is encoded by both SPI-1 & 2, and its effect is triggered by entry of pathogens into the host. SPI-1 plays a role in the invasion of non-phagocytic epithelial cells by S*almonella*, mediated by the translocation of effector proteins (Hansen-Wester *et al.*, 2002) and SPI-2 plays a critical role in the microbe's virulence, and survival within macrophages (Dai and Zhou, 2004).

Table 3.2: *Salmonella* Pathogenicity island [1328750 - 1356649].

Organism : *Salmonella enterica Ty2 chromosome I*
Length: 27899 bp
Number of genes : 34
tRNA: Val

id	Gene	Loci	Function
1	t1257	1328596 - 1329555	putative pathogenicity island protein
2	t1258	1329727 – 1330455	putative transcriptional regulator
3	SsrB	1330634 - 1331272	putative two-component response regulator
4	ssrA	1331303 – 1334065	putative two-component sensor kinase
5	spiC	1334484 – 1334867	putative pathogenicity islands secreted effector protein
6	spiA	1334869 – 1336362	putative outer membrane secretory protein
7	saaD	1336343 – 1337554	putative pathogenicity island protein
8	ssaE	1337562 – 1337804	putative sectretion system protein
9	sseA	1338007-1338333	putative pathogenicity island protein
10	sseB	1338340-1338930	putative pathogenicity island effector protein
11	sscA	1338927-1339400	putative type III secretion system chaperone protein
12	sseC	1339403-1340857	putative pathogenicity island effector protein
13	sseD	1340873-1341460	putative pathogenicity island effector protein
14	sseE	1341463-1341879	putative pathogenicity island effector protein
15	sscB	1341931-1342365	putative pathogenicity island protein
16	sseF	1342381-1343163	putative pathogenicity island effector protein
17	sseG	1343160-1343849	putative pathogenicity island effector protein
18	ssaG	1343943-1344158	putative pathogenicity island protein
19	ssaH	1344199-1344426	putative pathogenicity island protein
20	ssaI	1344438-1344686	putative pathogenicity island protein
21	ssaJ	1344683-1345432	putative pathogenicity island lipoprotein
22	t1278	1345450-1345998	putative pathogenicity island protein
23	t1279	1345995-1346669	putative pathogenicity island protein
24	ssaL	1346659-1347651	putative secretion system protein
25	ssaM	1347709-1348077	putative pathogenicity island protein
26	ssaV	1348062-1350107	putative type III secretion protein
27	ssaN	1350097-1351398	putative type III secretion ATP synthase
28	ssaO	1351401-1351778	putative type III secretion protein
29	ssaP	1351759-1352133	putative type III secretion protein
30	ssaQ	1352114-1353082	putative type III secretion protein
31	yscR	1353150-1353797	putative type III secretion protein
32	ssaS	1353794-1354060	putative type III secretion protein
33	ssaT	1354061-1354840	putative type III secretion protein
34	saaU	1354837-1355895	putative type III secretion protein

GI id 10 is clustered with essential genes that encode structural components of the type III secretion apparatus (Hensel *et al.*, 1997). Also within the region are membrane-bound sensor kinase (secretion system receptor A - ssrA) and cognate response regulator (secretion system receptor B - ssrB) genes that regulate the transcription of TTSS (Coombes *et al.*, 2007). The two-component regulatory system regulates genes in response to extracellular concentrations of divalent cations (Hansen-Wester and Hensel, 2001). Membrane-bound sensor kinase (ssrA) forms the first component of the system that senses changes in environment. Its phosphorylation is stimulated by low levels of Mg^{2+}/Ca^{2+}(Feng *et al.*, 2003; Ansong *et al.*, 2008). Upon its activation, the phosphoryl group gets transferred to the second component, the cognate response regulator which then regulates spiC effector protein (Feng *et al.*, 2003). SpiC is a virulence determinant encoded within SPI-2 which lies upstream of ssrA/B, and known to function as an inhibitor of intracelluar trafficking and fusion of the *Salmonella*-containing vacuole with lysosomes and endosomes (Hansen-Wester and Hensel, 2001; Marcus *et al.*, 2000; Uchiya *et al.*, 1999). SpiC protein has also been reported to be required for translocating effector proteins into target cells and the activation of signal transduction to alter functions of macrophages (Uchiya and Nikai, 2008; Freeman *et al.*, 2002). Thus, this promotes survival and virulence of *Salmonella* within phagolysosomes.

Other proteins that are encoded within the region are *Salmonella* secreted effectors designated secretion system effectors (sse) - A, B, C, D, E, F, and G. sseA is essential for the delivery of sseB to the external surface of the bacteria and also serves as its type III secretion chaperone (Zurawski and Stein, 2003). Secretion system effector proteins B, C and D predominantly reside on bacterial surfaces upon their secretion. The latter three effector proteins work as a functional unit of translocons, a unit required for the translocation of other effectors into infected cells (Kuhle and Hensel, 2002; Hansen-Wester and Hensel, 2001; Nikolaus *et al.*, 2001). Upstream of effector proteins sseF and G lies sscB, a chaperone that facilitates the secretion and

function of sseF within host cells (Dai and Zhou, 2004). Unlike the other secretion system effectors, sseF and sseG plays minor roles in intracellular replications and systemic pathogenesis (Dai and Zhou, 2004; Hensel *et al.*, 1998). Both sseF and sseG were found not to be required for translocations of SPI-2 effector proteins but they do function as translocated effectors that modulate aggregation of host endosomes (Kuhle and Hensel, 2002).

3.1.3 Pathogenicity island[1186900-1204199]

Table 3.3 shows 27 genes located in the GI id 8 that are essential for the virulence and survival of *Salmonella* within macrophages. *Salmonella* harbours many pathogenicity islets, each benefiting it in its own way. This indicates that succesful pathogens require multiple virulence genes to maintain their stability. Among the determinants that are located within the PAI is a cytolethal distending toxin (cdtB) encoding gene, widely known to be produced by disease causing microbes such as *Campylobacter jejuni* and *E. coli* (Pickett *et al.*, 1996). *Salmonella* expresses cdtB after it gets phagocytosed, indicating its necessity for adaptation and survival within human macrophages (Ansong *et al.*, 2008).

Other genes contained within the region that play an important role in *Salmonella* pathogenicity are macrophage survival gene - msgA, envelope gene - envE, virulence membrane gene - pagD and virulence membrane gene - pagC. Genes envE, and pagD are transcribed upstream of pagC and function as outermembrane envelope proteins required for intramacrophage survival, whereas pagC and msgA result in both survival within macrophages and systemic infections (Gunn *et al.*, 1995). Both pagC and pagD are regulated by PhoP/PhoQ two-component system (Gunn *et al.*, 1995), in response to low Mg (2+) concentrations within the cell. GI id 8 was found to be homologous to SPI-11 [1350481 - 1366166] (Chiu *et al.*, 2005) of *Salmonella enterica subsp. enterica serovar Choleraesuis str. A* serotype that causes swine parathyroid

and extra-intestinal infections in humans (Chiu *et al.*, 2006) also known to be the most invasive of all non-typhoidal *Salmonella* which are in possession of a total of five SPIs (Chiu *et al.*, 2005). It has also been indicated by BLAST similarity searches that other strains such as *S. enterica subsp. enterica serovar Typhi CT18* and *Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150* harbour such genes too. The presence of these genes (msgA, envE, pagD pagC) in the latter genomes provide selective advantage under unfavorable environmental conditions (Hacker and Carniel, 2001) and thus promote bacterial survival and virulence within host cells. The latter genes should therefore be characterised as fitness genes, because they offer fitness properties that are essential for the persistence of microbes within hosts (Hacker and Carniel, 2001).

Table 3.3: *Salmonella* Pathogenicity island [1186900 - 1204199].

Organism : *Salmonella enterica Ty2 chromosome I*
Length: 17299 bp
Number of genes: 27
tRNA name: Arg

id	gene name	region	description
1	t1102	1187029-1187418	putative regulator
2	t1103	1187422-1188231	putative regulator name : sirB1
3	kdsA	1188269-1189123	2-dehydro-3-deoxyphosphooctonate aldolase
4	tnpA	1189317-1189775	transposase for insertion sequence element IS200
5	t1106	1190280-1191404	putative bacteriophage protein
6	t1107	1192445-1192858	putative pertussis-like toxin subunit
7	t1108	1192875-1193603	putative pertussis-like toxin subunit - toxin subunit
8	t1109	1193795-1194337	conserved hypothetical protein
9	t1110	1194485-1194862	protein
10	cdtB	1194935-1195744	putative toxin-like protein
11	msgA	1196617-1196856	putative virulence protein
12	envE	1197047-1197568	putative lipoprotein
13	cspH	1197986-1198198	putative cold shock protein
14	pagD	1198330-1198593	putative outer membrane virulence protein
15	t1117	1198642-1198806	hypothetical protein
16	t1118	1198892-1199020	hypothetical protein
17	pagC	1199398-1199955	outer membrane invasion protein
18	t1120	1199966-1200136	hypothetical protein
19	t1122	1200688-1200861	hypothetical protein
20	t1123	1201189-1201530	putative secreted protein
21	t1124	1201602-1201691	hypothetical protein
22	t1125	1201746-1201835	hypothetical protein
23	t1126	1201911-1202036	putative lipoprotein
24	t1127	1202236-1202463	hypothetical protein
25	t1128	1202658-1203128	putative heat shock protein
26	t1129	1203237-1203386	hypothetical protein
27	t1130	1203467-1204510	hypothetical protein

3.2 GEI-DB and MGE analysis

The GEI-DB was used for further illustrations on the information and features that are possessed by *Salmonella* GI id 10 and GI id 8. The database not only houses the MGE coordinates and names of genes where they are harboured, but also contains the information on the evolutionary profiles that are entailed by each element. Before the database was populated with variable MGE information, several analyses were conducted (described in section 2.6), such as all *vs.* all BLAST and MCL similarity searches and, also searches for tRNAs were performed. *Salmonella* GIs id 8 and id 10 do not share nucleotide similarity (BLAST) with any other MGE in the database. The MCL analysis revealed that although these PAIs do not share any strong nucleotide sequence similarity with the other elements, they do share similarity in profiles of gene functions with the other members of *Salmonella* and *E. coli*, constituting similar sets of protein families. These observations illustrate that organisms may possess non-homologous sequences and yet share protein functional entities. Organism's nucleotide sequences are more susceptible to change because their compositions are mostly selected for by the differences in mutational pressures that act upon them and also the environments that they inhabit. The PAIs were further searched for tRNA sequences with the tRNASCAN tool to check the types of tRNA that they are associated with, because tRNAs are known to be target integration sites of mobile genomic islands, especially those that code for virulence factors. The search detected two components of tRNA at loci 27466-27390 and 27379-27303 of type Valine in GI id 10 and one tRNA of type Arginine in GI id 8 situated at loci 13460-13533.

3.3 BLAST SIG analysis

3.3.1 Phage genomes

Mobilization of genetic segments between organisms that are not of the same taxonomic unit has been shown to be the major driving force in the evolution and speciation of prokaryotes. The transfer of these genomic segments among bacterial species is mediated by vectors such as plasmids and bacteriophages (Wagner and Waldor, 2002). Bacteriophages are viruses that infect bacteria, and transfer foreign genetic material between bacteria by a process of transduction. They are important agents of horizontal gene transfer, and play a major part in bacterial chromosome evolution (Canchaya *et al.*, 2003). There are two groups of phages, designated as lytic and temperate (lysogenic). Lytic phages, just like other viruses, cannot make more copies of themselves when outside of host cells. They therefore parasitize bacterial hosts in order to reproduce maximal amounts of progeny phages and get released following the lysis of the host (Srividhya *et al.*, 2007). Temperate phages have the potential of either infecting their bacterial hosts and lyse them upon replication or integrate their DNA into hosts chromosomes as prophages to establish a stable and functional relationship (Canchaya *et al.*, 2003) for as long as their lytic genes are repressed. A prophage is a phage that has integrated its DNA material with its hosts choromosome through lysogenization, and bacteria with intact prophages are referred to as lysogens. Lysogenic prophages carry with them genes that can change the phenotypic character of their hosts by introducing new fitness factors (Srividhya *et al.*, 2007). The contribution of these novel phenotypic features to a host bacterium is known as lysogenic conversion (Brussow and Hendrix, 2002) and are of essential selective advantage. These features confer traits that help the bacterium to evade and survive environmental challenges. This group of phages are capable of lysogenically converting microorganisms from a state of being a commensal to becoming pathogenic making them contributors to the evolution and emergence of

virulent microbes(Gaidelyte *et al.*, 2007).

Lysogenic phages are composed of virion genes which are not constitutively expressed but get induced (Broudy *et al.*, 2001) when the host is exposed to unfavorable conditions that result in DNA damage. Under unfavorable conditions, a lysogenic phage becomes lytic, gets excised from the host chromosome, packages itself and searches for other suitable hosts to parasitize. Sometimes phages erroneously excise and package their hosts DNA particles with theirs, and transfer them to other organisms thus promoting major variations in gene compositions between species. One of the factors that make bacteriophages the major contributors in bacterial evolution is that they are the most abundant species in the biosphere (Brussow and Hendrix, 2002). Intensive research has been done on phage-host interactions and genes responsible for prophage integrations. Various *in silico* methods have been developed for the detection of prophages in bacterial genomes and to identify compositional features they possess. Some methods are carried out to detect prophages in bacterial genomes on the basis of similarities of their protein sequences with known phage genes (Lima-Mendez *et al.*, 2008*a*).

3.3.2 Prophinder vs Seqword Gene Island Sniffer

Genomic island compositions obtained from SeqWord Genomic Island Sniffer (SWGIS) were compared to a reference dataset of prophages that were predicted by Prophinder (Lima-Mendez *et al.*, 2008*a*) to evaluate its specificity in the detection of horizontally transfer events. Prophinder is an alternative tool that is not dependent on the analysis of sequence composition, it detects MGE by similarity searches of conserved prophage DNA pairs using BLASTP in accordance with known phage-like gene annotation features. Comparisons of the results that were predicted by both the latter methods showed consistency in many cases. From the comparisons it was illustrated that SWGIS failed to identify short and ameliorated MGE that were

efficiently detected by Prophinder, whereas Prophinder was deficient in the identification of horizontally transferred gene cassettes and truncated MGE that were precisely detected by SWGIS. SWGIS, just like any other composition-based method cannot identify MGE with ancient origins because it mainly relies on the detection of DNA fragments with atypical compositions which therefore only allows the identification of recently acquired genomic regions. The latter approach may suffer from several drawbacks but it provides insights beyond those overlooked by feature-based approaches, which only determine foreign inserts on the basis of sufficient homology searches. Prophinder also detects foreign inserts based on known gene annotation features. Although it detected genomic subsets that were missed by SWGIS, this method still has its own limitations as its utility is set by the availability of well annotated MGE in public databases, it therefore overlooks MGE that lack certain annotation features (Rajan *et al.*, 2007). Its outcome is thus likely to be affected by poor and less accurate annotations. Moreover, Prophinder could not identify insertion borders of MGE that SWGIS does. However, the approaches used by both SWGIS and Prophinder showed that bacterial genomes are in possession of large numbers of unique genomic segments that resemble horizontal transfer events, thus incorporation and synergistic usage of the latter approaches may be recommended as they could increase the efficiency of MGE detection. Table 3.4 has a list of the 15 hits obtained from the analysis ordered by the number of overlapping regions that are common between Prophinder and SWGIS predicted MGE . The first column of the table designates the id's of Prophinder prophages, the second column indicates the genomic elements with coding sequences similar to the ones found in prophages in corresponding rows, followed by columns of confidence scores presented as P-val (P value), E-val (E-value), and significance. *Bacillus subtilis subsp. subtilis str. 168* genomic element NC_000964 [2148850-2249149] of length 100299 bp (first row in Table 3.4) was found to be in possession of 155 coding sequences, and 154 of its 155 CDSs were found to overlap with some of the 186 CDSs that belong to a prophage

of id prophinder:45715. The region was assigned a significance score of 999.0, suggesting that it could be a prophage as it was found to be in possession of CDSs that match those of phage SPBc2. The last entry in the table (Table 3.4) (last row) listed as NC_004631 [4478900-4483699] of *Salmonella enterica subsp. serovar typhi Ty2* was determined by Prophinder to possess only 4 phage related CDSs that are similar to 4 of the 44 CDSs of prophinder:45433. These observations indicate that some of the regions that were identified by the both SWGIS Prophinder partly overlap whereas others overlap greatly. There are considerable differences in fragment sizes and numbers of coding sequences in genomic elements identified by Prophinder and SWGIS, indicating variations in results carried out by these methods.

Further analysis was carried out on *E. coli K - 12* [NC_000913] to show how combining the latter methods could be of use to increase the effiency of HGT events detection. *E. coli K - 12* was previously identified to be in possession of 11 prophages (Blattner *et al.*, 1997; Allison *et al.*, 2002; Casjens, 2003). Both SWGIS and Prophinder were used to scan this organism for mobile genomic elements. Results obtained from the methods were compared, and it turned out that Prophinder could only identify 5, whereas Seqword Gene Island Sniffer identified 8 phage sequences (Table 3.5), but failed to detect phages CP4-44, CPZ-55 and PR-X. SWGIS failed to identify the latter 3 prophages as they do not seem to show any deviance from the hosts OU signature pattern, which clearly means that their DNA composition is significantly uniform with that of the host. In this analysis, phage genomes were shown to be in posession of non-homologous compositional fragments, for example phage CP4-57 was identified by Seqword Gene Island Sniffer as 2 separate genomic determinants because other segments in this phage possess similar DNA features as the host. Prophinder was also deficient in detecting prophages CP4-44, CPZ-55 and PR-X, which possibly resulted from low sequence similarity and poor annotations since it is a problem that is suffered by most homology based methods. SWGIS identified 3 atypical phage genomic regions (KPLE2, CP4-6, CP4-57) which Prophinder failed

to identify. The latter serves as an indication of how a combination of composition-based and similarity-based methods could enhance one another in the detection of reliable phage sequences and other mobile elements, as they will both overcome limiting factors such as amelioration, poor sequence similarity and annotations. However, the accuracy of both methods may be affected by the fact that many phages tend to loose their morphology and thus loose their taxonomy characteristic features, whereas others are made up of genes that are of multiple ancestries which result in phylogenetic incongruencies and failure to classify.

3.4 Evolutionary and functional relationships between genomic islands

3.4.1 MGE protein families

Genomic islands are well characterized as DNA elements in genomes that resemble horizontal transfer, and contribute extensively to bacterial diversity and adaptation to different niches. But what is not fully understood is the mechanism involved in their mobility and exchange across different taxa. The latter results from lack of information on functional and evolutionary properties of genes that are hosted by mobile genetic elements. However, it is suggested that the study of relationships between genes from different genomic islands sequences may provide a better understanding on how genes segments get exchanged and mobilized between microorganisms. Therefore, this motivated a conserved module-based classification analysis of genomic elements which were identified in the study. The genomic elements were grouped into classes of families based on the hierachy of proteins that they have in

Table 3.4: Comparison of the MGE predictions obtained from Prophinder and SWGIS.

Prophinder id	**GI id**	**CDSs in prophages**	**CDSs in GI**	**CDSoverlap**	**P-val**	**E-val**	**Significance**
prophinder:45715	NC_000964 [2148850-2249149]	186	155	154	0	0	999.00
prophinder:44497	NC_006510 [535950-578449]	34	64	31	3.2e-90	8.4e-84	83.07
prophinder:46537	NC_006958 [1831500-1866749]	50	27	27	2.5e-85	6.6e-79	78.18
prophinder:45715	NC_000964 [2250550-2284399]	186	30	30	1.9e-74	4.9e-68	67.31
prophinder:43947	NC_000913 [1629150-1643899]	18	21	18	3.1e-66	8e-60	59.10
prophinder:42991	AC_000091 [1632850-1646399]	21	18	18	3.1e-66	8e-60	59.10
prophinder:44766	NC_007519 [1789150-1809799]	19	21	18	5.8e-65	1.5e-58	57.82
prophinder:45576	NC_003198 [1913250-1926849]	55	25	22	2.1e-64	5.4e-58	57.27
prophinder:45338	NC_006582 [1474100-1484249]	46	19	19	3.7e-60	9.6e-54	53.02
prophinder:42988	AC_000091 [562750-573599]	26	16	15	6.9e-51	1.8e-44	43.75
prophinder:45538	NC_006511 [2515150-2524799]	47	16	16	2.7e-50	7.1e-44	43.15
prophinder:43945	NC_000913 [572450-584899]	14	15	13	9.4e-49	2.5e-42	41.61
prophinder:42988	AC_000091 [572450-584899]	26	16	14	2.1e-46	5.5e-40	39.26
prophinder:45433	NC_004631 [4478900-4483699]	44	4	4	5.6e-13	1.5e-06	5.84

Table 3.5: *E. coli K12* prophages.

Number of phages: 11

Phage name	**Loci**	**Prophinder**	**SeqWord Gene Island Sniffer**
KPLE1	2464404 -2474619	2464567-2475651	2458800-2470399
KPLE2	4494108- 4534178	-	4497150-4505199
CP4-44	2064181 - 2077053	-	-
CP4-57	2753978- 2776007	-	2752450-2761549, 2767950-2774749
CP4-6	262182-296489	-	288000-295199
e14	1195443-1210646	1198902-1210402	1205450-1222299
Rac	1409966-1433025	1410024-1420753	1416850-1424699, 1428000-1434049
Qin	1630450-1646830	1631646-1643298	1629150-1588149
CPZ-55	2556791-2563350	-	-
PR-X	2165324-2166023	-	-
DLP12	564025-585326	572307-580602	562750-573599, 572450-584899

common and functional properties they possess. Measures of pairwise similarities of all the genomic element protein sequences were obtained by using BLASTP (Lima-Mendez *et al.*, 2008*b*). Proteins obtained from the search were clustered into families using a Markov clustering algorithm, by taking into account groups of proteins that share probable functional similarity (Enright *et al.*, 2002). The analysis was carried out as follows: an all *vs.* all sequence comparison was conducted on all the genomic islands CDSs, sequences that shared similarity were grouped into families and were annotated using ACLAME as the reference database (Leplae *et al.*, 2004). The compositions of genomic elements that constituted protein families were represented in the form of stochastic matrices, where rows represented genomic elements coding sequences and columns protein families (Lima-Mendez *et al.*, 2008*a*). A total of 9341 protein families were constructed. Upon the constructions, members of families were individually compared against each other and were then clustered into classes of MGE that share similar phylogenetic profiles (functional properties) (Lima-Mendez *et al.*, 2008*b*). The classes that shared functional properties were weighted and later assigned significance scores (SIG), designating their probable similarities in biochemical functions. However, classifications of these classes does not neces-

sarily imply significant sequence similarity. A class was denoted if SIG value for the group was bigger than 1.0. A total of 2,316 functional MGE classes were constructed. The majority of the classes appeared to be in possession of profiles such as integrases, transposases, transferases and lipopolysaccharide encoding enzymes that allow essential transfer, adherence and recombination of mobile elements with host chromosomes.

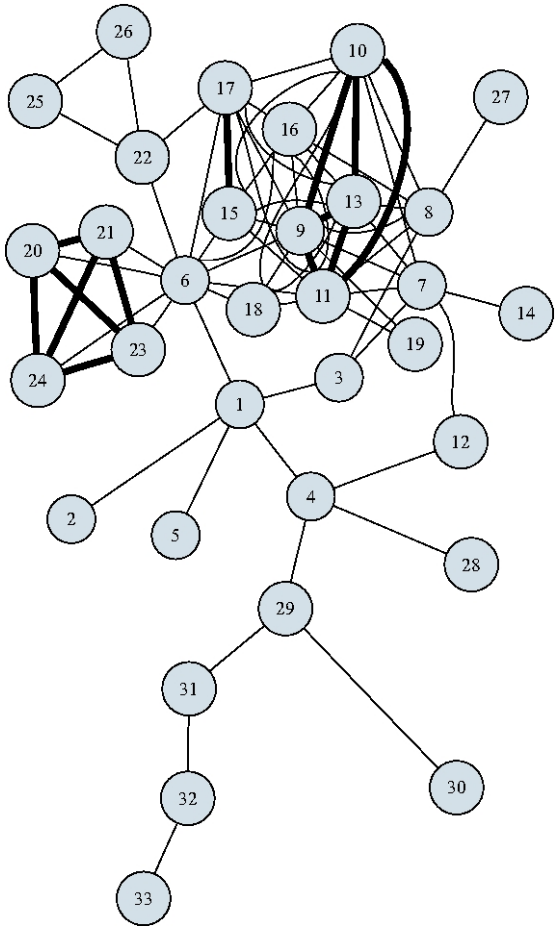


Figure 3.2: MGE class of functional properties that are shared among alpha, beta and gamma proteobacteria. The names of organisms that were used in the study are represented by circles. The names of the organims that as represented by the numbered circles are provided in Table 3.6, each circle number corresponds to the name of an organism. Organisms whose MGEs are similar in function are connected by lines. The MGEs that are connected with thin/dotted lines have significance values that are less than 10, and the MGEs that are connected by think lines have significance values that are beyond 50.

Table 3.6: Names of MGE in Figure 3.2

	Accessions	**Organisms**	**GEI coordinates**	**families**
1	NC_009092:15	*S. loihica PV-4 chromosome I*	[1602100-1635849]	family_4715
2	NC_004307:4	*B. longum NCC2705 chromosome I*	[1130000-1144849]	family_5706
3	NC_009052:15	*S. baltica OS155 chromosome I*	[3386600-3402399]	family_551
4	NC_006512:2	*I. loihiensis L2TR chromosome I*	[568200-593949]	family_7248
5	NC_007645:19	*H. chejuensis KCTC2396 chromosome I*	[2430200-2468149]	family_8534
6	NC_008570:11	*A. hydrophila subsp. hydrophila ATCC 7966 chromosome I*	[3225100-3265749]	family_3870
7	NC_007954:14	*S. denitrificans OS217 chromosome I*	[3177400-3199999]	family_3160
8	NC_007947:14	*M. flagellatus KT chromosome I*	[2142150-2171449]	family_9169
9	NC_003198:24	*S. enterica subsp. enterica serovar Typhi CT18 chromosome I*	[2113950-2136899]	family_619
10	NC_004631:5	*S. enterica Ty2 chromosome I*	[865850-887849]	family_619
11	NC_003197:24	*S. typhimurium LT2 chromosome I*	[2157200-2178449]	family_619
12	NC_009138:12	*H. arsenicoxydans chromosome I*	[1138600-1168399]	family_4810
13	NC_006511:5	*S. enterica ATCC9150 chromosome I*	[858550-886299]	family_619
14	NC_007908:5	*R. ferrireducens T118 chromosome I*	[1331500-1354549]	family_3160
15	NC_002695:28	*E. coli O157-H7-Sakai chromosome I*	[2772500-2788649]	family_6515
16	NC_004547:15	*E. carotovora SCRI1043 chromosome I*	[1605200-1635399]	family_6742
17	NC_002655:32	*E. coli O157-H7 EDL933 chromosome I*	[2842700-2858849]	family_1511
18	NC_007948:1	*P. sp. JS666 chromosome I*	[4184550-4221349]	family_9211
19	NC_006582:20	*B. clausii KSM-K16 chromosome I*	[3839000-3853549]	family_7350
20	NC_006932:4	*B. abortus biovar1 9-941 chromosome I*	[534700-557149]	family_1625
21	NC_003317:8	*B. melitensis 16M chromosome I*	[1447200-1468399]	family_5472
22	NC_007498:17	*P. carbinolicus DSM2380 chromosome I*	[2092800-2115349]	family_3825
23	NC_004310:2	*B. suis 1330 chromosome I*	[512200-535549]	family_1625
24	NC_007618:4	*B. melitensis biovar Abortus 2308 chromosome I*	[531600-553449]	family_644
25	NC_004431:28	*E. coli CFT073 chromosome I*	[2388300-2401549]	family_5783
26	NC_008253:20	*E. coli 536 chromosome I*	[2138500-2152299]	family_456
27	NC_007712:11	*S. glossinidius morsitans chromosome I*	[1606000-1623399]	family_8755
28	NC_007907:18	*D. hafniense Y51 chromosome I*	[3727750-3751699]	family_8980
29	NC_008820:1	*P. marinus MIT 9303 chromosome I*	[95500-145549]	family_4434
30	NC_005071:2	*P. marinus MIT9313 chromosome I*	[97750-125249]	family_6321
31	NC_005363:8	*B. bacteriovorus HD100 chromosome I*	[1608100-1637849]	family_6565
32	NC_006510:24	*G. kaustophilus HTA426 chromosome I*	[3138500-3155699]	family_7234
33	NC_007644:13	*M. thermoacetica ATCC 39073 chromosome I*	[785650-795899]	family_2984

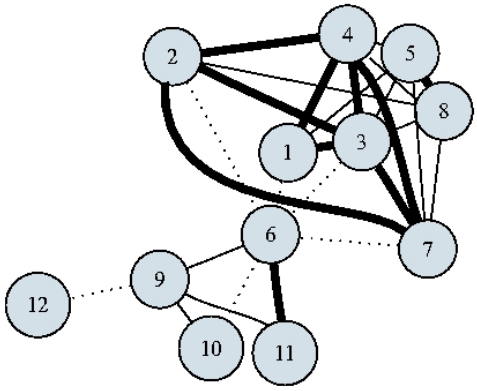
Three diagrams of MGE with similar functional profiles were constructed. The diagrams show classifications of functionally related MGE from similar and variable

evolutionary backgrounds. Figure 3.2 comprises of MGE of alpha, beta and gamma proteobacteria. MGE protein profiles are represented in the form of encircled numbers, the ones that share similar functional profiles are connected by lines. MGE that are connected by dotted lines designate significance values that are less than 10, the ones with thin solid lines designate significance values between 10 and 50, while those with thick solid lines designate significance values that are beyond 50. MGE with SIG values lower than 10 were removed from the diagrams for simplicity. In the figure (Figure 3.2), the bold lines appear to connect MGE that are of the same taxa, for example: MGEs 20, 21, 23 and, 24 represent strains of *Brucella*, MGEs 17 and 15 represent strains of *E. coli* and MGEs 9, 10, 11, 13 represent strains of *Salmonella*.

The latter could mean that the functional profiles harboured by these MGEs are completely similar, suggesting that they may be affected by similar evolutionary pressures and encode similar metabolic pathways. The MGE in the figure (Figure 3.2) share similarity in transposases such as IS711, ISGsu5 and ISPsy22 and enzymes such as O-antigen ligase, glycosyl transferase, ABC transporter (ABC-2 subfamily, ABC-type polysaccharide/polyol phosphate export systems) and rhamnosyl transferases that are involved in the biosynthesis and transport of polysaccharides. Similarity in profiles is also shared among MGE of different phyla, for example MGE 32 and 33 of *Geobacillus* and *Moorella* constitute a phyla and do not only share profiles among themselves. *Geobacillus* (32) shares similar profiles with MGE 31 of deltaproteobacter *Bdellovibrio*. Whereas, *Bdellovibrio* (31) shares functionality with MGE 29 of cyanobateria *Prochlorococcus,* that also relates to both MGEs 4 (*Idiomarina*) and 30 (*Prochlorococcus*) of gammaproteobacter and cyanobacteria. Figure 3.3 shows a classification of MGE that was acquired from *E. coli, Shigella* and *Salmonella* of the gamma-proteobateria.

These MGE are similar in profiles such as: type III secretion apparatus proteins, secretion system apparatus, secretion system effectors (sseF, sseD, sseA, sseG, sseL, sseM) and hypothetical proteins that most commonly occur in gram-negative pathogenic

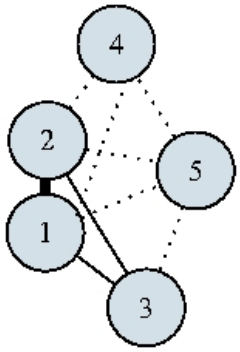
enterobacteria. Within this network, profiles with SIG values of over 50 seem to only be shared among MGE of the same species. Figure 3.4 constitutes of MGE from different *Salmonella* strains that share similarity in proteins that are involved in virulence traits, designated as *Salmonella* pathogenicity islands - SPI. The *Salmonellae* that make up this network are known to be intracellular facultative pathogens, known to cause diseases to both humans and animals (Hansen-Wester and Hensel, 2001). These MGE possess sets of proteins such as: putative outer membrane virulence proteins, putative toxin-like proteins, putative virulence proteins, putative lipoproteins and, putative bacteriophage proteins. Such proteins help pathogenic organisms to cope with the host's environmental conditions, and also with the production of toxins upon host cell invasion.



	Organisms	MGE coordinates	families
1	*Salmonella enterica Ty2 chromosome I*	[1328750-1356649]	family_826
2	*Salmonella enterica ATCC9150 chromosome I*	[1509200-1535549]	family_826
3	*Salmonella typhimurium LT2 chromosome I*	[1473850-1502249]	family_1615
4	*Salmonella enterica SC-B67 chromosome I*	[1511150-1539999]	family_7480
5	*Escherichia coli O157-H7-Sakai chromosome I*	[4587750-4625299]	family_6517
6	*Escherichia coli O157-H7 EDL933 chromosome I*	[3772500-3804399]	family_5019
7	*Salmonella enterica subsp. enterica serovar Typhi CT18 chromosome I*	[1622850-1650099]	family_826
8	*Escherichia coli O157-H7 EDL933 chromosome I*	[4656800-4694799]	family_2381
9	*Escherichia coli W3110 DNA*	[2981450-2997049]	family_1516
10	*Escherichia coli K12-MG1655 chromosome I*	[2981400-2995799]	family_1076
11	*Escherichia coli O157-H7-Sakai chromosome I*	[3705200-3736649]	family_2357
12	*Shigella sonnei Ss046 chromosome I*	[3154100-3168049]	family_1075

Figure 3.3: MGE class of functional properties that are shared among *Salmonella* and *E. coli*. The names of the organims that are represented by the numbered circles are provided in Table belowthe diagram. Each circle number corresponds to the name of an organism. Organisms whose MGEs are similar in function are connected by lines. The MGEs that are connected with thin/dotted lines have significance values that are less than 10, and the MGEs that are connected by think lines have significance values that are beyond 50

The network in Figure 3.3 illustrates that the virulence encoding MGE that are shared among the *Salmonella* in the group cannot be easily transferred to organisms of other genera except for *E. coli*. They tend to only be confined to *Salmonella*, hence the name SPI. Previous analysis also showed that *Salmonella* harbor virulence traits that set them apart from other species of enteric bacteria (Groisman *et al.*,



	Organisms	MGE coordinates	Families
1	*Salmonella enterica Ty2 chromosome I*	[1186900-1204199]	family_6028
2	*Salmonella enterica subsp. enterica serovar Typhi CT18 chromosome I*	[1774100-1791899]	family_2498
3	*Salmonella enterica ATCC9150 chromosome I*	[1658650-1675199]	family_2497
4	*Salmonella enterica SC-B67 chromosome I*	[1347150-1368499]	family_7479
5	*Salmonella typhimurium LT2 chromosome I*	[1322950-1337449]	family_5399

Figure 3.4: MGE class of functional properties that are shared among *Salmonella.* The names of the organims that are represented by the numbered circles are provided in Table belowthe diagram. Each circle number corresponds to the name of an organism. Organisms whose MGEs are similar in function are connected by lines. The MGEs that are connected with thin/dotted lines have significance values that are less than 10, and the MGEs that are connected by think lines have significance values that are beyond 50

1993; Mills *et al.*, 1995; Ochman and Groisman, 1996).

3.5 BLASTN

Early sequence comparisons of MGE revealed that they are genetic mosaics, where regions with sequence similarity alternate with unrelated regions (Lima-Mendez *et al.*, 2008*b*). To study the networks of MGE sharing similar DNA sequences, the BLASTN algorithm was used, where MGE individual fragments were compared against one another in a pairwise alignment fashion to search for nucleotide fragments that share similarity. The BLASTN pairwise aligned regions were computationally processed in a way that the overlapping or adjacent segments longer than 100 bp were fused into long regions and stored to the database along with the information of all the

counterpart DNA fragments obtained from other MGE that share sequence similarity. MGE with similar sequence features obtained from the search were assembled into groups in association with their corresponding annotations retrieved from the NCBI database. The similarity searches were conducted in order to group genomic islands according to the compositional features that they have in common and to also allow evaluations and inferences on homology between MGE obtained from varying bacterial lineages.

3.5.1 BLASTN groups

Following the all-against-all MGE BLASTN searches, 7302 BLAST matching regions were obtained from 1570 of the 3518 identified MGE. A total of 1328 groups were created from all the obtained matching regions and stored in genbank flat files (GBFF) that are available for download from ftp://milliways.bi.up.ac.za/SeqWord/MGE/ . In average 2.08 matches were obtained per MGE. The average length of sequences that were obtained from the combined overlapping and adjacent fragments was 1865 bp, acquired from MGE of lengths 100 to 40248 bp. The files were sorted according to the total number of MGE that each group entail. The largest group: Group #1, comprises of 2648 MGE hits from 31 genera. The MGE in the group share similarities in genes that are involved in a variety of functions, such as transposable elements and membrane proteins involved in the motility of gene fragments and bacterial virulence.

The other sets of genes shared in the group are characterized as either hypothetical or putative. Some of the most often occurring ones are: putative inner and outer membrane proteins, transposases and IS elements. Sets of such genes are the most shared within the group, yet the fact that they are uncharacterized, deprive us of the understanding of the functional characteristics that are conserved in and between MGE. However, many MGE in the group show a significant large number of sequence matches with *E. coli*, comprising a total of 1089 hits followed by *Salmonella* with

757 hits that probably resulted from a biased overrepresentation of these organisms among completely sequenced bacteria. The MGE that constitute this group belong to organisms of diverse evolutionary backgrounds, such as: alpha, beta, delta and gamma Proteobacteria and Chlorobi (see Figure 3.5). The genes that are shared among MGE whose genes are assigned functions are essential in bacterial virulence and the transfer of genomic segments among species. Among the latter are outer membrane porin proteins that enable gram negative bacteria to cope with variable environmental conditions, and allow solute diffussion through the pores that form on outer cell membranes.

These proteins are also involved in bacterial invasiveness and the modulations of pathogen-host cell interactions during the early stages of virulence (Massari *et al.*, 2003). Within the group, porins and the other outer-membrane proteins appear to only be conserved and shared among *E. coli* and *Shigella* species. The MGE of *Shigella*, *E. coli* and *Salmonella* tend to be the most that are in possession of a wide range of genes known to be involved in different stages of pathogenicity. They encode fimbria that allow bacteria to adhere to host cell surfaces; invasion protein for invading host cells; toxin subunits for toxicity, and lipoproteins to confer resistance to bactericidal effects and survival within phagolysosomes. The latter organisms also share other virulence associated factors such as UDP-glucose/GDP-mannose dehydrogenase and UDP-glucose 6-dehydrogenase with organisms from other backrounds such as: *Aeromonas, Pseudomonas, Polaromonas, Shewanella, and Marinobacter*, indicating that transfer of genes that encode pathogenic traits do also occur across species borders. The transposition insertion sequences such as: is1 orf2, is600 orf2, is1 orf1, is2 orf2, is600 orf1, is911 orf2, is2 orf1, is1 transposase InsAB, is1 protein InsB and is2 OrfB protein that are widely known as elements that take part in gene translocations also occured in great numbers in these MGE. The latter elements were identified in *E. coli, Salmonella* and *Shigella* only, illustrating that the frequency at which these elements get transferred is low, and that their spread is only restricted

within a genus.

Most other MGE sharing similar sequences are of the *Neisseria* genus in group #2. The MGE of this genus are composed of 180 BLAST matching regions that comprise virulence associates such as FRPC, MAF adhesin proteins. Among some of the genes that are present in the group are insertion sequence elements IS1016c2 transposases that are mainly found in proteobacteria and also remnants of IS11060a3 first described in variants of subgroup III serogroup A *Neisseria meningitidis* (Tzeng *et al.*, 2003; Zhu *et al.*, 2003). Group #3 is also composed of MGE that only show sequence similarity between *Salmonella, Shigella, Escherichia* and *Erwinia*. The functional genes of these MGE are the membranous proteins that can only be found in gram negative bacteria, and are essential in the lipopolysaccharides biosynthesis, and evasion of host cell immune responses during virulence.

3.5.2 Exchange of laterally acquired gene islands of group#1 between genera

The schema in Figure 3.5 is a representation of the gene stock exchange (GSE) of genetic materials that occurs across microbial lineages. The relational lines that connect bacteria in the MGE network diagram represents the number of BLAST hits between MGE from genomes of different species. Organisms with the largest number of matching sequences (>50 matching regions) are represented by thick solid lines, whereras those with fewer than 10 matching regions are represented by dotted lines. Figure 3.5 indicates the intricacies of gene transfer within the microbial biosphere of different phlya. Organisms of the gamma Proteobacteria such as: *Salmonella, Idiomarina* and *Hahella* share compositional similarity in DTDP-glucose 4,6-dehydratase genes, known to be involved in different types of molecular pathways such as: nucleotide sugar metabolism and polyketide antibiotic biosynthesis

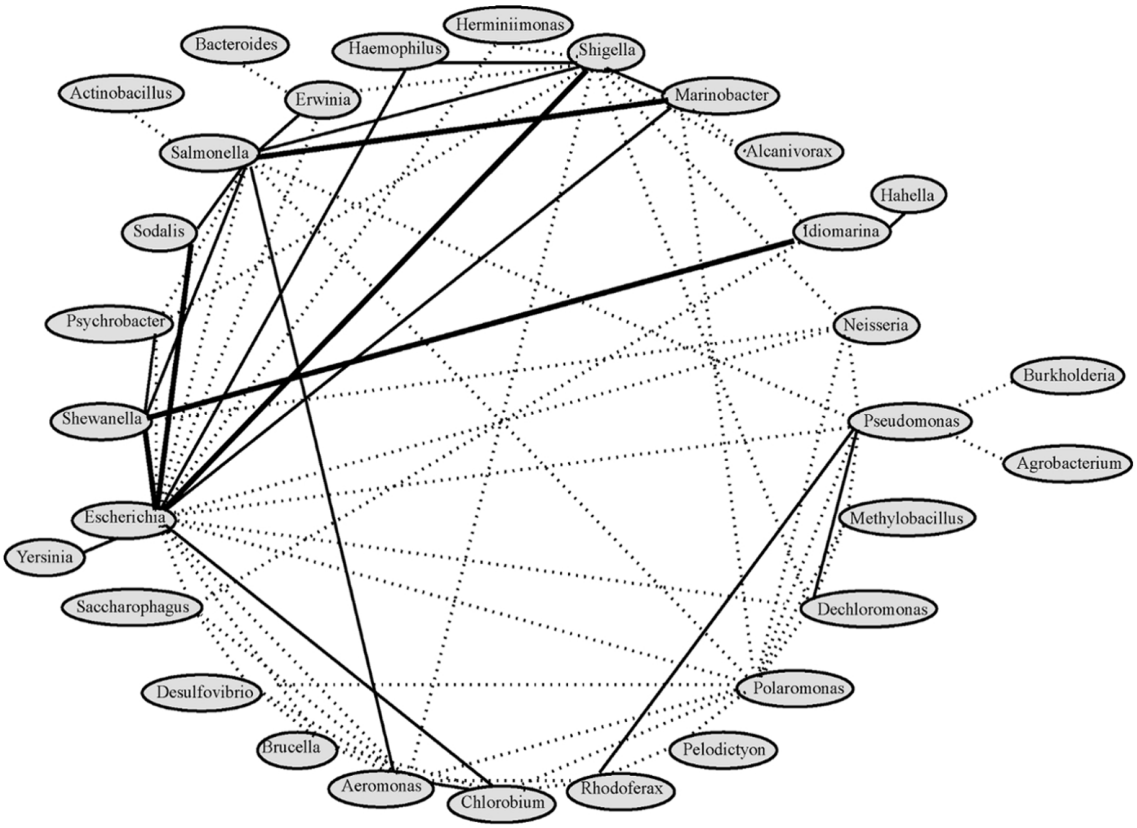


Figure 3.5: BLASTN gene exchange network of organims of variable evolutionary backgrounds.

(Hashimoto *et al.*, 2006). It is illustrated in the figure (Figure 3.5) that *Shewanella* shares compositional features with *Idiomarina* and *Idiomarina* with *Hahella* and *Alcanivorax,* and all are marine bacteria. *Hahella* is shown to have similar features with *Idiomarina* only, and not all the other MGE in the network. Furthermore, *Pseudomonas* of class gamma Proteobacteria was also shown to exchange DNA sequences with diverse organisms such as *Rhodoferax, Dechloromonas, Burkholderia, Neisseria, Methylobacillus* of betaproteobacteria, and also *Agrobacterium* of alpha Proteobacteria, collectively the latter organisms inhabit habitats such as soils and plants and yet they harbour similar OU patterns of genomic sequences.

3.5.3 Group#1 MGE phylogenetic inferences

Upon classifying MGE into groups using the BLASTN method (above), further analysis was conducted to reconstruct and further study the evolutionary history of protein sequences shared among organisms of group 1 (previous section), because clusters of proteins that share similar sequence patterns may define families that perform common biochemical functions or share common evolutionary histories. Therefore, a total of 8513 MGE proteins were extracted, and were similarity searched against one another using BLASTP. The largest of BLAST outputs was found to be in possession of 280 hits comprising of *E. coli*, *Shigella*, *Salmonella* of enterobacteria, sharing similarity in transposases and insertion elements. In this instance groups were selected not according to the number of hits that each entail but according to the number and variety of organisms that each group constitutes. Thus, only 3 groups were selected, comprising of Chlorobi, alpha, gamma and beta Proteobacteria. The MGE protein sequences of each group were multiply aligned using CLUSTALX (Thompson *et al.*, 1997) to group pairs of sequences that share similar compositions together, the alignments were then saved in PHYLIP (J, 1993) output format. The latter was followed by the construction of phylogenetic trees (using PHYLIP). The protein sequence maximum likelihood method with molecular clock parameters were used for a phylogenetic reconstruction, as it is normally used to measure rates of molecular evolution and amino acids substitutions among diverse taxa.

3.5.3.1 Findings

Upon the BLASTP analysis that was performed above (subsection 3.5.3), a tree topology was constructed and viewed with dendroscope (an interactive tool for viewing large phylogenetic trees and networks). These trees were specifically selected because they showed consistency in occurrences of different groups of clades that constitute proteobacteria of variable phyla, illustrating the frequency of gene exchange be-

tween organisms. This obervation also shows that the proteins that are shared among these MGE are preferentially transferred during gene exchange events, for proteins of their kind tend to also appear in great numbers in the analysis that were performed in the above sections. Figure 3.6 represents a tree that was constructed from a BLAST group that constituted transferases of alpha, delta, beta Proteobacter and Chlorobi that showed similarity with glucose-1-phosphate thymidylyl transferase-rfba of *E. coli* AC_000091:23. The glucose-1-phosphate thymidylyl transferase-rfba is a gene that is involved in the biosynthesis of surface polysaccharides and lipopolysaccharides in pathogenic bacteria that mediate virulence and host tissue adhesion. Among the genes that shared similarity with the latter were glucose-1-phosphate transferase-rmlA, TDP glucose pyrophosphorylase-rfbA, mannose-1-phosphate isomerase, mannose-6-phosphate guanylyl transferase, TDP-rhamnose synthetase and GDP-mannose 4,6-dehydratase. Most of these genes are also involved in encoding virulence traits for bacteria, for example GDP-mannose 4,6-dehydratase serves as an enzyme that converts GDP-mannose to GDP-L-fucose, a product that acts as a precursor for surface antigens such as extracelluar polysaccharides in bacteria. Collectively, the latter genes possess catabolic functions that lead to formation of new metabolic pathways that help bacteria to cope with certain environmental conditions, thus these genes are likely preferred in mechanisms of HGT. The tree in Figure 3.6 illustrates the grouping of proteins from various organisms, for example the cluster 1 shows the grouping of gamma (*Salmonella, Erwinia* and *Aeromonas*) and beta Proteobacteria (*Polaromonas*), followed by the cluster 2 that only constitutes of gamma-proteobacteria. Species that are of the same phylogenetic background (family) mostly appear to have branches of same lengths suggesting similar evolutionary substitution rates. The cluster 3 shows an interesting relationship shared among chlorobi (*Chlorobium* and *Pelodictyon*) and betaproteobacteria (*Rhodoferax*), it is unexpected for *Rhodeferax* and *Pelodictyon* to form a clade, they appear to be more closer than *Chlorobium* is to *Pelodictyon.* Moreover, *Rhodoferax* and *Pelodictyon*

share the same branch lengths, also suggesting a similar evolutionary rate.

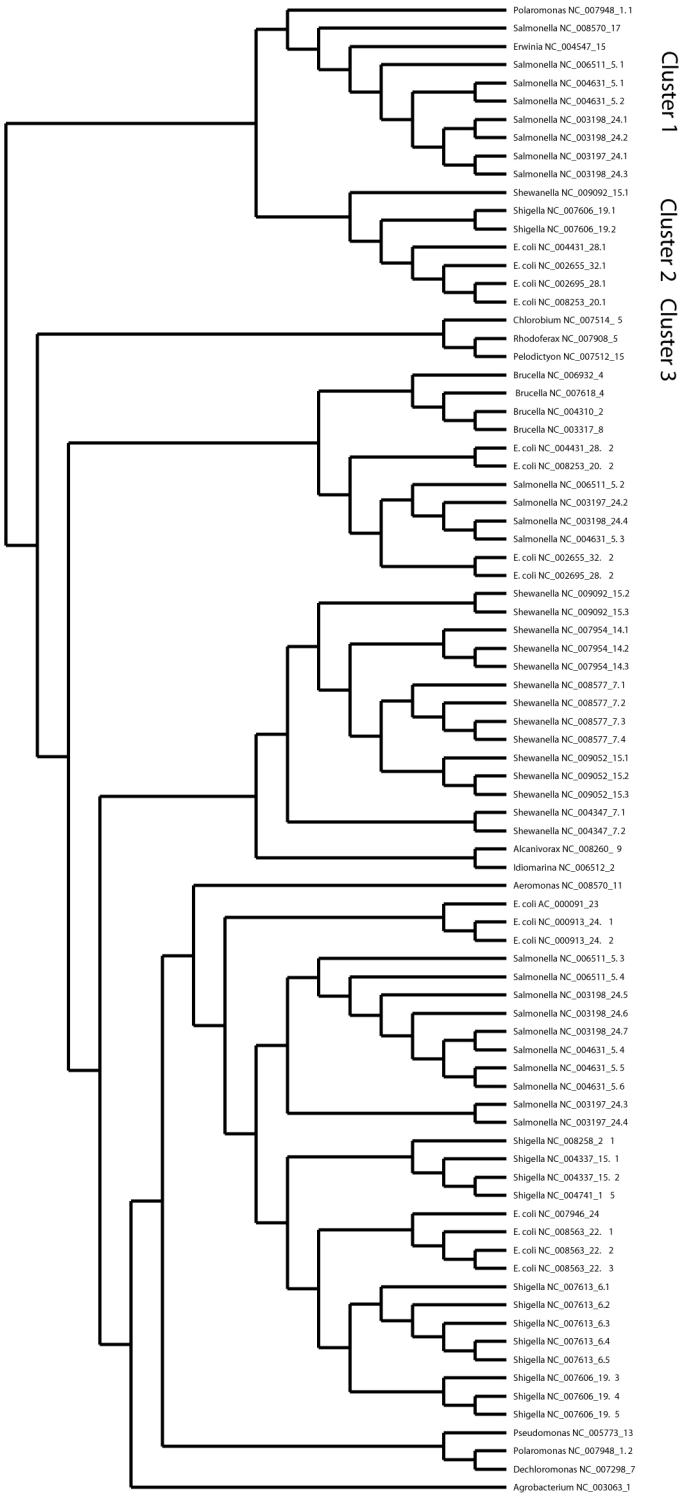


Figure 3.6: A tree of MGE proteins that share amino acid similarity with glucose-1-phosphate thymidylyltransferase (rfbA) of *Escherichia coli W3110* [2110408-2114599] /AC_000091:23 (AC_000091:23 resemble *Escherichia coli W3110* genomic island number 23 as in the GEI-DB database).

3.5.4 Determination of gene order conservation in SWGIS MGE

Further MGE analysis was conducted in order to study the features and evolutionary relationships of their gene entities. The analysis was carried out by comparing *Brucella melitensis 16M chromosome I* MGE [1447200 - 1468399] with MGEs of other *Brucella* strains and those of organisms outside the *Brucella* family. This was conducted to measure evolutionary relationships between various MGE based on the order and conservation of genes that each entail. *Brucella melitensis 16M* was selected for the analysis because it harbours genes that are essential for virulence and persistence within hosts cells (Zygmunt *et al.*, 2006). It belongs to the *Brucella* spp. which are gram negative facultative intracellular bacteria that induce abortion in wild stock and undulant fever in humans (DelVecchio *et al.*, 2002). The *Brucella* spp. harbour virulence determinants that essentially allow defense against microbiocidal mechanisms and therefore influence microbial survival and persistance within macrophages (Godfroid *et al.*, 1998; Zygmunt *et al.*, 2006). The *Brucella melitensis 16M chromosome I* MGE [1447200 - 1468399] possesses genes that encode lipopolysaccharide (O-antigen), GDP-mannose 4,6-dehydratase, mannosyltranferases and transposases, known to be major determinants involved in brucella pathogenecity (Godfroid *et al.*, 2000) that also contribute in manipulations of the host cell immune responses. BLAST analysis of the *Brucella* MGE [1447200 - 1468399] with a cut-off E-value of 0.00001 was conducted, and showed that it shares significant sequence similarities with MGE of other *Brucella* strains and also *Escherichia coli 0157:H7* [2772500-2788649], *Escherichia coli 0157:H7_EDL 933* [2842700-2858849], *Chlorobrium chrolorochromatii CaD3* [914300-950899], *Desulfovibrio desulfuricans G20* [2859700-2891149], and *Aeromonas hydrophilia ATCC 7966* [3225100-3265749].

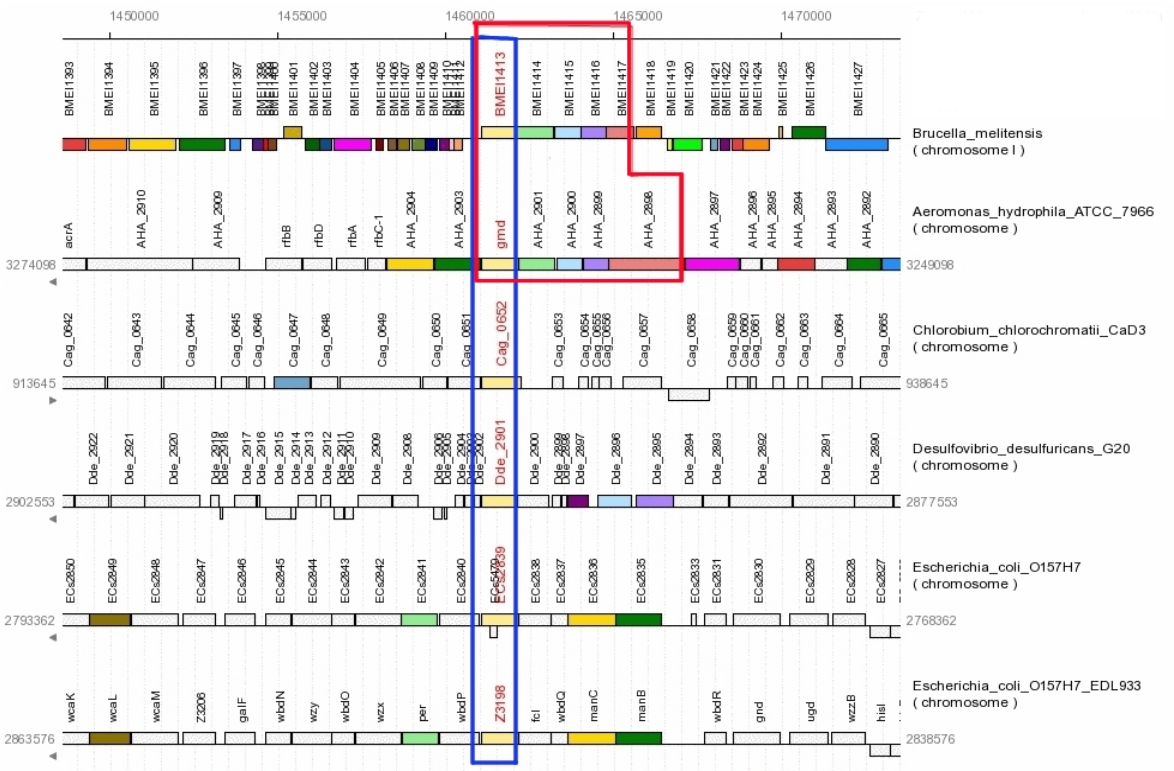


Figure 3.7: Gene order conservation in *Brucella* and *Aeromonas*. The conservation of gene order analysis was performed with PSAT - Prokaryotic sequence homology analysis tool Fong *et al.*, 2008. The figure is generated after the genes between the genomes of study are compared against each other Fong *et al.*, 2008.

These MGE were retrieved from the search and further analysed to check if the entities that they entail share any evolutionary relationship with those of *Brucella melitensis 16M chromosome I* [1447200 - 1468399]. This was conducted using PSAT (Prokaryotic sequence homology analysis tool) (Fong *et al.*, 2008), a web tool that performs searches for the conserved neighbouring genes among multiple prokaryotic genomes and displays the results in the form of a visual context (Figure 3.7). The order of genes displayed in Figure 3.7 indicate that these MGE share one common gene (included in the blue block). Thus, having one gene in common does not provide enough evolutionary history shared among these organisms. However, *Aeromonas hydrophilia ATCC 7966* appears to be the only MGE that posessess over two genes

that are common to those of the *Brucella* spp. It shares five genes with *Brucella melitensis 16M,* that are clustered in the same order (Figure 3.7, included in the red block). The enlightening part is that these organisms belong to different lineages, yet they have a shared synteny that implicate common regulatory mechanisms (Table 3.7) (Tamames *et al.*, 1997).

The evolutionary history of gene order in *Brucella* and *Aeromonas* is not known, however their synteny may indicate either xenology or orthology. Gene order conservation in bacteria of different lineages cannot always be referred to as: acquisition from a common source, as the order may be a result of factors such as different evolutionary pathways, gene acquisition by different transfer mechanisms or similar selection processes (Tamames, 2001). Apart from the conservation of order maintained by the five genes, there are several other genes (Figure 3.8) that are common in both MGE, but the ones that are harboured by *Aeromonas* appear to have been rearranged, to possibly increase the functional and fitness traits of the *Aeromonas* genome (Figure 3.8 (B)). The rearrangement may have occurred upon deletion of several other genes, as it appears to have lost quite a number of transposases in reference to *Brucella* (illustrated in Figure 3.8(A)). It is widely known that conservation of gene order is well maintained and preserved in species that are related and it is thus used as a measure to study evolutionary relationships between organisms as revealed in Figure 3.9. Thus, the factors that keep clusters of genes into functional units among related and unrelated species are still not fully understood (Tamames, 2001). However, the latter observations serve as an indication that horizontal gene transfer is a universal event, and also indicates that the genes that are shared or harbored by MGE are not a result of random acquisition.

Table 3.7: *Brucella melitensis 16M chromosome I* genomic island [1447200 - 1468399] and *Aeromonas hydrophila ATCC 7966* statistics.

***Aeromonas*[3225100-3265749]**		***Brucella 16M*[1447200-1468399]**	
AHA_2894	Glycosyl transferase, group 1	BMEI1404, BMEI1393	Mannosyltransferase, Mannosyltransferase C
AHA_2897	Glycosyl transferase, group 1 family protein	BMEI1404	Mannosyltransferase
AHA_2898	Putative glycosytransferase	BMEI1417	Perosamine synthetase WBKB
AHA_2899	RfbE, O-antigen export system ATP-binding	BMEI1416	O-antigen export system ATP-Binding protein RFBB
AHA_2900	ABC transporter, permease protein 33039	BMEI1415	O-antigen export system permease protein RFBD
AHA_2901	Perosamine synthetase, Per protein 34870	BMEI1414	Perosamine Synthetase
AHA_2902	GDP-mannose 4,6-dehydratase 35841	BMEI1413	GDP-mannose 4,6-dehydratase
AHA_2903	Phosphomannomutase	BMEI1396	Phosphomannomutase
AHA_2904	Mannose-1-phosphate	BMEI1395	Mannose -1-phosphate Guanylyl Transferase

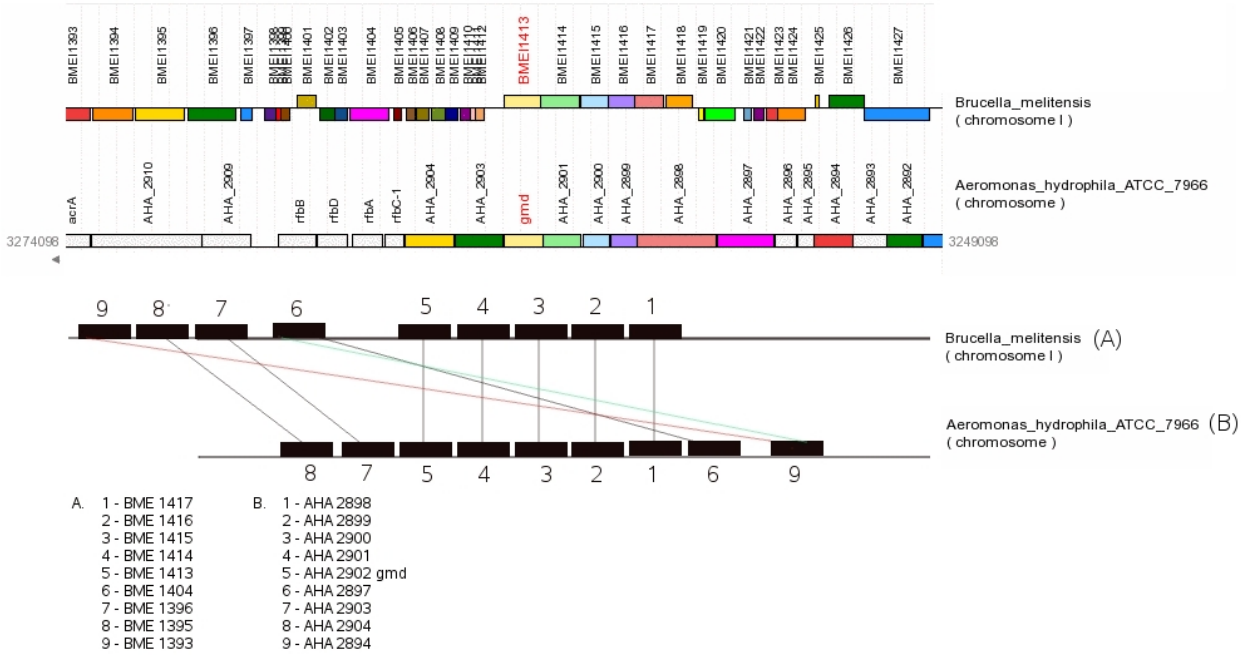


Figure 3.8: Rearrangement of *Aeromonas* genes. The genes that are common between *Brucella* and *Aeromonas* are represented with the same colours and are arranged in the same order. The figure shows the conservation in order between genes that are found in GIs of both *Aeromonas* and *Brucella*.

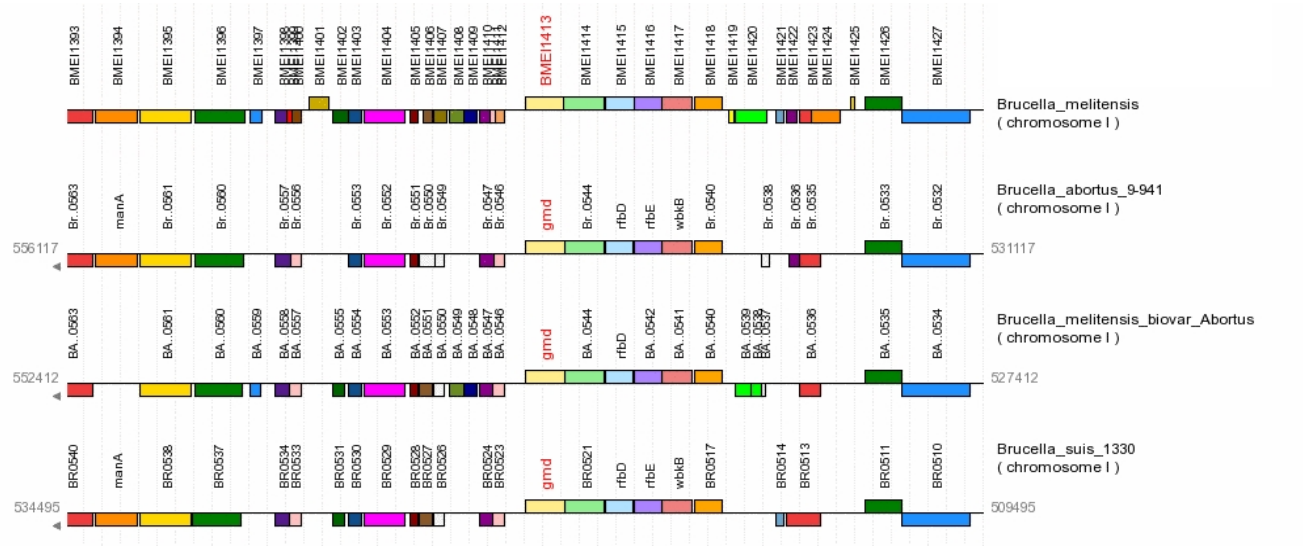


Figure 3.9: Preserved gene order in *Brucella*. The figure illustrates the conservation of gene order that is maintained in the GIs of *Brucella*.

3.6 Conclusion

The results obtained from the above analysis provide insights in the different evolutionary aspects of mobile genomic elements. They are an illustration of the many constraints that are hidden within genomic islands, including their diversity. Comparisons of elements on the basis of DNA and protein compositional parameters reveal the conservation, evolutionary significance and functional properties that they entail. The entities that are harboured by mobile elements showed that they are not just products of random acquisitions. Suggesting that MGE have specific sets of genes that they require to evolutionary advance organisms.

Chapter 4

MetaLingvo: Algorithm for OU pattern similarity search

4.1 Background

Comparative analysis methods have illustrated that many microbial genomes contain gene entities that they have acquired through lateral transfer events from other organisms. The acquisition of such genes is regarded as an evolutionary factor that alters the bacterial biochemical functions and also allows them to adapt to different enviroments. However, the oligonucleotide composition of these laterally acquired genes vary throughout the genome, resembling base compositions of different donor genomes. Several computational methods were developed based on the analysis of gene portions with compositions that deviate away from those of the core genes in a given genome. Most of these methods studied variances of gene compositions by searching for genes with variable GC content and unusual codon or amino acid usages. Developments of several other methods based on the preferences of oligonucleotide usage (OU) patterns by bacteria followed, as the latter ones could not discriminate between species. The OU usage concept was first introduced by Karlin

(1998), and has since been widely used. The ou concept showed that more closely related microbial species are made up of a common signature pattern. Therefore, the base compositions of organisms can be distinguished according to their OU patterns. The base compositions that are exhibited by laterally acquired genomic islands are believed to resemble those of their donor genomes. Thus by computing the associations in OU patterns of various organisms, the putative donors could easily be identified as genomic signature is species specific. Similar methods are used in several metagenomics projects, as an attempt to assign environmental reads to their specified taxa and predict their source genomes. Most such methods use similarity based approaches, that group or perform binning of reads based on the known sequences that are available in public databases. Thus a read or reads sharing similar composition with the organism that is available in the database is assigned to that paricular organism. In most cases a sample of interest is collected from an environment and gets sequenced to obtain reads, these reads are then blasted using the BLAST algorithm against NCBI databases to identify their source organisms. The latter method has its own limitation as it is mainly dependent only on the organisms that are available in the database. Therefore such limitations have been complemented by development of composition based methods. In this study, MetaLingvo, an algorithm for OU pattern similarity search was developed. MetaLingvo attempts to identify putative donors of genomic islands by the classifications of genomes with the closest OU patterns.

4.2 MetaLingvo

The MetaLingvo algorithm was developed in the Python programming language. Its easy to use graphical interface was implemented in TurboGears, a Python-based framework that allow the rapid development of web applications. MetaLingvo is

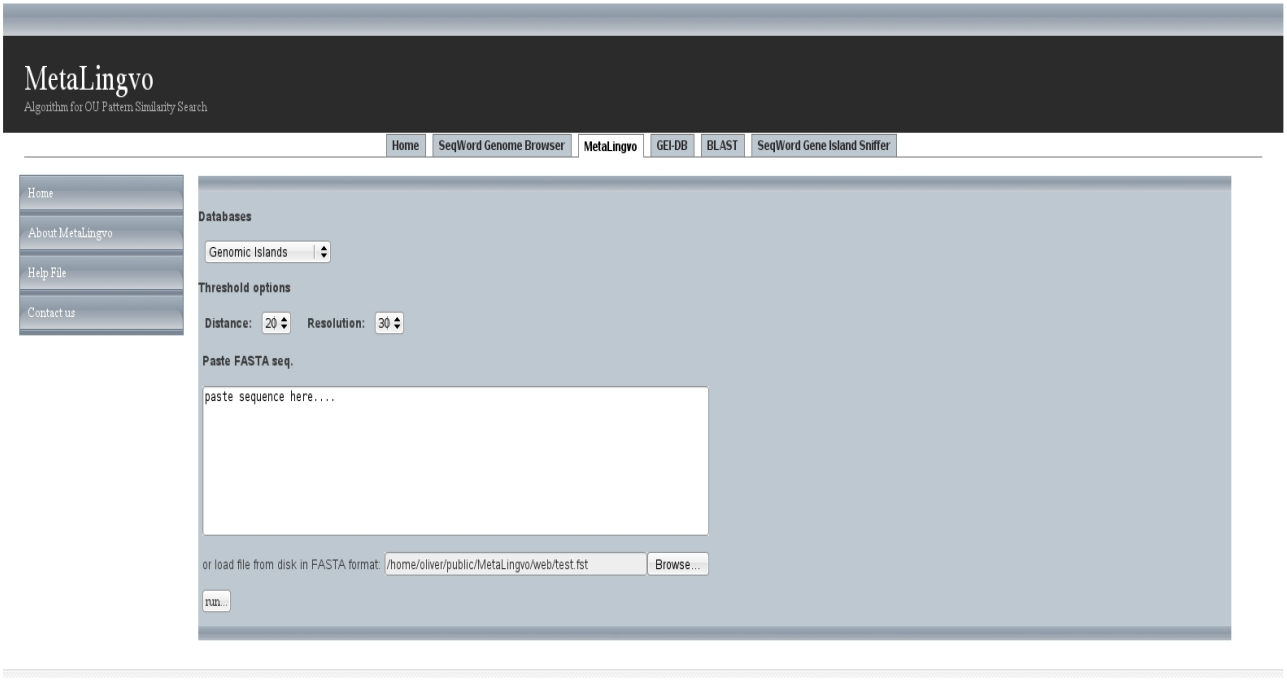


Figure 4.1: Graphical user interface of MetaLingvo: Algorithm for OU pattern similarity search.

composed of different databases made up of oligonucleotide usage patterns that were calculated for genomic islands, plasmids, bacterial chromosomes and bacteriophages. The databases of plasmids and bacteriophages help to illustrate the mechanisms in which some of the fragments were dispersed in microbial communities, especially in cases of lateral transfer events where they act as major players of distributing entities across organisms that are not related. MetaLingvo takes as input bacterial sequences in a FASTA format. Upon supplying it with the file of a query sequence, the sequence gets broken down into fragments of OU patterns that are compared with the ones available in the databases to determine the source organism. Since this work is mainly focused on mobile genomic elements the MetaLingvo genomic islands database was widely used to identify putative donors or source organisms of most elements and also groups of genomic islands sharing similar patterns. Figure 4.1 illustrates the MetaLingvo web interface. The page offers two options of loading sequences prior to analysis, users can either paste their query sequences in the pro-

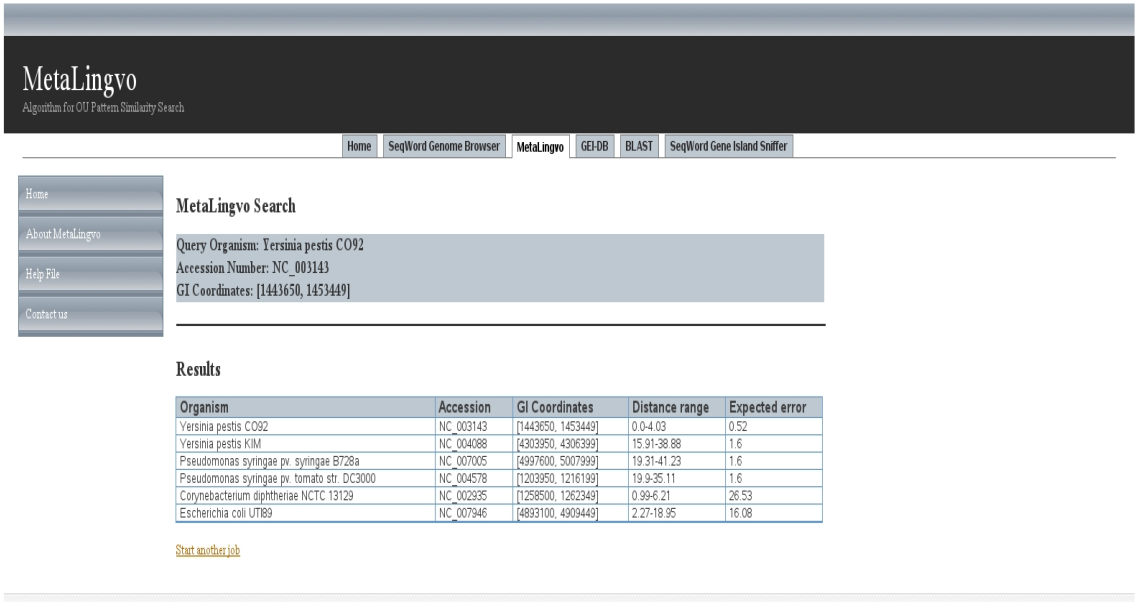


Figure 4.2: MetaLingvo output of GIs that share tetranucleotide pattern similarity with *Yersinia pestis* GI [1443650-1453449] .

vided text area window or load them as files from disk. Several parameters are to be set before execution, the database to perform a search against has to be selected, followed by selecting the preferred values for the distance and resolution required for the search. The distance values determine the stringency of the searches and comparisons performed between patterns of different organisms. The resolution is the standard deviation threshold that is used by the algorithm during the search as a filtration method that avoids ambiguities of correlating clusters. Upon the setting of parameters, the run button is pressed to allow the execution of the analysis. Figure 4.2 presents the output of results that are obtained from the search, with names of putative source organisms together with values that illustrate the significance of the results, designated as distance range and expected error.

Yersinia pestis CO92 genomic region [1443650-1453449] was used as a test template. The genomic region was similarity searched against the genomic islands database of

MetaLingvo using a distance value of 20 and a resolution of 30. The same region from the same organism designated *Yersinia pestis CO92* genomic island [1443650-1453449] appeared to be the first hit obtained from the search, accompanied by distance values ranging between 0.0-4.03 illustrating that these patterns are the same and less apart from each other. The distance values are the determination of how similar or how far apart the sequence patterns are from one another, the lesser the distance values of patterns the more closer they are to each other. Supporting the latter is also the Expected error value of 0.52. The significances of the results are also determined by the expected error values that are closer to zero. Following *Yersinia pestis CO92* are genomic islands of *Yesrinia pestis KIM, Pseudomonas, Corynebacteria* and *E. coli* with varying values, each of which could be a potential donor of the genomic island that is harboured by *Yersinia pestis CO92* or it could either be that these organisms share a common of mobile genetic elements. The distance values of *Pseudomonas* and *Yersinia KIM* are high as compared to other, but their expected error values are significantly acceptable, as compared to those of *Corynebacteria* and *E. coli* even though their distance values are less. Although these organisms have different values of expected error and distance they do cluster together in terms of the ou patterns that each is made up of.

4.3 Conclusion

Several similarity based methods that are dependent on BLAST and also the availability of known sequences in public data repositories have been developed. These methods suffer several drawbacks as they only rely on sequences that have been deposited into public databases, such as the NCBI since it is the one that is mostly used. The developments of composition based methods have come in handy as they outperform the latter methods. Such methods do not require any public databases to search against and their use has been shown to be more accurate as compared

to similarity based methods by just clustering organisms according to their genomic signatures. The use of OU patterns offers a valuable tool as they have also been widely applied in detections of foreign genomic elements in the bacterial world.

Chapter 5

Concluding discussion

Horizontal transfer plays a pivotal role in the evolution of organisms across all kingdoms of life. The concept of gene exchange among organims that are not of the same genera was rejected in the past, until the emergence of incongruencies that were observed in phylogenetic tree topologies. Comparative analysis methods also made it apparent that horizontal transfer is a continual event that occurs in both prokaryotes and eukaryotes. Studies of horizontal transfer events are increasingly carried out on prokaryotes as they have previously been found to possess a complexity of mobile fragments that are made up of virulence, fitness, drug resistance, and symbiosis genes, collectively known as mobile genetic elements (MGE). These elements can be beneficial to bacteria by changing their lifestyle and may also be detrimental to mankind. A variety of methods have been developed to identify MGE in prokaryotic genomes, and most of them involve the study of phylogenetic relationships and tracing down inconguent lines among species lines, and also methods that search for fragments with unusual codon and GC%. The uses of GC content and codon usage bias fall among the most simplest methods used to compare genomes between different bacteria. Although GC content influences codon usage patterns of many bacteria, they cannot reliably serve as phylogenetic signals between lineages. It was

previously thought that bacteria from different lineages could be classified according to their GC composition. However, it has recently been shown that factors such as temperature and environment have a serious impact in organism's GC content distribution. Closely related organims that have been exposed to different environmental conditions show marked differences in their GC content, indicating that GC does not necessarily serve as a phylogenetic signal (Chen and Zhang, 2003; Foerstner *et al.*, 2005). The usage of oligomers in studying distributions of genomic patterns that are of potential horizontal origin provides an essential method that reliably measures variances between different sets of genes in genomes of variable lengths. Yet, previous studies indicated that use of tetramers for the study of oligonucleotide usage distributions in bacteria is sufficient to uncover constraints that are distintive in genomes and between genera. Thus, in this study a novel tool (SeqWord Gene Island Sniffer) that examines and detects variances in frequencies of oligonucleotides to efficiently trace down the distributions of mobile genomic elements across genomes by patterns of 4-bases long words was developed. SeqWord Gene Island Sniffer is a computational tool for an automated identification of MGE in bacterial and plasmid DNA sequences. The method of tetramers instead of GC and codon bias was preferrably used in SeqWord Gene Island Sniffer (SWGIS) as our previously published methods (Reva and Tummler, 2004, 2005) based on the similar principle made it evident that OU usage patterns reliably differentiate DNA compositions. A total of 3518 MGE were retrieved upon a search of horizontally transferred genes throughout 637 complete bacterial genomes with SWGIS. MGE that were acquired from the search were further classified into homologous groups using BLAST based on similar DNA composition features. Groupings and classifications were performed as to illustrate genes and compositional features that are shared among the MGE that are hosted by organisms of various phylogenetic backgrounds. Also, these were performed in order to determine the types of genes that are likely transferred during horizontal transfer events. Upon the latter, MGE were converted into coding sequences and

compared against one another based on their amino acids compositions. MGE that shared similar features were clustered into families, afterward the latter clustering the proteins were once against compared against one another based on common functional parameters that they entail. Proteins that were found to share functionality were once more grouped into classes to study the evolutionary relationships of functional properties that are harboured by MGE. Entities that were observed from all the latter analysis were compared. Comparisons of all the major groups, families, and classes that were obtained from the analysis showed to have common enzymes and properties. The most abundant genes were found to be the ones involved in the synthesis of lipopolysaccharides, O-antigen, ABC transporters, regulatory and restriction-modification system proteins. Among these genes were also transposases and integrases encoding genes, elements that allow successful transfer of gene segments across species borders. Although these elements seem to be spread across variable MGE, they do not necessarily entail acquisition from a common ancestor. It could necessarily mean that MGE comprise of a common and preferential evolutionary approved combination of genes that perform specific molecular functions. However, it was also observed that not all organisms can acquire and transfer their segments to other organisms. As illustrated in Figure 3.5 (BLAST network), MGE of organisms such as *Burkholderia, Actinobacillus, Hahella, Yersinia, Burkholderia* and *Agrobacteria* appear as acceptors rather than donors, suggesting that the transfer and exchange of some MGE may be restricted to a particular line of organisms.

More analysis was conducted on MGE in search of order conservations shared among their gene entities. Measures of this analysis were conducted using homologous genomic segments that were obtained from BLAST. *Brucella melitensis 16M* was used as a reference genome. MGE that showed to be the closest homologues of *Brucella melitensis 16M* upon BLAST similarity search were compared against one to search for a shared synteny. Shared synteny among profiles serves as a measure that is widely used to study evolutionary relationships in organisms. The analysis illus-

trated that conservation of genes is extensively maintained in MGE harboured by species that are related. But conservation of genes is not only restricted to close phylogenetic distances, as the analysis also revealed that the conservation can also be shared between MGE of unrelated organisms. *Aeromonas* showed to share a well conserved synteny, however, some of its genes appeared rearranged but, they still are clustered closer to one another. Forces that conserve the order of genes among organisms are not well understood. It was previously stated that HGT cannot always account for instances of order conservations (Tamames, 2001) in phylogenetic distant microorganisms, as it is believed that there exists other factors that maintain the order of genes. Thus, future work facilitated by the findings in this study would help unravel some of the parameters that are hidden within MGE properties.

Summary

Horizontal gene transfer, well characterized as the transfer of genomic material between organisms contributes hugely in the evolution and speciation of bacteria. The transfer of such material brings about bacteria that are virulent and also in possession of genes that render them resistant to antibiotics. This helps to spread about and recombine genes of their kind to other bacteria. Horizontally acquired genomic elements exhibit compositional features that are deviant from the rest of the other genes in a recipient genome. They possess features such as unusual GC%, atypical codon usage, oligonucleotide usage bias and direct repeats at their flanks that can be used to distinguish them from native genes in a genome. This work focused on the developments of statistical and computational methods to aid with the detection of genes that have undergone horizontal transfer, to help track down genes that could be of medical and environmental importance. Therefore, SeqWord Gene Island Sniffer (SWGIS), a statistically driven computational tool for the prediction of genomic islands, and GEI-DB, a comprehensive database of horizontally transferred genomic elements were established. The SWGIS tool allows the precise predictions of precise inserts of horizontally acquired gene clusters in prokaryotic genomic sequences. Thus, the GEI-DB stores all the foreign genomic inserts that have been detected in the study, together with their annotations and evolutionary measures, such as groups of genomic islands that share similarities in DNA and amino acids features.

Bibliography

Allison, G. E., Angeles, D., Tran-Dinh, N. and Verma, N. K. (2002) Complete genomic sequence of SfV, a serotype-converting temperate bacteriophage of *Shigella flexneri*. *J Bacteriol* **184**, 7, 1974–1987.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**, 3, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 17, 3389–3402.

Ansong, C., Yoon, H., Norbeck, A. D., Gustin, J. K., McDermott, J. E., Mottaz, H. M., Rue, J., Adkins, J. N., Heffron, F. and Smith, R. D. (2008) Proteomics analysis of the causative agent of typhoid fever. *J Proteome Res* **7**, 2, 546–557.

Auchtung, J. M., Lee, C. A., Monson, R. E., Lehman, A. P. and Grossman, A. D. (2005) Regulation of a *Bacillus subtilis* mobile genetic element by intercellular signaling and the global DNA damage response. *Proc Natl Acad Sci U S A* **102**, 35, 12554–12559.

Baldi, P. and Baisnee, P. F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* **16**, 10, 865–889.

Beiko, R. G., Harlow, T. J. and Ragan, M. A. (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A* **102**, 40, 14332–14337.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2003) GenBank. *Nucleic Acids Res* **31**, 1, 23–27.

Betley, M. J. and Mekalanos, J. J. (1985) Staphylococcal enterotoxin A is encoded by phage. *Science* **229**, 4709, 185–187.

Blanco, M., Gutierrez-Martin, C. B., Rodriguez-Ferri, E. F., Roberts, M. C. and Navas, J. (2006) Distribution of tetracycline resistance genes in *Actinobacillus pleuropneumoniae* isolates from Spain. *Antimicrob Agents Chemother* **50**, 2, 702–708.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. and Shao, Y. (1997) The complete genome sequence of *Escherichia coli K-12. Science* **277**, 5331, 1453–1462.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 1, 365–370.

Bohlin, J., Skjerve, E. and Ussery, D. W. (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol* **4**, 4.

Broudy, T. B., Pancholi, V. and Fischetti, V. A. (2001) Induction of lysogenic bacteriophage and phage-associated toxin from group a streptococci during coculture with human pharyngeal cells. *Infect Immun* **69**, 3, 1440–1443.

Brussow, H. and Hendrix, R. W. (2002) Phage genomics: small is beautiful. *Cell* **108**, 1, 13–16.

Butler, J. E., He, Q., Nevin, K. P., He, Z., Zhou, J. and Lovley, D. R. (2007) Genomic and microarray analysis of aromatics degradation in *Geobacter metallireducens* and comparison to a *Geobacter* isolate from a contaminated field site. *BMC Genomics* **8**, 180.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L. and Brussow, H. (2003) Phage as agents of lateral gene transfer. *Curr Opin Microbiol* **6**, 4, 417–424.

Carbone, A., Zinovyev, A. and Kepes, F. (2003) Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* **19**, 16, 2005–2015.

Casjens, S. (2003) Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol* **49**, 2, 277–300.

Charkowski, A. O. (2004) Making sense of an alphabet soup: the use of a new bioinformatics tool for identification of novel gene islands. Focus on "identification of genomic islands in the genome of *Bacillus cereus* by comparative analysis with *Bacillus anthracis*". *Physiol Genomics* **16**, 2, 180–181.

Chen, L.-L. and Zhang, C.-T. (2003) Gene recognition from questionable ORFs in bacterial and archaeal genomes. *J Biomol Struct Dyn* **21**, 1, 99–109.

Chiu, C.-H., Chuang, C.-H., Chiu, S., Su, L.-H. and Lin, T.-Y. (2006) *Salmonella enterica serotype Choleraesuis* infections in pediatric patients. *Pediatrics* **117**, 6, e1193–e1196.

Chiu, C.-H., Tang, P., Chu, C., Hu, S., Bao, Q., Yu, J., Chou, Y.-Y., Wang, H.-S. and Lee, Y.-S. (2005) The genome sequence of *Salmonella enterica serovar Choleraesuis*, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res* **33**, 5, 1690–1698.

Choi, I.-G. and Kim, S.-H. (2007) Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A* **104**, 11, 4489–4494.

Conte, L. L., Brenner, S. E., Hubbard, T. J. P., Chothia, C. and Murzin, A. G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* **30**, 1, 264–267.

Coombes, B. K., Brown, N. F., Kujat-Choy, S., Vallance, B. A. and Finlay, B. B. (2003) SseA is required for translocation of *Salmonella* pathogenicity island-2 effectors into host cells. *Microbes Infect* **5**, 7, 561–570.

Coombes, B. K., Lowden, M. J., Bishop, J. L., Wickham, M. E., Brown, N. F., Duong, N., Osborne, S., Gal-Mor, O. and Finlay, B. B. (2007) SseL is a salmonella-specific translocated effector integrated into the SsrB-controlled salmonella pathogenicity island 2 type III secretion system. *Infect Immun* **75**, 2, 574–580.

Dai, S. and Zhou, D. (2004) Secretion and function of *Salmonella* SPI-2 effector SseF require its chaperone, SscB. *J Bacteriol* **186**, 15, 5078–5086.

Daubin, V., Lerat, E. and Perriere, G. (2003) The source of laterally transferred genes in bacterial genomes. *Genome Biol* **4**, 9, R57.

Davids, W. and Zhang, Z. (2008) The impact of horizontal gene transfer in shaping operons and protein interaction networks–direct evidence of preferential attachment. *BMC Evol Biol* **8**, 23.

DelVecchio, V. G., Kapatral, V., Elzer, P., Patra, G. and Mujer, C. V. (2002) The genome of *Brucella melitensis*. *Vet Microbiol* **90**, 1-4, 587–592.

Deschavanne, P., Giron, A., Fagot, J. V. G. and Fertil, B. (2000) Genomic Signature is Preserved in Short DNA Fragments *IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE'00)* , 6, 161.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. and Fertil, B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* **16**, 10, 1391–1399.

Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K. and Nattkemper, T. W. (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics* **10**, 56.

Dobrindt, U., Blum-Oehler, G., Nagy, G., Schneider, G., Johann, A., Gottschalk, G. and Hacker, J. (2002) Genetic structure and distribution of four pathogenicity islands (PAI I(536) to PAI IV(536)) of uropathogenic *Escherichia coli* strain 536. *Infect Immun* **70**, 11, 6365–6372.

Dobrindt, U., Hochhut, B., Hentschel, U. and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms *Nat Rev Microbiol* **2**, 5, 414–424.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A. and Deschavanne, P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res* **33**, 1, e6.

Dutta, C. and Pan, A. (2002) Horizontal gene transfer and bacterial diversity.

Enright, A. J., Dongen, S. V. and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 7, 1575–1584.

Ermolaeva, M. D. (2001) Synonymous codon usage in bacteria. *Curr Issues Mol Biol* **3**, 4, 91–97.

Feng, X., Oropeza, R. and Kenney, L. J. (2003) Dual regulation by phospho-OmpR of ssrA/B gene expression in *Salmonella* pathogenicity island 2. *Mol Microbiol* **48**, 4, 1131–1143.

Foerstner, K. U., von Mering, C., Hooper, S. D. and Bork, P. (2005) Environments shape the nucleotide composition of genomes. *EMBO Rep* **6**, 12, 1208–1213.

Fong, C., Rohmer, L., Radey, M., Wasnick, M. and Brittnacher, M. J. (2008) PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. *BMC Bioinformatics* **9**, 170.

Frech, G. and Schwarz, S. (1998) Tetracycline resistance in *Salmonella enterica subsp. enterica serovar Dublin*. *Antimicrob Agents Chemother* **42**, 5, 1288–1289.

Freeman, J. A., Rappl, C., Kuhle, V., Hensel, M. and Miller, S. I. (2002) SpiC is required for translocation of *Salmonella* pathogenicity island 2 effectors and secretion of translocon proteins SseB and SseC. *J Bacteriol* **184**, 18, 4971–4980.

Gaidelyte, A., Vaara, M. and Bamford, D. H. (2007) Bacteria, phages and septicemia. *PLoS One* **2**, 11, e1145.

Ganesan, H., Rakitianskaia, A. S., Davenport, C. F., Tummler, B. and Reva, O. N. (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* **9**, 333.

Garcia-Vallve, S., Guzman, E., Montero, M. A. and Romeu, A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res* **31**, 1, 187–189.

Godfroid, F., Cloeckaert, A., Taminiau, B., Danese, I., Tibor, A., de Bolle, X., Mertens, P. and Letesson, J. J. (2000) Genetic organisation of the lipopolysaccharide O-antigen biosynthesis region of *Brucella melitensis 16M* (wbk). *Res Microbiol* **151**, 8, 655–668.

Godfroid, F., Taminiau, B., Danese, I., Denoel, P., Tibor, A., Weynants, V., Cloeckaert, A., Godfroid, J. and Letesson, J. J. (1998) Identification of the perosamine synthetase gene of *Brucella melitensis 16M* and involvement of lipopolysaccharide O side chain in *Brucella* survival in mice and in macrophages. *Infect Immun* **66**, 11, 5485–5493.

Goshorn, S. C. and Schlievert, P. M. (1989) Bacteriophage association of streptococcal pyrogenic exotoxin type C. *J Bacteriol* **171**, 6, 3068–3073.

Groisman, E. A. and Aspedon, A. (1997) The genetic basis of microbial resistance to antimicrobial peptides. *Methods Mol Biol* **78**, 205–215.

Groisman, E. A., Sturmoski, M. A., Solomon, F. R., Lin, R. and Ochman, H. (1993) Molecular, functional, and evolutionary analysis of sequences specific to *Salmonella*. *Proc Natl Acad Sci U S A* **90**, 3, 1033–1037.

Gunn, J. S., Alpuche-Aranda, C. M., Loomis, W. P., Belden, W. J. and Miller, S. I. (1995) Characterization of the *Salmonella typhimurium* pagC/pagD chromosomal region. *J Bacteriol* **177**, 17, 5040–5047.

Hacker, J., Blum-Oehler, G., Hochhut, B. and Dobrindt, U. (2003*a*) The molecular basis of infectious diseases: pathogenicity islands and other mobile genetic elements. A review. *Acta Microbiol Immunol Hung* **50**, 4, 321–330.

Hacker, J. and Carniel, E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep* **2**, 5, 376–381.

Hacker, J., Hentschel, U. and Dobrindt, U. (2003*b*) Prokaryotic chromosomes and disease. *Science* **301**, 5634, 790–793.

Hansen-Wester, I. and Hensel, M. (2001) *Salmonella* pathogenicity islands encoding type III secretion systems. *Microbes Infect* **3**, 7, 549–559.

Hansen-Wester, I., Stecher, B. and Hensel, M. (2002) Analyses of the evolutionary distribution of *Salmonella* translocated effectors. *Infect Immun* **70**, 3, 1619–1622.

Hartman, A. B., Essiet, I. I., Isenbarger, D. W. and Lindler, L. E. (2003) Epidemiology of tetracycline resistance determinants in *Shigella* spp. and enteroinvasive

Escherichia coli: characterization and dissemination of tet(A)-1. *J Clin Microbiol* **41**, 3, 1023–1032.

Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K. F., Ueda, N., Hamajima, M., Kawasaki, T. and Kanehisa, M. (2006) KEGG as a glycome informatics resource. *Glycobiology* **16**, 5, 63R–70R.

Hensel, M., Shea, J. E., Raupach, B., Monack, D., Falkow, S., Gleeson, C., Kubo, T. and Holden, D. W. (1997) Functional analysis of ssaJ and the ssaK/U operon, 13 genes encoding components of the type III secretion apparatus of *Salmonella* Pathogenicity Island 2. *Mol Microbiol* **24**, 1, 155–167.

Hensel, M., Shea, J. E., Waterman, S. R., Mundy, R., Nikolaus, T., Banks, G., Vazquez-Torres, A., Gleeson, C., Fang, F. C. and Holden, D. W. (1998) Genes encoding putative effector proteins of the type III secretion system of *Salmonella* pathogenicity island 2 are required for bacterial virulence and proliferation in macrophages. *Mol Microbiol* **30**, 1, 163–174.

Hsiao, W., Wan, I., Jones, S. J. and Brinkman, F. S. L. (2003*a*) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* **19**, 3, 418–420.

Hsiao, W., Wan, I., Jones, S. J. and Brinkman, F. S. L. (2003*b*) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* **19**, 3, 418–420.

Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**, 3, 377–386.

Huson, D. H., Richter, D. C., Mitra, S., Auch, A. F. and Schuster, S. C. (2009) Methods for comparative metagenomics. *BMC Bioinformatics* **10 Suppl 1**, S12.

J, F. (1993) PHYLIP (PHYLogeny Inference Package) version 3.6a2, Distributed by the author, Department of Genetics, University of Washington, Seattle, WA. , 9.

Jain, R., Rivera, M. C. and Lake, J. A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* **96**, 7, 3801–3806.

Jeffrey, H. J. (1990) Chaos game representation of gene structure. *Nucleic Acids Res* **18**, 8, 2163–2170.

Johnson, L. P., Tomai, M. A. and Schlievert, P. M. (1986) Bacteriophage involvement in group A streptococcal pyrogenic exotoxin A production. *J Bacteriol* **166**, 2, 623–627.

Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1**, 5, 598–610.

Karlin, S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* **9**, 7, 335–343.

Karlin, S. and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**, 7, 283–290.

Karlin, S., Ladunga, I. and Blaisdell, B. E. (1994) Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci U S A* **91**, 26, 12837–12841.

Karlin, S., Mrazek, J. and Campbell, A. M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**, 12, 3899–3913.

Kent, W. J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res* **12**, 4, 656–664.

Klockgether, J., Wurdemann, D., Reva, O., Wiehlmann, L. and Tummler, B. (2007) Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in *Pseudomonas aeruginosa*. *J Bacteriol* **189**, 6, 2443–2459.

Koski, L. B., Morton, R. A. and Golding, G. B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* **18**, 3, 404–412.

Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., Edwards, R. A. and Stoye, J. (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* **36**, 7, 2230–2239.

Kuhle, V. and Hensel, M. (2002) SseF and SseG are translocated effectors of the type III secretion system of *Salmonella* pathogenicity island 2 that modulate aggregation of endosomal compartments. *Cell Microbiol* **4**, 12, 813–824.

Lawrence, J. G. (1999) Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol* **2**, 5, 519–523.

Lawrence, J. G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**, 4, 383–397.

Lawrence, J. G. and Ochman, H. (2002) Reconciling the many faces of lateral gene transfer. *Trends Microbiol* **10**, 1, 1–4.

Leplae, R., Hebrant, A., Wodak, S. J. and Toussaint, A. (2004) ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res* **32**, Database issue, D45–D49.

Li, J. and Sayood, K. (2005) A genome signature based on markov modeling. *Conf Proc IEEE Eng Med Biol Soc* **3**, 2832–2835.

Lima-Mendez, G., Helden, J. V., Toussaint, A. and Leplae, R. (2008*a*) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* **24**, 6, 863–865.

Lima-Mendez, G., Helden, J. V., Toussaint, A. and Leplae, R. (2008*b*) Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol* **25**, 4, 762–777.

Lowe, T. M. and Eddy, S. R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 5, 955–964.

Mantri, Y. and Williams, K. P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res* **32**, Database issue, D55–D58.

Marcus, S. L., Brumell, J. H., Pfeifer, C. G. and Finlay, B. B. (2000) *Salmonella* pathogenicity islands: big virulence in small packages. *Microbes Infect* **2**, 2, 145–156.

Massari, P., Ram, S., Macleod, H. and Wetzler, L. M. (2003) The role of porins in neisserial pathogenesis and immunity. *Trends Microbiol* **11**, 2, 87–93.

McHardy, A. C. and Rigoutsos, I. (2007) What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol* **10**, 5, 499–503.

Merlin, C., Springael, D. and Toussaint, A. (1999) Tn4371: A modular structure encoding a phage-like integrase, a *Pseudomonas*-like catabolic pathway, and RP4/Ti-like transfer functions. *Plasmid* **41**, 1, 40–54.

Mills, D. M., Bajaj, V. and Lee, C. A. (1995) A 40 kb chromosomal fragment encoding *Salmonella typhimurium* invasion genes is absent from the corresponding region of the *Escherichia coli K-12* chromosome. *Mol Microbiol* **15**, 4, 749–759.

Nikolaus, T., Deiwick, J., Rappl, C., Freeman, J. A., Schroder, W., Miller, S. I. and Hensel, M. (2001) SseBCD proteins are secreted by the type III secretion system of *Salmonella* pathogenicity island 2 and function as a translocon. *J Bacteriol* **183**, 20, 6036–6045.

Noble, P. A., Citek, R. W. and Ogunseitan, O. A. (1998) Tetranucleotide frequencies in microbial genomes. *Electrophoresis* **19**, 4, 528–535.

O'Brien, A. D., Marques, L. R., Kerry, C. F., Newland, J. W. and Holmes, R. K. (1989) Shiga-like toxin converting phage of enterohemorrhagic *Escherichia coli* strain 933. *Microb Pathog* **6**, 5, 381–390.

Ochman, H. (2001) Lateral and oblique gene transfer. *Curr Opin Genet Dev* **11**, 6, 616–619.

Ochman, H. and Groisman, E. A. (1996) Distribution of pathogenicity islands in *Salmonella* spp. *Infect Immun* **64**, 12, 5410–5412.

Ochman, H., Lawrence, J. G. and Groisman, E. A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 6784, 299–304.

Ogawa, N. and Miyashita, K. (1995) Recombination of a 3-chlorobenzoate catabolic plasmid from *Alcaligenes eutrophus NH9* mediated by direct repeat elements. *Appl Environ Microbiol* **61**, 11, 3788–3795.

Pezzella, C., Ricci, A., DiGiannatale, E., Luzzi, I. and Carattoli, A. (2004) Tetracycline and streptomycin resistance genes, transposons, and plasmids in *Salmonella enterica* isolates from animals in Italy. *Antimicrob Agents Chemother* **48**, 3, 903–908.

Philippe, H. and Douady, C. J. (2003) Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol* **6**, 5, 498–505.

Pickett, C. L., Pesci, E. C., Cottle, D. L., Russell, G., Erdem, A. N. and Zeytin, H. (1996) Prevalence of cytolethal distending toxin production in *Campylobacter jejuni* and relatedness of *Campylobacter* sp. cdtB gene. *Infect Immun* **64**, 6, 2070–2078.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. and Blaser, M. J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**, 2, 145–158.

Rajan, I., Aravamuthan, S. and Mande, S. S. (2007) Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics* **23**, 20, 2672–2677.

Reva, O. N. and Tummler, B. (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* **5**, 90.

Reva, O. N. and Tummler, B. (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* **6**, 251.

Rivera, M. C., Jain, R., Moore, J. E. and Lake, J. A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* **95**, 11, 6239–6244.

Sandberg, R., Winberg, G., Branden, C. I., Kaske, A., Ernberg, I. and Coster, J. (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res* **11**, 8, 1404–1409.

Sharp, P. M. and Li, W. H. (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 3, 1281–1295.

Spanier, J. G. and Cleary, P. P. (1980) Bacteriophage control of antiphagocytic determinants in group A streptococci. *J Exp Med* **152**, 5, 1393–1406.

Srividhya, K. V., Alaguraj, V., Poornima, G., Kumar, D., Singh, G. P., Raghavenderan, L., Katta, A. V. S. K. M., Mehta, P. and Krishnaswamy, S. (2007) Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. *PLoS ONE* **2**, 11, e1193.

Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A* **48**, 582–592.

Sullivan, J. T., Trzebiatowski, J. R., Cruickshank, R. W., Gouzy, J., Brown, S. D., Elliot, R. M., Fleetwood, D. J., McCallum, N. G., Rossbach, U., Stuart, G. S., Weaver, J. E., Webby, R. J., Bruijn, F. J. D. and Ronson, C. W. (2002) Compar-

ative sequence analysis of the symbiosis island of Mesorhizobium loti strain R7A. *J Bacteriol* **184**, 11, 3086–3095.

Tamames, J. (2001) Evolution of gene order conservation in prokaryotes. *Genome Biol* **2**, 6.

Tamames, J., Casari, G., Ouzounis, C. and Valencia, A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol* **44**, 1, 66–73.

Tamames, J. and Moya, A. (2008) Estimating the extent of horizontal gene transfer in metagenomic sequences. *BMC Genomics* **9**, 136.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. and Glockner, F. O. (2004*a*) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**, 9, 938–947.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. and Glockner, F. O. (2004*b*) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163.

Thomas, C. M. and Nielsen, K. M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 9, 711–721.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997) The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 24, 4876–4882.

Top, E. M. and Springael, D. (2003) The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Curr Opin Biotechnol* **14**, 3, 262–269.

Tzeng, Y.-L., Noble, C. and Stephens, D. S. (2003) Genetic basis for biosynthesis of the (alpha 1–>4)-linked N-acetyl-D-glucosamine 1-phosphate capsule of *Neisseria meningitidis* serogroup X. *Infect Immun* **71**, 12, 6712–6720.

Uchiumi, T., Ohwada, T., Itakura, M., Mitsui, H., Nukui, N., Dawadi, P., Kaneko, T., Tabata, S., Yokoyama, T., Tejima, K., Saeki, K., Omori, H., Hayashi, M., Maekawa, T., Sriprang, R., Murooka, Y., Tajima, S., Simomura, K., Nomura, M., Suzuki, A., Shimoda, Y., Sioya, K., Abe, M. and Minamisawa, K. (2004) Expression islands clustered on the symbiosis island of the *Mesorhizobium loti* genome. *J Bacteriol* **186**, 8, 2439–2448.

Uchiya, K., Barbieri, M. A., Funato, K., Shah, A. H., Stahl, P. D. and Groisman, E. A. (1999) A *Salmonella* virulence protein that inhibits cellular trafficking. *EMBO J* **18**, 14, 3924–3933.

Uchiya, K. and Nikai, T. (2008) *Salmonella* virulence factor SpiC is involved in expression of flagellin protein and mediates activation of the signal transduction pathways in macrophages. *Microbiology* **154**, Pt 11, 3491–3502.

van Passel, M. W. J., Bart, A., Thygesen, H. H., Luyf, A. C. M., van Kampen, A. H. C. and van der Ende, A. (2005) An acquisition account of genomic islands based on genome signature comparisons. *BMC Genomics* **6**, 163.

Wagner, P. L. and Waldor, M. K. (2002) Bacteriophage control of bacterial virulence. *Infect Immun* **70**, 8, 3985–3993.

Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L. and Rapp, B. A. (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* **30**, 1, 13–16.

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C. S., Sutton, G., Frazier, M. and Venter, J. C. (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One* **3**, 1, e1456.

Willner, D., Thurber, R. V. and Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol*.

Yoon, S. H., Hur, C.-G., Kang, H.-Y., Kim, Y. H., Oh, T. K. and Kim, J. F. (2005) A computational approach for identifying pathogenicity islands in prokaryotic genomes. *BMC Bioinformatics* **6**, 184.

Zhu, P., Klutch, M. J., Derrick, J. P., Prince, S. M., Tsang, R. S. W. and Tsai, C.-M. (2003) Identification of opcA gene in *Neisseria polysaccharea*: interspecies diversity of Opc protein family. *Gene* **307**, 31–40.

Zurawski, D. V. and Stein, M. A. (2003) SseA acts as the chaperone for the SseB component of the *Salmonella* Pathogenicity Island 2 translocon. *Mol Microbiol* **47**, 5, 1341–1351.

Zygmunt, M. S., Hagius, S. D., Walker, J. V. and Elzer, P. H. (2006) Identification of *Brucella melitensis 16M* genes required for bacterial survival in the caprine host. *Microbes Infect* **8**, 14-15, 2849–2854.