CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Integrated Approach for Computational Prediction of Bacterial Pathogenicity Islands

by

Hajra Qayyum

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Health and Life Sciences
Department of Bioinformatics and Biosciences

2019

In the memory of my beloved father Malik Abdul Qayyum (Late). I wish You were here to read this.

CERTIFICATE OF APPROVAL

Integrated Approach for Computational Prediction of Bacterial Pathogenicity Islands

by

Hajra Qayyum

(MBI173001)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Rehan Zafar Paracha	NUST, Islamabad
(b)	Internal Examiner	Dr. Shaukat Iqbal	CUST, Islamabad
(c)	Supervisor	Dr. Syeda Marriam Bakhtiar	CUST, Islamabad

Dr. Syeda Marriam Bakhtiar
Thesis Supervisor
October, 2019

Dr. Sahar Fazal
Head
Dept. of Bioinformatics and Biosciences
October, 2019

Dr. Muhammad Abdul Qadir
Dean
Faculty of Health and Life Sciences
October, 2019

Author's Declaration

I, **Hajra Qayyum** hereby state that my MS thesis titled "**Integrated Approach for Computational Prediction of Bacterial Pathogenicity Islands**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Hajra Qayyum)
Registration No: MBI173001

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Integrated Approach for Computational Prediction of Bacterial Pathogenicity Islands**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been dully acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Hajra Qayyum)
Registration No: MBI173001

Acknowledgements

After expressing my gratitude to the supreme power of Allah (S.W.T), the creator of all creations, for enabling me to complete my thesis in time, I would like to express my sincerest thanks to my supervisor, **Dr. Syeda Marriam Bakhtiar**, for her continuous support, guidance and great insight, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my MS study.

Besides my supervisor, I would like to thank Dr. Amjad Ali, Associate Professor, ASAB, National University of Science and Technology, Islamabad, for his positive suggestions and guidance throughout my thesis. I would also like to extend my gratitude to Mr. Siomar C. Soares, Assistant Professor, Institute of Biological and Natural Sciences, Federal University of Triângulo Mineiro, Brazil, the pioneer in the field of genomic islands for sharing his knowledge regarding the subject.

I further thank Dr. Sahar Fazal, Dr. Shaukat Iqbal Malik, Ms. Fatima Khan and all other faculty members of the Department of Health and Life Sciences, Capital University of Science and Technology, Islamabad, for their honest opinions on academic life and science. I would like to acknowledge my colleagues Mr. Saboor Ahmed, Ms. Sadia Zaman, and Ms. Shabana Yasmeen for facilitating me in every possible way throughout. Last but not the least, I would like to thank my father for having faith in me, for making me strong enough to struggle for myself, my mother for always supporting me and standing out for me and my friends for motivating me, especially Anum Munir and Sana Masood for always being standing with me and for being the constant source of encouragement and motivation throughout this time period.

(Hajra Qayyum)

Registration No: MBI173001

Abstract

Lateral Gene Transfer refers to the transfer of genes from one organism to another other in lateral fashion. The regions that are exchanged between the genomes as a result of such events are known as Mobile genetic elements. These elements include transposons, integron, prophage, genomic islands, and insertion sequence elements. Genomic islands (GIs) are large regions of the chromosomes that constitute a flexible gene pool. These are the regions of the DNA that are transferred from one organism to another by other mobile genetic elements. Such genomic regions encode genes that confers adaptability and versatility advantages to a bacterium. If such regions confer virulent properties to the bacterium, these are called

pathogenicity islands (PAIs). Characterizing GIs gives insights into the nature of bacterial species as to why some of the strains could tolerate extreme living conditions while others do not, why some of the strains are resistant towards a particular antibiotic while, others do not. Focus of this study is on PAIs which facilitate to give insight into the pathogenic nature of the bacterial species concerning why even within the same species, some of the bacterial strains are pathogenic in nature while, others are not. Therefore, identification of PAIs constitutes one of the critical tasks for understanding the nature of pathogenic species benefitting biomedical research. Identification of PAIs can lead towards better diagnosis and antibiotics designing, ultimately, contributing to human health. As the biological experiments only contributes a meager fraction of information for PAIs in the sequenced genome, computational approaches seem to be the better option. Currently, two computational approaches are available for the identification of GIs i.e. sequence-based and comparative genomics-based pipelines. Each of the two available approaches have their own limitations resulting in false predictions limiting the accuracy of these approaches. Therefore, in this study an integrated approach has been proposed which could overcome the short-comings of each approach while identifying PAIs in particular, with improved accuracy. Proposed integrated approach is based on the existing pipelines with certain modifications that suggest broadening up the

subset of genomic signatures leading to the application of more stringent criteria in decision making. This study has focused on organisms which comparatively weaker known basis of pathogenicity by selecting an opportunistic pathogen *Streptococcus sanguinis SK36* as a case, suggesting the application of k-means clustering approach for determining the non-pathogenic strains leading to better selection of subject genome and comparative genome set. Suggested integrated approach has led to more accurate identification of GIs/PAIs and has shown better prediction results as compared to the other conventional pipelines.

Contents

List of Figures

List of Tables

Abbreviations

AIs	Antimicrobial Resistance Islands
cPAI	candidate PAthogenecity Island
DR	Direct Repeats
FN	False Negative
FP	False Postive
GI	Genomic Islands
GI-GPS	Genomic Islands Genomic profile Scanning
GIST	Genomic Island Suit of Tools
IS	Insertion Sequence
IE	Insertion Elements
IVOM	Interpolated Variable Order Motifs
ICE	Integrative and Conjugative Elements
ICEs	Integrative and Conjugating Elements
LGT	Lateral Gene Transfer
MGE	Mobile Genetic Elements
NCBI	National Center for Bioinformatics
ORF	Open Reading Frames
PAIs	PAthogenecity Islands
SVM	Support Vector Machine
TP	True Positive
TN	True Negative
TCP	Toxin Co-regulated Pilus
T3SS	Type III Secretion System
T4SS	Type IV Secretion System
VFDB	Virulence Factor Database

Chapter 1

Introduction

Microbes are the most diverse organisms that accounts for 60% of Earth biomass [1]. Among microbes, bacteria are the most widely spread organisms that are found nearly everywhere. This ubiquity is due to the adaptive nature of the bacteria towards various environments. Besides the bacterial conventional mode of gene transfer, bacterial cells also have the ability of transferring genes laterally. This lateral transfer of genes makes a strain adaptive in comparison to others within a same species and induces strains specific properties.

1.1 Lateral Gene Transfer

Adaptive bacterial nature is due to the ability of bacteria to acquire genes of horizontal origin from various sources including prokaryotes, viruses and even eukaryotes via a process called Lateral Gene Transfer (LGT). LGT refers to genes transfer from an organism to another in a lateral fashion. LGT facilitates evolution and has been accepted as one of the important evolutionary mechanism of life [2]. LGT takes place by following processes: conjugation, transformation and transduction (Figure 1.1).

Transformation involves direct uptake, incorporation and expression of an exogenous DNA by a bacterium from its external environment [3]. A cell that is capable of up-taking a DNA is known as competent. Such state of competency is

usually inducible by external stimuli such as pulse heat shock etc. The process of transformation initiates with the attachment of DNA (double stranded) with the specific binding sites present on cell surface. Followed by conversion of foreign DNA into single stranded form by the action of series of proteins such type IV pili and type II secretion system proteins, it is then translocated into the cell [4].

In contrast to that, conjugation is the transfer of exogenous DNA by establishing a physical link (mating pillus) between recipient and a donor. Transferred genetic material is usually a plasmid or a transposon. Elements encoding conjugation machinery are considered self-transmissible whereas, those that rely upon externally encoded conjugation systems are referred to as mobilizable. This activity takes place by the projection of sex pillus directing towards recipient cell from donor cell at considerable distance. The single stranded DNA is then transferred into the recipient cell by forming a replicating rolling circle by the action of secretion systems such as type IV [4].

Transduction refers to DNA transfer carried via a virus infecting a prokaryotic cell called phage. Phages are categorized broadly based on their infection follow up strategy i.e. whether they enter lytic phase or become dormant. The dormant or temperate phages integrates their own DNA within a bacterial genome becoming a prophage and keeps on replicating for many generations along with the host's genome. Induction (spontaneous or environmental change) helps a temperate phage to come out of dormancy and enter a complete lytic cycle. Transduction can be generalized as well as specialized. Generalized transduction includes a phage particle packaging host's DNA fragments randomly during its lytic phase of the cycle. While, specialized transduction takes place when host DNA having an integrated is replicated during phage induction and becomes integrated into the phage particle [4].

These LGT events facilitates the transfer of genetic material causing genomic alteration via gene loss and gain and are major source of evolution. Such acquisition of genes plays a crucial role in the adaptive evolution of prokaryotes conferring beneficial traits to bacteria under particular growth and environmental conditions [2].

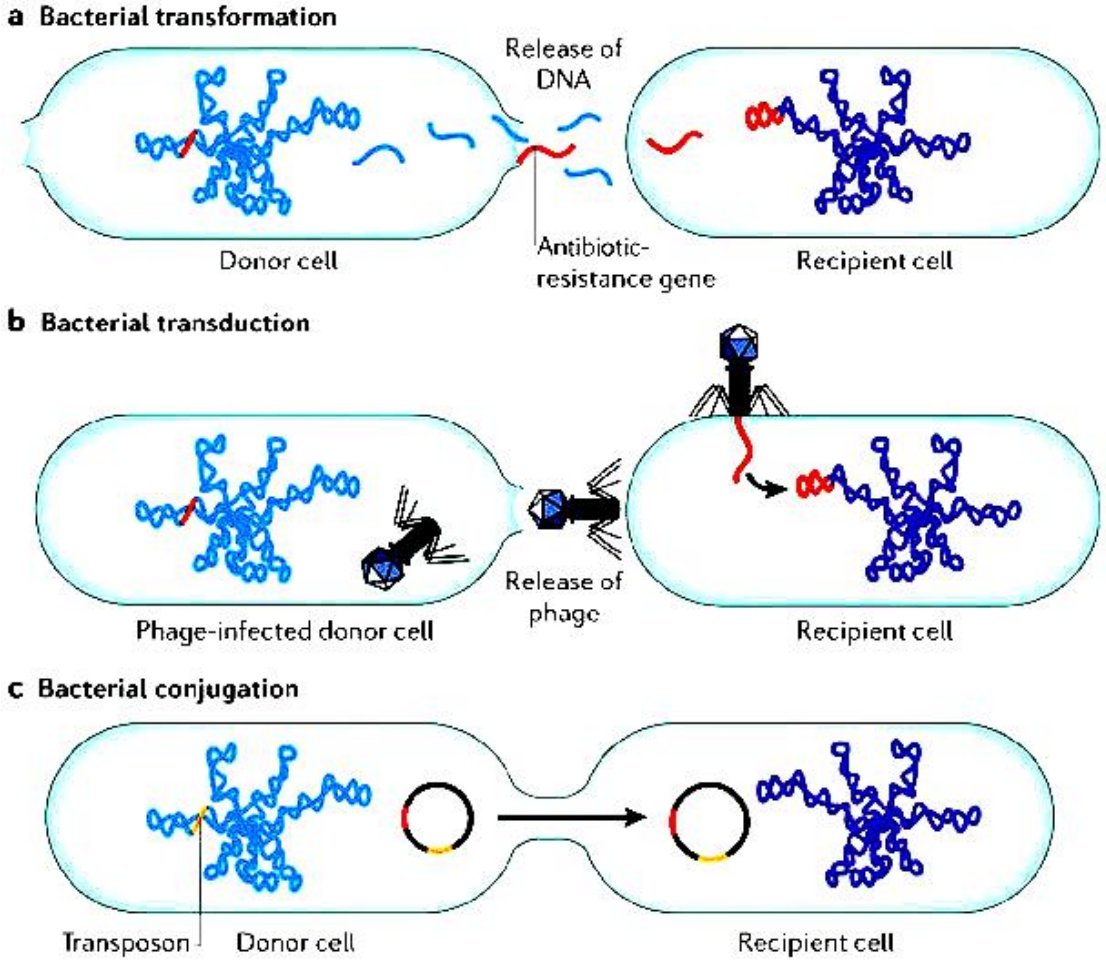

FIGURE 1.1: Bacterial modes of horizontal gene transfer

1.2 Mobile Genetic Elements

Regions exchanged between the genomes as a result of LGT events are known as Mobile Genetic Elements (MGEs) and are described as the blocks of DNA that can translocate on a chromosome and can be exchanged between different chromosomes and even different species [5]. MGE includes transposons, integron, prophage, genomic islands (GIs), and insertion sequence (IS) elements [4].

Genomes of both the prokaryotes and eukaryotes contains a large amount of repeated DNA sequences which are mobile in nature and are referred as transposons (TNs). TNs greatly affect the genomes in positive or negative manner due to their mobility. They have a crucial part in genome evolution by

promoting inactivation of genes, managing the gene expression levels through recombination which results in the change in genome size during evolution [6]. The size of transposons is about 5kb, containing approximately 38 to 40 bps with inverted repeats at both ends from which 5 bp of repeated DNA sequences is generated during insertion [7].

Integrons on other hand, are the hereditary elements that permit the start and stop of gene expression They are generally known due to their important role in the development of drug resistance, especially among Gram-negative bacteria. Clinically, integron are a most common portion of bacterial genomes occurring in all environmental conditions, can move among species and easily pass to lineages through evolution, and they have access to a huge pool of novel genes whose functions are not known [8]. All integron consists of three primary components to capture exogenous genes: a gene (intI) that encodes integrase (tyrosine-recombinase family), a primary recombination site (attI), and an outward-orientated promoter (Pc) for initiation of transcription of exogenously acquired genes [9].

Interplay between lysogenic and lytic cycle of phage during transduction is considered as the major source of variations in genomic sequence pattern of bacterial strains. Prophage's genome could account for 10-20 % of genes in bacterial genomes. Number of virulence factors contributing towards bacterial pathogenesis is mobilized by phages and is considered as principal factor in the evolution emerging pathogens. Prophage regions usually includes an integrase and certain phage related genes. To identify the prophage integration, flanking direct repeats or presence of tRNA is considered as a supportive evidence [4].

Insertion Elements (IEs) are yet another kind of MGEs that are the small segments of DNA usually less than 2.5-kb, with a simple organization of genetic sequence and ability to insert at more than one space in a genome at a time. These constitutes components with RNA as intermediates, for example, the retroviruses, retrotransposons, DNA components including conjugative transposons and components of bacteriophage Mu, Tn7, and transposons of the Tn554 type, many IEs have complex structure and show multiple drug resistance [10]. Number of IEs increases or decreases due to inactivation of genes and

decay, genome modification and reduction in genome, therefore, affects the life of host [11].

Genomic Islands are the blocks of genes that are being transferred laterally as a result of LGT events and are the focal point of this study. GIs usually range from 5 to 500 kb [12] [4] and encode various genes that confer adaptability advantages of medical and ecological importance [12]. Depending upon the type of genes, a GI could be defined as Antimicrobial Island or Pathogenicity Islands etc [13].

1.3 Genomic Islands (GI)

GIs are large regions of the chromosomes constituting a flexible gene pool. These are the regions of the DNA that are transferred from one organism to another by other MGEs. Such genomic regions are possessed by certain bacteria and are absent from its closely related strains. GIs encode genes that confers adaptability and versatility advantages to a bacterium and are associated with tRNA and flanked direct repeats structures. They are also characterized by mobility genes encoding integrases or transposases necessary to integrate and excise the chromosome. They have a significant part in the dynamic character of bacterial chromosomes and can be excised and transferred from one chromosome to other [14].

Research on GIs have remained an area of extensive interest due to its role in genomic variability and evolution of pathogenic bacteria. The concept of GI was originated for pathogenic bacteria and was associated with pathogenicity, but later on due to the observation of these region in related non-pathogenic bacteria and the different ecological context-based functions GIs play in genome leads to more relaxed term- the GIs. This generalization differentiates the term pathogenicity island (PAIs) from GIs while merely correlating it with the pathogenesis. On the other hand, depending upon the types of adaptability advantages they confer GIs get classified as pathogenicity, symbiosis, fitness, metabolic and resistance islands [14].

GIs play a crucial role in bacterial survival and fitness. Bacterial fitness is defined as the resultant of properties that intensify its survival and transmission in a particular niche. In context to that, GIs confer one evolutionary advantage in a way that large blocks of genes can be exchanged between two organisms conferring new traits to the recipient. Another advantage is the maintenance of genetic flexibility while shifting parts of host chromosomal DNA when it is excised from the host genome, thus, transferring parts of host chromosomes to recipient results in successful adaptation and enhanced fitness to a specific niche [14][15]. In addition to that, along with the facilitation in translocation of GIs and GIs- encoded products, it also aids in exchange of host's chromosomal DNA. Genetic variability is another role it plays, as it undergoes recombination with the host's chromosome that has a crucial impact on the host-bacterium evolution. Irrespective of genes these genomic regions have the potential to drastically alter the life styles of the bacterium due to its acquisition or rapid loss from a genome facilitating evolution by "quantum leaps". It should also be noted that along with many known GI-encoded genes, these regions also possess many of the novel and hypothetical proteins with unknown functions that have no traceable homologues in other species, however, sometimes confer selective advantage to the host [15].

Keeping in view the magnificent role of GIs, it is highly significant to detect and identify the locations and contents of GIs as such finding will greatly benefit the biomedical research and are of clinical importance. For example, design and production of GI- based identification of antibiotics. While on the other side, GIs containing beneficial metabolites can be subjected to large-scale production [13],[15].

With the rapid increase in the genomic sequence data, huge amount of information regarding genomic structures needs to be validated. In context to that, biological experiments contribute only meagre amount of information for GIs in all sequenced genomes, therefore, identification of such regions through computational means is a quick and best suitable solution available [13].

Due to the fact that GIs originate from different bacterial lineage, sequence composition of the GIs differs significantly from the host genome. This difference

is exploited by the computational approaches for the detection of GIs. Following are the characteristics of GIs which marks the detection of GIs (Figure 1.2).

Uneven distribution is the most prominent property of genomic islands. Such genomic sequences are of diverse origin and their phyletic pattern is different than rest of the host genome. GIs are present in few isolates of a respective species or strain. They are unstable and are reported to be excised sporadically even within a specific strain [16].

GIs also differ in G+C content in terms of percentage and oligonucleotide lengths (2-9 nucleotides). For example, dinucleotide bias is measured via counting on pairs of nucleotides and analyzing if there is any change in the number of each pair as compared to expected in certain region (probability of one-sixteenth per pair) or is compared to the average of a genome [16].

On the other hand, codon usage also varies for GIs. It is a measure of oligonucleotide of length three. Such calculations are also compared to the average sequence composition of the host genome, as an approach to predict GIs and to detect LGT events [16].

As far as the size of GIs has been proposed as a GI predictor, typical threshold of 8 genes or 8 kb is suggested by many methods as minimum size of a GI though this minimum threshold lacks biological evidences [16].

Studies show presence of some genomic elements associated with GIs. These elements include tRNA and flanking direct repeats. Thus, these elements could be utilized as markers while identifying GIs. Transfer RNA genes represent phage integration sites whereas direct repeats appear as a consequence when a phage is inserted into tRNA gene. Certain types of tRNA genes, such as transfer messenger RNA (tmRNA) gene and the genes encoding tRNASer, tRNAArg, tRNALeu, tRNAThr and tRNASec, are favorably exploited as insertion sites for the phages. However, flanking tRNA could not always be used as GI predictor as significant number of GIs does not possess them [16].

GIs also contain certain types of protein encoding region that could be used as markers for their identification. Most of these genes are associated with mobility of MGE and includes integrases and transpoases. Such genes are termed as “mobility

genes" and show whether a GI is self- mobilized or is leftover of other embedded MGEs which are frequently found in GIs such as insertion elements (IS). Specific classes of functional genes are also found over-expressed in GIs. These classes include genes that encode cell surface proteins, host interaction proteins, DNA-binding proteins, phage-related proteins and those associated to the mobility of MGE. The strongest indicator of GI counts disproportionate presence of proteins encoding genes, that have no homologues or have unknown functions [16].

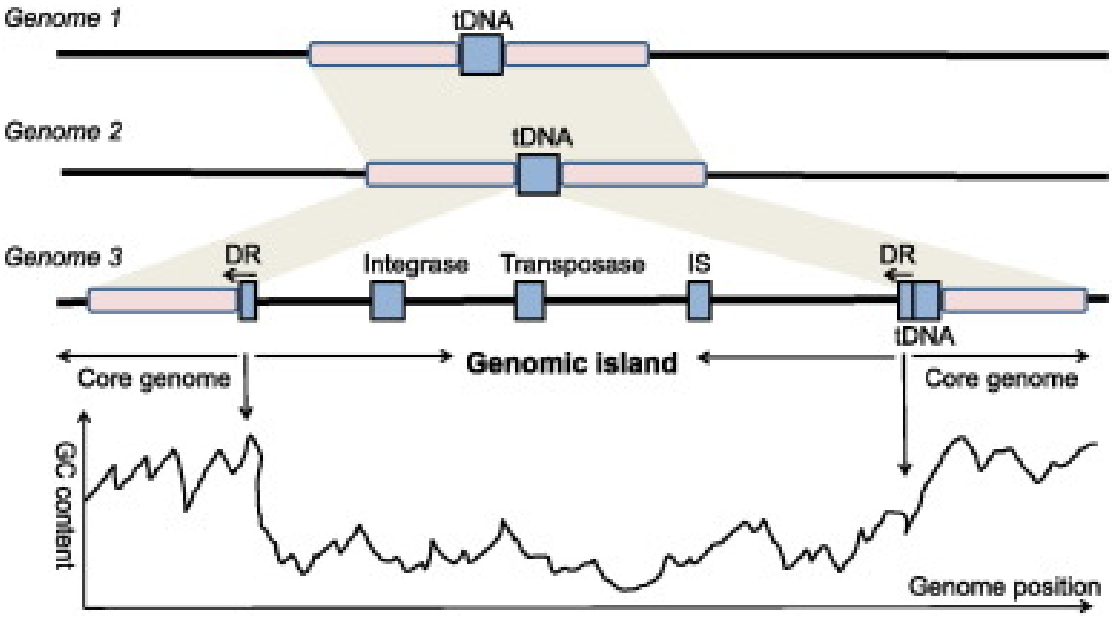

FIGURE 1.2: Structure of genomic island [15]

GI is flanked by direct repeats and possesses integrase, transposases and insertion sequences. Graph in Figure 1.2 is illustrating difference in G+C content between core genome and GI region.

All of the above- mentioned features characterize a region as a GI, though, it is not necessary that each GI contains all of these features, the simultaneous existence of a subset of these features could provide strong evidence for LGT event. However, some recent research suggests most important feature as, sequence composition bias, length of the region and the presence of integrase and phage-related genes, apart from phylogeny-based methods, whereas gene density, tRNA genes and the presence of flanking direct repeats could also be identified [16].

1.4 Pathogenicity Islands (PAIs)

Pathogenicity corresponds to the disease-causing expression of factors that exist in the pathogenic bacteria while being absent from non-pathogens. Usually a bacterial genome has a "core" or conserved region encoding basic cellular functioning information and a "flexible" gene pool encoding supplementary traits that are useful under specific conditions. This variable gene pool contains resistance, toxins and virulence encoding genes. The organization of core genome remains conserved among the closely related organisms whereas, variable genome represents variable chromosomal regions including certain MGEs like bacteriophages, plasmids, pathogenicity islands, insertion sequence elements, transposons and integron. In pathogenic bacteria, genome plasticity is a significant feature as it allows acquisition of multiple genes by LGT facilitating the single-step inheritance of complex disease-related characteristics [15].

Virulence factors in a pathogenic bacterium are usually located on MGEs including the pathogenicity islands [15]. Pathogenicity Islands (PAIs) refers to the distinct genetic elements present on the chromosomes of pathogenic bacteria [17]. PAIs are considered as the large continuous segments of genes that exist only in the pathogenic bacterial strains while being absent in other related strains of the same species [17]. Identification of such regions is of great medical interest as such regions carry multiple genes which contribute to the pathogenic virulence as well as potential vaccine candidates could also be located within PAIs [2].

PAIs may differ in structure and function but some of the features remain same. They include one or more virulent genes and covers large area on the chromosomes. These PAIs vary in size and may range from 10-100 kb. Sometimes, a bacterial genome harbors smaller pieces of DNA referred to as "pathogenicity islets". PAIs vary from the host genome in terms of G+C content as well as the codon usage. PAIs can be identified by flanking tRNA or direct repeats (DR) on one side. Furthermore, PAIs, frequently encodes factors responsible for genetic mobility i.e. integrases, transposases, phage genes and origins of replication [17]. The general structure of PAI is shown in Figure 1.3.

Notable characteristics of PAIs includes variation in the G+C content. G+C content refers to the percentage of the bases guanine and cytosine and usually varies between the host genome and PAIs [18]. These PAIs not only differ in their base composition but also vary in codon usage. Reason for such discrepancy is not yet known, however, the conservation of a genus- or species-specific base composition is a remarkable bacterial characteristic [19].

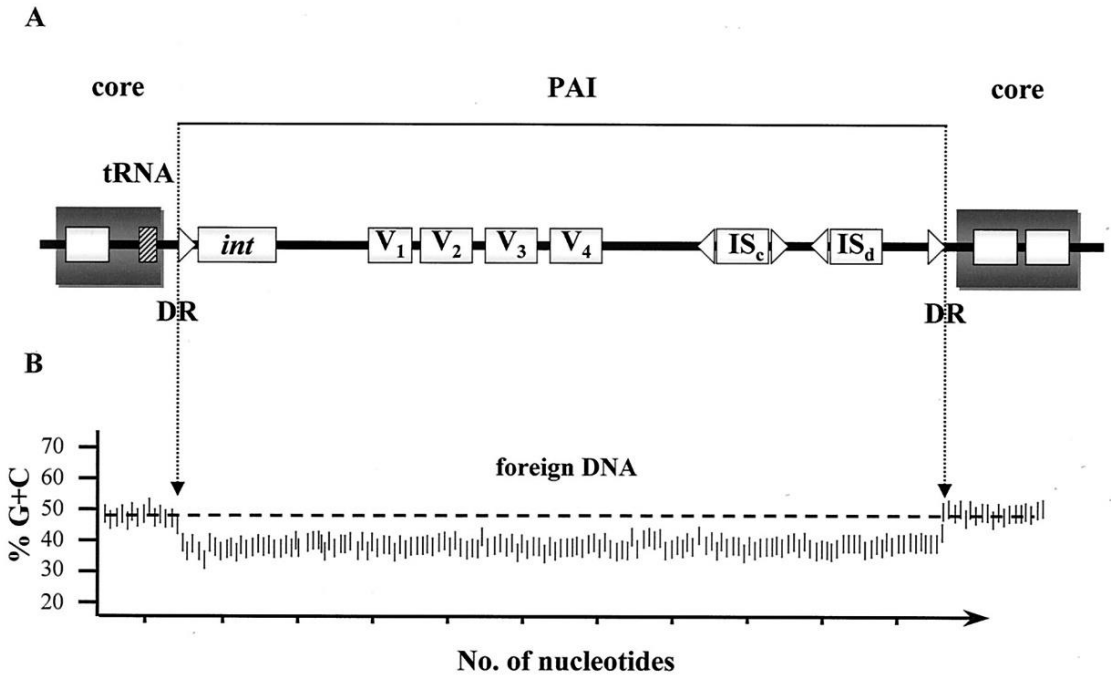

FIGURE 1.3: General structure of PAI flanked by direct repeats.

K-mer frequency is another such feature which is used to differentiate the two genomes. Such measurement of dinucleotide or high-order nucleotide frequencies has been extensively used for detecting PAIs. K-mer frequency is also used as a parameter for PAIs prediction by different tools such as AlienHunter and Centroid [18].

Presence of virulence genes is the most significant feature of PAIs. Such genes are functionally categorized as adherence factors, siderophores, exotoxins, invasion genes and type III and IV secretion systems. Adherence factors empower bacteria to get attached to the host surfaces and aid the process of infection such as P-related pilli, S-fimbriae, vibrio cholerae toxin coregulated pilus (TCP) and intimin. Siderophores like aerobactin are used to deliver essential iron into microbial cells. Exotoxins including alpha-haemolysin, and enterotoxins are the compounds that

can have impact on the function of eukaryotic cell. Invasion genes facilitates the bacterial cell entry into eukaryotic epithelial cells such as *inv* genes of *Salmonella* spp. Type III and IV secretion system are needle like structures that aids delivery of bacterial effector proteins (to modulates host functions) into the host cells. Type III (T3SS) and Type IV secretion systems (T4SS) are present in *Salmonella*, *Agrobacterium tumefaciens* respectively [17],[18]. Findings indicate presence of more aggressive virulence determinants in PAI regions of pathogens as compared to non-pathogenic bacterial species. PAI can be detected for virulence genes by using tools like VirulentPred or manually using BLAST search [18].

Transfer RNA genes depict the ideal site for the exogenous DNA integration, including PAIs. The tRNA genes are usually identical to the attachment sites of bacteriophages at their 3' end, hence serves as ideal integration site for certain plasmids and phages in various bacteria. Example of PAIs insertion into tRNA specific loci includes PAI I and PAI II of UPEC 536. PAI II is incorporated into the locus of the leucine tRNA gene (leuX), wheras, PAI I was found to be integrated into the tRNA gene of selenocysteine (selC). However, in case of some pathogens, insertion of PAIs can be found as much frequently as other tRNA sites [17].

PAIs often encode functional mobility or cryptic genes which includes phage-like integrase genes, referred as *int*, or genes for transposases. Some other PAIs possess genes homologous to the integrase and transposons resolvase genes of phages. Products of such genes are responsible for integrating and excising out the DNA regions by recombination between flanking DRs, IS elements, or within regions of homologous sequences. Hence, subsets of PAIs for some pathogenic bacteria may get deleted spontaneously, making genome unstable. Whereas, in some cases PAIs becomes integrated into the chromosomes permanently [17].

PAIs are characterized by flanking Direct Repeats (DR) regions that are described as sequences of DNA comprising of 16-20 bp with absolute or nearly absolute sequence repetition. These repeats are usually homologous to phages attachment sites and have been originated during the chromosomal integration of MGEs at hosts site via specific recombination yielding duplicated integration site. DRs serve as the recognition sites for the enzymes taking part in excision of MGEs, ultimately, enhancing the genomic instability of the island [17].

Insertion Sequences (IS) are the small mobile genetic elements that keep on transposing themselves, within and between prokaryotic genomes. IS offers sites for inverted repeats for homologous recombination and can mediate integration of MGEs into the chromosomes yielding PAIs. These IS elements can bring excision or instability to PAIs. For example, in *Yersinia pestis*, the *pgm* PAI is flanked by DRs of IS100. IS100 has approximately 30 copies in the genome of *Yersinia* and can mediate the integration of plasmids into the chromosomes [17].

As phage transduction and integration are the key processes of LGT events, abundance of phage-related genes is observed in PAIs. Thus, presence of phage-related genes is also used as a parameter to identify PAIs. Phage-related genes can be searched in protein databases such as Pfam using HMMER3 [18].

When compared to host genome, PAIs are characterized by the presence of relatively higher number of proteins with unknown functions. The reason for their unknown nature is the unavailability of cultured and sequenced donor genomes along with functional annotation. Hypothetical proteins can be detected by the acquisition of Open Reading Frames (ORFs) with the aid of gene-detecting tools such as GeneMark and Glimmer followed by the subtraction of proteins with known functions determined by NCBI-nr/nt, Pfam, UniprotKB, or COG database. Interestingly, it has also been found that inter-genic distance of island regions is longer as compared to the core regions [18].

Above mentioned properties of PAIs forms the basis of their computational identification. Most of the features, if not all, have been made basis for detecting PAIs in pathogenic species following computational approach providing efficient and accurate results in less time.

1.5 Purpose

As the number of sequenced genome is increasing, role of GIs in prokaryotic evolution is becoming more revealing, whereas, identifying such DNA blocks utilizing bioinformatics approach is becoming a fundamental aspect of the microbial evolution and functions [4],[18]. Characterizing GIs gives insights into

the nature of bacterial species as to why some of the strains could tolerate extreme living conditions while others do not, why some of the strains are resistant towards a particular antibiotic while, others do not whereas, focus of this study is on PAIs which facilitate to get insight into the pathogenic nature of the bacterial species concerning why even within the same species, some of the bacterial strains are pathogenic in nature while, others are not. Therefore, identification of PAIs constitutes one of the critical tasks for understanding the nature of pathogenic species benefitting biomedical research.Better identification of PAIs leads towards better diagnosis and antibiotics designing, ultimately, contributing to the human health [3],[4].

As the biological experiments only contributes a meager fraction of information for PAIs in the sequenced genome [13] computational approaches seem to be the better option. Currently, computational approaches for identifying GIs fall into two major categories i.e. sequence based and comparative genomics-based pipelines, whereas, surprisingly, no special pipeline has been designed for the identification of PAIs in particular which stimulated our interest in this area. Moreover, each of the two available approaches has its own limitations (discussed in detail in the literature review section) resulting in false positive results, therefore, the aim of this study is to design an integrated pipeline minimizing the limitations of each approach while identifying PAIs in particular, with more precision and accuracy.

1.6 Problem Statement

Pathogenicity islands are repeatedly found to contain virulence genes in bacteria that holds medical importance. Therefore, prediction of such regions is of significant interest. Computational identification of such regions is preferred over biological methods because of time constraints.Existing computational approaches focuses on the identification of GIs in general and not specifically on PAIs. Moreover, both the existing GI prediction approaches faces limitation that reduces the prediction accuracies of the existing approaches yielding false results. Prevailing approaches also fail to predict GIs/PAIs for a pathogen with closest neighbors of unknown pathogenicity.Therefore, there is a need of designing an

approach that can minimize the limitations faced by both of the existing approaches achieving improved prediction accuracy and is applicable to every pathogen irrespective of the unknown pathogenic nature of its closely related strains minimizing high rates of false results which misleads the identification of these critical areas.

1.7 Proposed Solution

Designing an integrated approach to minimize the limitations faced by existing two approaches has been proposed which would predict PAIs with more accuracy, specificity and sensitivity.

1.8 Scope

This study has a wide range of scope as it deals with designing a new approach for the identification of PAIs which would help in determining the pathogenic nature of the bacterial species along with their transmission in human. Moreover, it will also provide an insight into the bacterial evolution and emergence of new pathogens helping in the identification of new antibacterial drug targets. This research also focuses on the pathogenic nature of the opportunistic pathogen *Streptococcus sanguinis* which persists to be confusing. *Streptococcus sanguinis* is a member of oral microflora and one of the leading causes of dental plaques and infective endocarditis in humans. Identifying PAIs in *Streptococcus sanguinis* will help in understanding the patho-mechanism of IE suggesting the potential drug targets.

1.9 Aims and Objectives

This research aims at designing an integrated approach for the prediction of PAIs while evaluating the existing approaches with respect to prediction accuracy. The objectives are stated as follow:

1. Accuracy analysis of the two available computational approaches for PAIs identification.

2. Derivation of an integrated approach for the identification of PAIs.

3. Evaluation of the designed integrated approach.

4. Application of the designed integrated approach to the case study of *Streptococcus sanguinis*.

5. Comparison of designed integrated approach and the available two approaches in terms of prediction accuracy.

Chapter 2

Literature Review

This section deals with the review of the literature following computational pipelines established for the prediction of genomic islands. It broadly includes the pipelines established for the GI predictions along with the enumeration of the existing tools based on these pipelines. The chapter also provides insight into the limitations associated with each approach leading towards the gap analysis. A point that is worth mentioning here is the availability of the tools for GI prediction in general rather than PAIs in specific.

2.1 Computational Identification of GIs

Currently, there are two approaches for the identification of the GIs, broadly classified as sequence composition-based and comparative genomics-based approach. The former method relies upon the differences in the genome of host and GIs sequence composition whereas, later relies on detecting unique regions that are absent from the genome of several related isolates.

2.1.1 Sequence Composition Based Approach

This approach relies upon the differences in the sequence composition between the host genome and the GIs. All the genomic regions belonging to the host genome

tends to share some genomic signatures and are supposed to be different from the exogenous genomic regions acquired laterally. Therefore, detecting any genomic region with different gene signatures or content represents lateral mode of transfer defining that region as GI. This approach enables to compare the genomic regions within a single genome to determine various genomic markers. These genomic markers consist of G+C content, dinucleotide frequency, codon usage, tRNA genes, mobility genes, virulence factors, flanking direct repeats, and other characteristics of GIs and PAIs (discussed in section 1.3 and 1.4). This approach provides with the benefit of relying only on query genome sequence and does not require a comparative set of closely-related genomes [13], [18]. Tools based on the sequence composition approach are summarized in Table 2.1. It is worth mentioning here that most of the available tools predicts GIs generically whereas only few such as PIPS identifies PAIS in particular.

AlienHunter is one of the GI prediction tool but its utility is limited to newly sequenced genomes as it relies upon genomic sequences instead of existing annotation or gene position information. It includes highly expressed genes, which results in higher IVOM values and hence the high rate of false-positive results.Underlying principles state a direct relation between IVOM score and GI segments [18].

Centroid employs word frequencies to identify distinct genomic regions within a genome. This program works by dividing the query genome into non-overlapping groups of equal lengths. For each of the given group, centroid calculates the frequencies of all possible words (A, T, G, C) with the given length "m". Average frequency of the whole genome sequence is then calculated and is taken as centroid. Based on these frequencies, outlier regions are detected by estimating the distance between the centroids and the genomic regions considering them to be the GIs [18].

EGID is an ensemble-based algorithm for the identification of GIs. This tool gets output from multiple GI prediction tools (such as AlienHunter, IslandPath, SIGI-HMM, INDeGenIUS, and PAI-IDA), simultaneously, and then generate consensus result based on voting algorithm. Use of multiple tools makes ensemble algorithm

and EGID comparatively better than other GI prediction software in terms of prediction accuracy and precision [18].

Genomic Island Suite of Tools (GIST) is a suite of tools developed with the purpose to provide a user-friendly interface to the researchers. It facilitates third-party programs to get embedded in different GI prediction tools such as EGID. It also provides the users with the facility to download genomic sequences automatically using FTP server of National Center for Biotechnology Information (NCBI)[18].

GIDetector is another ensemble algorithm-based GI prediction tool. It uses J48-based decision tree-bagging model for island prediction. Bagging model was found out to be best fit classifier after testing various similar algorithms such as adaBoost, bagging, multi-boost, and random forest. The model was trained by using IVOM scoreing method, size, and insertion point of the genomic region, genes number per kb, repeats, integrase, phage and non-coding RNA. GIDetector also enables the user to download genome sequences available in public repositories and predicts the GIs by using its training model [18].

Genomic Island Genomic Profile Scanning (GI-GPS) is another GI prediction model and is based on Support Vector Machines (SVMs). This SVM is designed based parameter's information such as codon usage frequency, dinucleotide frequency, codon adaption index; and GC content. GI-GPS trims the whole genome into the specified sized segments and classified them as candidate regions for GIs utilizing SVMs. It then integrates these small segments into one large segment. Several filtering steps are then carried depending upon the segment length and presence of mobile genetic elements. GI-GPS refines predicted GIs boundaries, as final step, by identifying the tRNA genes and repeat elements positions [18].

GIHunter predicts GI by exploiting gene information along with sequence information and inter-genic distances. This tool is modeled over the training set of 113 genomes and a decision tree-based bagging model that predicts GIs. Parameters such as gene information (such as highly expressed genes), phage genes, mobility genes, tRNAs, gene density and inter-genic distance have been utilized to further refine the GIs prediction accuracy [18].

Improved N-mer based Detection of Genomic Islands Using Sequence-clustering (INDeGenIUS) is a hierarchical-clustering based approach for GI detection. It employs the principle of hierarchical clustering to determine the "centroid" by splitting the query genome into "n" overlapping segments of same length. For each of the segment, frequencies of word length "k" are calculated and a vector of 4k words is estimated. The word enumeration process for each group, thus, can generate "n" clusters. Distances for all the possible pairs are then calculated from the centroid utilizing hierarchical clustering approach and groups are merged iteratively into certain number of clusters. Cluster fulfilling the threshold criteria is referred as "major" else "minor" cluster. Relying on the members of "major cluster", the algorithm determines the exact centroid for the host genome and utilizes it for GI identification [18].

IslandPath utilizes DNA signals and genome annotation features for GI prediction. It incorporates certain additional features to improve the prediction accuracy. These features include such as % G+C content for predicted open reading frames, dinucleotide bias for gene-clusters, location of known or probable mobility genes and the location of tRNAs [18].

PAI-IDA utilizes the genomic signature's differences to determine GIs. This tool considers a region as GI if it is different from the host genome in terms of three parameters: G+C content, dinucleotide frequency and codon usage. PAI-IDA contains a small database of known PAI from seven genomes and utilizes this resource to build up the training dataset. Constructed dataset generates list of parameters of linear functions that fetch the different region from the rest of the genome. The discriminant function is improved through iteration by taking additional predicted anomalous regions into account [18].

Pathogenicity Island Prediction Software (PIPS) is the only software targeting the prediction of PAIs in particular. This tool utilizes multiple PAIs signatures for the prediction. These PAIs signatures include G+C content, codon usage deviation, virulence factors, hypothetical proteins, transposases, flanking tRNA, and its absence in nonpathogenic organisms [18].

SIGI-HMM utilizes the fact that each genome prefers different codon usage and as the GIs are of external origin, they can be easily detected. This approach is

called codon bias. The backend algorithm of this tool analyzes the codon usage for each gene and assigns it a score, thus separating the alien genes. Such way of detecting GIs that is codon score-biased, is termed as SIGI. Later on, researchers employ Hidden Markov Model (HMM) approach to SIGI improving the prediction accuracy and approach got the name of SIGI-HMM. While the GIs commonly have a specified length, HMM was utilized to predict GIs on the gene level. Such technique is specifically accurate for the prediction microbial GIs [18].

GIPSy is a standalone GI prediction tool which provides researchers with user-friendly interface. It is based on PIPS, the PAI prediction tool. It exploits commonly used shared genomic signatures such as G+C content, codon usage, presence of transposase genes, virulence factors, metabolism related genes, antibiotics-related genes, flanking tRNA genes and absence of region in other closely-related species, to predict GIs [20].

Zisland explorer is non-supervised algorithm- based tool. It divides the whole genome sequence into fragments by implementing G+C profile tool for further analysis. It exploits homogeneity of sequences within each island and heterogeneity in different regions of genome. As an output, Zisland explorer yields a static plot depicting G+C content throughout the genome, spotlighting the potential GIs entailing their size and total no. of genes present in them [12].

PredictBias predicts PAIs and GIs based on sequence composition-based approach. It analyses the genomic signatures such as presence of insertion elements and virulent genes in order to predict GIs. Virulent genes are predicted by utilizing an internal database called VFDB which searches the database for presence of virulent genes via executing RPS-BLAST (Reversed Position Specific- Basic Local Alignment Search Tool) in the candidate GI regions. Predict Bias uses annotation approach for detecting tRNA and mobility genes like integrases and transposases [12].

GIDetector is based on J48 decision-bagging model for island detection. This model has been trained on the basis of certain features such as IVOM score, insertion point, size of the genomic region, number of genes per kb, repeats, integrase, phage, and non-coding RNA. This tool collects the open source genomic sequences from the websites and detect GIs based on its training model.

In addition to the above discussed tools, number of other tools are also available that follows sequence composition approach and are summarized in Table 2.1.

TABLE 2.1: Overview of the currently used tools based on sequence composition-based approach for GI prediction

No	**Tool**	**Sequence Composition Bias**	**Type**	**URL**	**Year**	**Ref**
1	Alien Hunter	G+C content and oligonucleotides composition	Desktop	https://www.sanger.ac.uk/science/tools/alien-hunter	2006	[12],[16]
2	Centroid	G+C content and oligonucleotides composition	Desktop	http://eraplatform.virtusa.com/tools/centroid	2007	[12],[16]
3	EGID	G+C content, dinucleotide frequencies, trinucleotide frequencies, oligonucleotide frequencies, codon usage	Desktop	http://www5.esu.edu/cpsc/bioinfo/software/EGID/	2011	[16]
4	GIST	Dinucleotide, codon usage, k-mers, IVOM	Desktop	http://www5.esu.edu/cpsc/bioinfo/software/GIST/	2012	[16],[21]
5	GI Detector	G+C content, dinucleotide frequencies, codon usage, tRNA, repeat elements, region length	Desktop	https://omictools.com/gidetector-tool	2010	[16]
6	GIHunter	Inter-genic distance, mobile genes, phage genes, tRNA, gene density	Desktop	http://www5.esu.edu/cpsc/bioinfo/software/ GIHunter/	2014	[16],[13]
7	INDeGenIUS	k-mers	Desktop	Available on request	2010	[16]
8	Island Path	G+C content, dinucleotide, mobile genes, codon usage	Desktop	https://github.com/brinkmanlab/islandpath	2005	[12],[16]
9	PAI-IDA	G+C content, dinucleotide frequencies, codon usage	Desktop	Available on request	2003	[12],[16]

10	PIPS	G+C content, codon usage deviation, virulence factors, hypothetical proteins, transposases, flanking tRNA, and absence in non-pathogenic organisms	Desktop	http://www.genoma.ufpa.br/lgcm/pips	2012	[16],[22]
11	SIGI-HMM	HMM on codon usage	Desktop	http://www.uni-goettingen.de/en/research/185810.html	2006	[12],[16]
12	GIPSy	G+C content, codon usage deviation, transposases genes, class specific factors, tRNA genes, absence in related organisms	Desktop	http://www.bioinformatics.org/groups/?%20groupid=1180	2015	[19]
13	Zisland explorer	G+C content, codon usage, amino acid	Desktop	http://tubic.tju.edu.cn/ZislandExplorer/	2016	[12]
14	Predict Bias	Virulent genes, mobility genes, tRNA genes	Online	http://www.bioinformatics.org/sachbinfo/predictbias.html	2008	[12]
15	PAI-Finder	G+C content, codon usage deviation	Online	http://www.paidb.re.kr/paifinder.php?m=f	2015	[23]
16	MTGI pick	Tetranucleotide	Desktop	http://bioinfo.zstu.edu.cn/MTGI/software.html	2016	[23]

17	MSGIP	Oligonucleotide	Desktop	https://github.com/msgip/msgip	2016	[23]
18	GI-SVM	k-mer frequency	Desktop	https://github.com/icelu/GIPrediction	2015	[23]
19	Sighunt	Tetranucleotide	Desktop	https://www.iba.muni.cz/index-en.php?pg=research–data-analysis-tools–sighunt	2014	[23]
20	GC-profile	G+C content	Online	http://tubic.tju.edu.cn /GC-Profile/ and http://www.zcurve.net/	2014	[23]
21	SVM-AGP (HGT)	GC, oligonucleotide, codon usage, amino acid, position-based frequency	Desktop	http://svm-agp.bioinf.mpi-inf.mpg.de/	2014	[23]
22	GI-POP	GC, oligonucleotide, codon usage, codon adaptation index	Online	http://gipop.life.nthu.edu.tw	2013	[23]
23	CGS (HGT)	k-mers	Desktop	available on request	2012	[23]
24	IGIPT	k-mers (2–6), codon usage, amino acid	Online	http://bioinf.iiit.ac.in/IGIPT/	2011	[23]
25	MJSD	k-mers	Desktop	http://cbio.mskcc.org/aarvey/mjsd/	2009	[23]
26	Design-Island	GC, oligonucleotide	Desktop	http://www.isical.ac.in/rchatterjee/Design-Island.htm	2008	[23]

27	RVM	k-mers, IVOM	N/A	no implementation available	2008	[23]
28	SIGI-CRF (HGT)	Tetranucleotide	Desktop	http://www.uni-goettingen.de/en /research/185810.html	2006	[23]
29	Wn-SVM (HGT)	k-mers	Desktop	available on request	2005	[23]

2.1.2 Comparative Genomics Based Approach

Comparative genomics-based method relies upon the comparison of multiple genome sequences in order to predict GIs. It contrasts the variation of the genetic tree with the respective species tree. This approach works on the principle of detecting a cluster of genes in one genome that is absent from several closely-related genomes. Such regions can easily be detected exploiting whole genome sequence alignment software including MUAVE and Mummer etc. The regions that get aligned over multiple genome sequences are said to be conserved with possibly vertical origin of transfer whereas, regions that are distinctive to a specific isolate is considered to potentially have lateral mode of transfer [4]. This hypothesis is defended by the uneven distribution of GIs within closely related species as shown by degree of sequence divergence in 16S rRNA or other orthologs [12]. This implies that this pipeline relies heavily on the choice of the query genome as well as the availability of subject genomes being tested [4], [18]. Currently, there are two tools available that have specifically been designed for GI prediction based on the comparative genomics approach and are as follow:

MobilomeFINDER is a tool that predicts GIs that are bounded by tRNA - the site of integration for most of the GIs. The algorithm identifies tRNAs that are shared among related genomes and exploits these findings using Mauve tool to detect GIs lying upstream and downstream regions of these orthologous tRNAs. Utilizing this approach makes MobilomeFINDER limited to predicting GIs that are present in the vicinity of the tRNA. Whereas, not all GIs have tRNAs as their

insertion points making such GIs unable to be detected by this tool. Moreover, this tool entails manually inputting the query genome and the subject genome that is vulnerable to inconsistent selection of genomes due to lack of knowledge about intra-genera phylogenetic distances [4], [12], [23].

IslandPick uses method of comparative genomics for the prediction of GIs. It looks for a genomic region that exists in query genome but is absent from several other related species or strains. To detect such regions this tool uses muaveAligner and then BLAST to look for unique region's present in other related genomes. The tool also uses an in-house database named MicroDB at the backend [4].

DarkHorse detects GIs by identifying horizontally acquired proteins and subject them to BLAST analysis. BLAST compares these protein sequences with NCBI non-redundant database while calculating phylogenetic difference between the query and the subject genome. This tool followed combined approach by employing comparative genomics approach along with phylogenetics to identify LGT candidates at various taxonomic levels [23], [24].

TABLE 2.2: Overview of the existing tools based on comparative genomics approach

No	**Tool**	**Type**	**URLs**	**Year**	**Ref**
1	MOBILOME Finder	Desktop	http://db-mml.sjtu.edu.cn/ MobilomeFINDER	2006	[4], [12]
2	IslandPick	Online	http://www.brinkman .mbb.sfu.ca/ mlangill /islandpick/	2008	[4]
3	DarkHorse (HGT)	Desktop	http://darkhorse.ucsd .edu/	2007	[23]

2.2 Databases and Other Computational Resources

Other than the above discussed tools and software, several databases and genome viewers are also available that facilitates the identification of the GIs. These

resources are as follow:

MOSAIC is a whole genome alignments database that contains pre-computed alignments. It facilitates the user to browse the required downloadable alignments and analyze the conserved and variable regions, where variable regions are considered as candidate GIs [4]. Using this resource as GI predictor could be tricky as MOSAIC when identifies variable regions do not consider the condition of insertion into tRNA which deals inversions and translocations between genomes as strain specific regions and hence false positive GI detection [12].

IslandPath is GI viewer which facilitates the users with the graphical user interface for manual detection of GIs. It represents each gene with a small color-coded circle where color assignment is based on the significant level of deviation from the G+C content of the host genome. Genes possessing atypical dinucleotide bias are represented by strikethrough symbol. This tool also illustrates tRNA and mobility genes with different shapes. On the whole, it yields a clickable graphical view of genome spotlighting different genomic signatures related to GIs facilitating manual detection of GIs [4].

Islander is another GI database that consists of 84 GIs along with integration sites for 106 genomes. GIs have been predicted based on tRNA and tmRNA where tRNA and tmRNA are predicted via tRNAscan-SE and BRUCE in a BLAST search. This search filters out the regions that do not possess integrase genes. Islander facilitates researchers to browse GI by name, organisms or site of integration [4].

PAIDB is a database that works on simple principle of homology. It identifies GIs based on their available information provided that is homologous to known PAIs and consider these GIs as potential PAIs. These candidate PAIs are labeled as cPAIs based on their G+C %. This resource allows users to browse GIs based on species, text search or BLAST search [4].

Virulence Factor Database (VFDB) is virulence factors repository that holds curated list of virulence genes as well as PAIs related information about several species. It also enlists candidate virulence genes based on their homology with

known virulence factors which can be searched by species names, text or BLAST and PSI-BLAST search [4].

GIV is Genomic Island Visualization Tool and is customized version of Circos which is one of the eminent genomic information visualization tools. GIV displays the location of the GIs along with the corresponding feature values in genome making study of GIs easier and meaningful [25].

IslandViewer is one of widely used servers for GI predictions and visualizations. It follows an integrated approach by integrating three of the available tools named GIST, SIGI-HMM and IslandPick (as summarized in Table 2.1 and Table 2.2). It is the only integrated webserver to provide integrated results [26].

ICEberg is a database featuring ICEs in 363 bacterial isolates that are extracted either from with the experimental data via literature or directly from GenBank or is predicted using bioinformatics approaches. The browser of ICEberg displays detailed information on 460 ICEs along with genome context view, sequence information as well as respective publications [23].

2.3 Challenges with Current Approaches

2.3.1 Limitations of Sequence Composition Based Approach

Despite of the wide availability of the sequence composition-based tools and techniques designed for GI prediction, the prediction accuracy has still not reached the desirable level. Analysis reveals that sequence composition-based tools predict GIs with accuracy of 82-86% [12]. Causes for this low prediction accuracy are the challenges being faced by this approach that hinder the prediction of all existing GIs. These challenges include:

Amelioration of genomes: Sometimes, sequence composition of PAIs and core genome becomes similar over a period of time known as amelioration of genome. Though the bacterial genomes illustrate deviation in G+C contents, genes within

single species have similar base composition [13]. For example, gen. *Salmonella* lineage showed that former genes acquired via LGT show more similarity in sequence composition with the host genome as compared to recently acquired genes depicting genes amelioration over time [1].

Extremity in the variations of GI: Although, it seems easy to differentiate GIs from rest of the host genome based on the genomic signatures, all genome signatures are not present in each GI. Thus, the mosaic and extremely variable nature of the GIs makes them complex to detect, leading to false negative results. With the advent of evolutionary events, GIs undergo various transformations such as gene loss or rearrangements, consequently, composition, structure and functions of the GIs vary. This variation in the nature of GIs hinder the integration of multiple genomic signatures to be used as predictors [1], [2].

Lack of benchmark datasets: Despite of the availability of multiple GI prediction tools, there still lies lacking in setting up the benchmark datasets for validation of the prediction methods. Though, a few databases, such as Islander, PAIDB and ICEberg etc., are available but they provide information limited to specific types of GI, such as, tDNA-borne GIs (GIs that are inserted at tRNA or tmRNA genes sites) PAIs and ICEs (Integrative and Conjugating Elements) [2].

Uncertain nature of origin: As the GIs adapt their genomic signatures over the period of time, it is complex to determine their origin by comparing them with genomic signatures of other organisms. Likewise, two distantly related organisms may share same codon usage because of tRNA bioavailability [1].

Demarcating the boundaries of GIs: It is difficult to mark the boundaries of GI, as, some of the GIs consists of several kilo bases (kb) while others cover region of hundreds of kb(s), which makes setting up the standard size of GIs, complex [13].

Existence of abnormal sequence composition: Sometimes, the host genomes contain abnormal sequence composition such as ribosomal regions and taking sequence composition-based approach into account, it can lead to the false positive detection of the GI [13].

2.3.2 Limitations of Comparative Genomics Based Approach

Comparative genomics approach relies heavily on the genome in question and the related genomes. Inclusion of distantly related genomes could make the alignment complex yielding false positive predictions. Likewise, including more recently diverged genome in the analysis could yield more robust results. Moreover, including too closely related specie results in un-identification of GIs that has been inserted prior to divergence of the genomes. As the approach is based on multiple related sequenced genomes, it becomes unsuccessful for a specie that has no closely related sequenced genomes to perform the comparison [4].

Chapter 3

Material and Methods

This chapter provides an overview of the methodology adopted for analysis of the existing approaches as well as for designing an integrated approach for the prediction of genomic islands.

3.1 Identification of Computational Approaches for Prediction of GIs/PAIs

Research on the computational prediction of GIs have made substantial progress. There are number of techniques and tools that have been proposed and designed for identifying GIs and can be found in the published literature (summarized in Table 2.1 and Table 2.2 in literature review section). Literature survey has been performed in order to figure out the computational approaches using literature searching platforms like Google Scholar, PubMed and Polysearch2. Keywords like "computational identification of GIs", "Genomic islands", "Pathogenicity islands", "Horizontal gene transfer", "Genome plasticity" were used.

3.2 Extraction of Pipelines from Existing Tools

To extract a pipeline for the sequence composition-based as well as the comparative genomics-based identification of GIs, analysis of the existing tools and techniques was performed. Analysis yielded that most of the tools available for the sequence composition-based approach exploits a set of maximum four genomic signatures to predict GIs (summarized in Table 2.1 in literature review section). Whereas, GIPSY is the only tool that follows detailed pipeline and analyzes seven to eight genomic signatures before it predicts GIs and PAIs [20]. Besides GIPSY, PIPS is another detailed analysis tool that solely exists for the prediction of PAIs in specific. Later on, in 2016, GIPSY was revised and integrated features of PIPS [20],[27], enabling GIPSY to predict GIs and PAIs both. Therefore, the pipeline for sequence composition-based analysis was extracted by keeping GIPSY as reference.

On the other hand, literature represents several different approaches for performing comparative genome analysis for the identification of GIs. Most of the cited work utilizes MuaveAligner followed by BLAST search to identify the "conserved" and "unique" regions of the query genome by comparing it with closely related organisms. In case of PAIs, the query genome is compared with the closely related non-pathogenic species to identify the unique regions responsible for causing pathogenicity, hence, PAIs [2], [12], [18].

3.3 Accuracy Analysis of Existing Approaches

3.3.1 Selection of Organism for Case Study

For the evaluation of prediction accuracy of the existing computational approaches, *Streptococcus sanguinis* was chosen as a case. Among *S. sanguinis* strain SK36 was selected because of its important role as an opportunistic pathogen of infective endocarditis. *S. sanguinis* SK36 is a gram-positive bacterium, normally constituting the oral microflora in human [28], [29] and is an active colonizer of dental plaques [30]. It follows the route of mouth and if gets enter into the blood stream through a minor cut or a wound causing infective

endocarditis (IE). IE is a life-threatening endovascular infection caused by adherence of bacteria from the bloodstream to the damaged heart valve [28]. It is characterized by the vegetative growth embedded by infection-causing microorganisms along with fibrin and platelets at the site of infection [31], [32], [33]. This vegetation provides microorganisms with a site to embed and multiply, disrupting normal patterns of blood circulation within the heart, whereas, the intensity and destruction of the tissue depends upon the bacterial species [34], [35]. In spite of the recent advancements in the medical, surgical and critical care interventions, IE continues to remain a leading cause of morbidity and mortality, not only in West but also in Asian countries like Pakistan where it has the mortality rate of 27.3% [36]. Higher incidence rate between IE and isolation of *S. sanguinis* has been reported in patients undergoing dental procedures [30], [37].

While the world of literature is rich in research spotlighting various aspects of IE with *S. sanguinis* SK36 being the most well-known pathogen, the patho-mechanism of this opportunistic pathogen remains to be revealed. Moreover, no information is reported about the occurrence of LGT events and the presence of PAIs in the genome. Therefore, there is a need to investigate the origin of pathogenicity in the strain by identifying and investigating the PAIs it possesses.

3.3.2 Sequence Composition-Based Approach

Sequence composition-based method exploits different sequence based genomic features (discussed in section 2.2.1). It is the most widely used approach for genomic islands prediction as it claims to be independent of the related genome selection. After the extraction of the pipeline for this approach, a tool (PIPS) was identified that can perform all of the mentioned steps. Steps for the pipeline followed for sequence composition-based approach is summarized in Figure 4.1 (results and discussion section). This pipeline has been extracted from the two most commonly used GI prediction tools named GIPSY and PIPS. These two tools were ideal for extracting pipeline as both of these tools use maximum of the documented GI and PAI properties. Therefore, GIPSY tool(http://www.bioinformatics.org/groups/?% 20groupid=1180) was exploited along with the

standard parameter settings for the implementation of the pipeline as it already has integrated multiple tools for the detection of the genomic features and hence provides identification of multiple genomic signatures via one platform. GIPSY accommodate two genome annotation files as input in either Genbank or ensemble format. One genome is required as query genome, whereas, the other one is used as subject genome. For the identification of the PAIs specifically, the subject genome needs to be most closely related non-pathogenic strain of the query genome [20]. *S. sanguinis* has 62 strains, among which only 6 are complete. Among these 6 strains, 3 strains (SK36, NCTC11085, NCTC10904) have confirmed pathogenicity, whereas rest of the 3 (NCTC11086, NCTC7863 and NCTC3168) are newly identified strains with unknown pathogenicity. Therefore, we used NCTC11086, NCTC7863 and NCTC3168 strains as subject genomes one by one and identified PAIs generating True Positive and False Positive datasets.

3.3.3 Comparative Genomics-Based Approach

Very few tools employ comparative genomics-based approach. Among the tools summarized in Table 2.2 in literature review section, DarkHorse was not available due to maintenance purpose (from April 2019 to date), MobilomeFINDER has become obsolete, whereas, IslandPick was the only choice left. IslandPick has recently been integrated into IslandViewer, therefore, the putative GIs in SK36 were predicted by the IslandViewer4 software tool which has the highest prediction accuracy i.e. 88% [23] and involves three different GI identification approaches: sequence composition-based approaches using SIGI-HMM and IslandPath-DIMOB, and the comparative genomics approach using IslandPick. As the focus is on the comparative genomics- based approach, results from only IslandPick method were considered.

As the comparative genomics-based approach for the identification of PAIs is based on the selection of multiple genomes to perform comparative analysis with respect to query genome, selection of comparative genome set is an important aspect. IslandPick (http://www.brinkman.mbb.sfu.ca/ mlangill/islandpick/) does not allow the user to select comparison genomes by choice and choses

genome set by default. Therefore, analysis was continued with its default selection. IslandPick, by default makes inter-specie selection of comparison genomes based on minimum phylogenetic distance from the query genome. For this analysis, the selected comparison genomes were all complete genomes and included *Streptococcus gordonii* strain KCOM 1506 (ChDC B679), *Streptococcus constellatus* subsp. *pharyngis* C232, *Streptococcus constellatus* subsp. *pharyngis* C818, *Streptococcus* sp. oral taxon 431, *Streptococcus parasanguinis* FW213 and *Streptococcus pneumoniae* AP200 with distances 0.116, 0.194, 0.194, 0.209, 0.211 and 0.213 respectively.

For *Streptococcus sanguinis* SK36, the pre-computed GI analysis as well as the manually predicted GIs were taken into account. Inclusion of both manual and pre-computed results is supported by the reason that the system was last updated two years back (2017) and has not undergone any recent up-gradation event and therefore, may include variable results now.

Furthermore, the predicted GIs were then further filtered by removing GIs with genomic length less than 10 kb. Likewise, the predicted putative GIs from IslandViewer4 (http://www.pathogenomics.sfu.ca/islandviewer/)were further inspected manually using PAIDB (http://www.paidb.re.kr/about-paidb.php) [38].

3.4 Proposing an Integrated Approach for PAIs Prediction

An integrated approach was proposed based on the survey of previous studies on GI prediction in various bacterial species. Studies included in the survey were mandatory to have followed computational means of prediction. These studies majorly include the reviews conducted to compare and evaluate the performance of the existing tools and techniques available for GIs prediction [20], [27], [38], [40], [43], [44].

An integrated approach refers to integrating both of the existing approaches i.e. sequence composition as well as comparative genomics-based approach such that

limitations of one approach gets compensated by the other and vice versa. Previously, such simple integration of approaches is implemented in IslandViewer4 (http://www.pathogenomics.sfu.ca/islandviewer/). It is the only tool that follows integrated pipeline by integrating 2 sequence composition (SIGI-HMM and IslandPath-DIMOB) and 1 comparative genomics-based (IslandPick) tools [23]. Though this software has highest prediction accuracy, it faces two general limitations. Firstly, for sequence composition-based approach, it exploits only four genomic signatures i.e. G+C content deviation, codon usage, presence of tRNA and frequency of dinucleotides which narrow downs the decision criteria for GIs. Second is the auto-selection of comparative genome set by IslandPick (https://www.brinkman.mbb.sfu.ca/mlangill/islandpick/index.html) for comparative genome based approach.IslandPick was designed initially to allow the users to chose comparative genome set of their own choice as well as to use auto-selections but currently, the server does not provide the facility of genome selection anymore and the user is compelled to work with auto-selection thus limiting the analysis.

In this study, a modified integrated approach was proposed with two basic modifications minimizing the limitations caused by existing approaches. First modification was proposed in sequence composition-based approach based on the observation that different tools uses different genomic signatures to make decision about a GI. This deduces that different tools vary in stringency criteria for decisions making. It was also observed that, maximum number of genomic signatures exploited by any sequence composition-based tool except GIPSY is four (as shown in Table 2.1). Decision based on such smaller number of genomic signatures is proportional to lose stringency criteria resulting in higher probability of false prediction results. Broadening up the subset of genomic features, could therefore be looked upon as a way to increase the stringency on decision making, minimizing the probability of false predictions.

Second modification was proposed in comparative genomics approach, based on the fact that for an organism with related strains of unknown pathogenicity, k-means clustering approach could be used to cluster pathogenic and non-pathogenic strains. This clustering is based on the related known virulence genes found in an organism. The selection of comparative genome set or the

subject genome based on such clustering method provides more accurate selection of comparative genomes rather than assumption or brute force selections as this approach is greatly influenced by the selected comparison genome set.

Such modified integrated approach helps to minimize the currently faced challenges by the GI prediction pipelines, minimizing the false prediction results and solves the problem of predicting GIs/PAIs for organisms that do not have a completely known pathogenic basis. The proposed integrated approach is shown in Figure 4.6.

3.5 Application of Integrated Pipeline to Case Organism *S. sanguinis* SK36

Proposed integrated approach was applied to the case organism *S. sanguinis SK36* with the aim to predict GIs and PAIs more precisely, while, validating the results obtained by sequence composition as well as comparative genomics-based approaches in parallel. Application of pipeline and generation of results was followed by manual validation to counter-check the positive and negative results. In order to identify the PAIs, virulence genes related to IE were determined through VFDB (http:// www.mgc.ac.cn/VFs/) and published literature. Resultant PAIs were then scanned for these pre-determined virulence genes. PAIs containing the searched virulence genes were considered as confirmed PAIs, while, those missing these genes were removed from the list of PAIs and were condiered as GIs. Results obtained from both approaches were taken into account and compared to ensure non-redundant results. Previously generated datasets were re-evaluated for validation purpose. Integrated dataset is then validated using the computational resource PAIDB (http://www.paidb.re.kr/about-paidb.php).

3.5.1 Determining Non-Pathogenic Group of *S. sanguinis* Strains

3.5.1.1 Enlisting Virulence Determinants

Reported virulence determinants were enlisted with the help of previously published literature [15],[29],[30],[32],[34],[41],[42],[43],[44],[45],[46] and VFDB (http:// www.mgc.ac.cn/VFs/main.htm). Information was collected exploiting Polysearch2 (http://polysearch.ca/) and PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) platforms using multiple keywords such as infective endocarditis, endocarditis, bacteremia, dental plaque, oral microbiome and biofilms.

3.5.1.2 Clustering of *S. sanguinis* Strains

Applying comparative genomics approach requires selection of comparative genome set. In order to find out the PAIs, it is worth comparing a pathogenic strain with a set of non-pathogenic strains within a specie so that the unique regions causing pathogenicity could be identified.

With the selected case organism *S. sanguinis* SK36, several strains are available but pathogenic nature of many of the strains is unknown which makes it difficult to select an accurate genome set for comparison. Whereas, selection of in-accurate genome set for comparison can lead to highly false results. Therefore, in the proposed integrated approach, pre-processing for such organisms is suggested which states adopting the patho-genomic comparison approach to determine the non-pathogenic strains of the respective specie.

In the discussed case, IE virulent determinants are critical for *S. sanguinis* specie for its pathogenicity. Therefore, a search of pre-enlisted virulent factors among *S. sanguinis* strains could differentiate pathogens from non-pathogens. Hence, a patho-genomic comparison approach was used to examine the conservation of the virulent factors within *S. sanguinis* species. 62 strains of *S.sanguinis* are publicly available in NCBI genome database (https://www.ncbi.nlm.nih.gov/genome/genomes/1345). Out of these 62 strains, 6 are complete, 22 are scaffold and 34 are contigs. For the study, only complete strains have been

used and were compared through Nucleotide BLAST [47] and UniProt [48] to check the presence of the enlisted virulent factors.These strains included NCTC11085, NCTC11086, NCTC7863, NCTC10904, NCTC3168 and the query genome SK36. These strains were then clustered into two group i.e. pathogenic strains and non-pathogenic strains, using k-mean clustering approach, provided that a strain is IE pathogen only if it possesses majority of the enlisted virulent factors whereas strains which are void of the virulent determinants were considered non-pathogenic. Pathogenic cluster of strains was further divided into sub-clusters of "severe" and "moderate" pathogens depending on the criteria defined as follow: severe pathogens - least distance a strain has with reference strain SK36, more pathogenic it would be. Moderate pathogens – more the distance a strain has with the reference strain, more moderate pathogen it would be. Distance was determined based on the number of virulence determinants a strain possess in reference to SK36 and the criteria was developed based on the idea of Euclidean distance shown in equation 3.1.

$$d(i,j) = Sqrt(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2) \tag{3.1}$$

[48], where x_i and x_j refers to the total number of virulence determinants possessed by SK36 and other strains respectively. SK36 was fixed as centroid object for clustering, thus x_i remained fixed, and the number of desired clusters was two.Strains included were NCTC11085, NCTC11086, NCTC7863, NCTC10904 and NCTC3168.

3.5.1.3 Selection of Subject Genome and Integration of Results

Once, the non-pathogenic cluster of the *S. sanguinis* strains was identified, it became easier to choose one non-pathogenic genome from the cluster and use it as subject genome for sequence composition-based analysis. This determination of non-pathogenic strains could also be exploited when predicting GIs/PAIs by comparative genomics approach depending upon the context of the study as if the comparison is needed to be carried out at inter-specie or intra-specie level.

From the previous step (section 3.5.1.2), NCTC7863 was chosen as most closely non-pathogenic strain because as compared to other strains, it is a complete genome, so GIs/PAIs set predicted using NCTC7863 in section 4.3.1 was selected and integrated with the set of GIs/PAIs acquired by comparative genomics approach. Resultant dataset was then processed and evaluated for accuracy.

3.6 Accuracy Evaluation of Approaches Under Study

In order to evaluate the accuracy of discussed pipelines for the identification of GIs/PAIs (section 3.3), statistical measure of accuracy was chosen. Accuracy refers to how close a measured value is to the actual value. This statistical measures is the indicative of the correctness of the designed model i.e. sensitivity and specificity. Higher the value of accuracy, more accurate are the results. It is measured as follow [50]:

$$Accuracy = \frac{TP + TN}{Total} \tag{3.2}$$

Where, TP refers to true positive results and are indicative of truly predicted PAIs, FN stands for false negative results and refers to the GIs that were actually PAIs but have been predicted falsely. Likewise, FP stands for false positive results that refers to those falsely predicted PAIs that do not contain any virulence gene and are actually GIs.

For the accuracy evaluation of the approaches, positive and negative datasets were constructed. Constructing datasets for a genome with no GI/PAI information in literature or GI database was a challenging task for which a simple yet insightful approach was adopted. For dataset constructions, genome viewer PAIDB was exploited. Proposed approach for dataset construction was based on the most significant characteristics of PAIs i.e. presence of virulence genes as virulence genes make a PAI different from a GI.

GIs and PAIs predicted by the existing approaches were used as the basis for this search. Detected GIs and PAIs for the virulence genes were examined , as

a GI must contain at-least one virulence genes in order to be considered as PAI [51], if a putative GI showed presence of virulence gene or its homolog, it has been considered as "False Negative", else, "True Negative". Likewise, if a PAI contained virulence gene, it has been considered as "True Positive" else "True Negative".

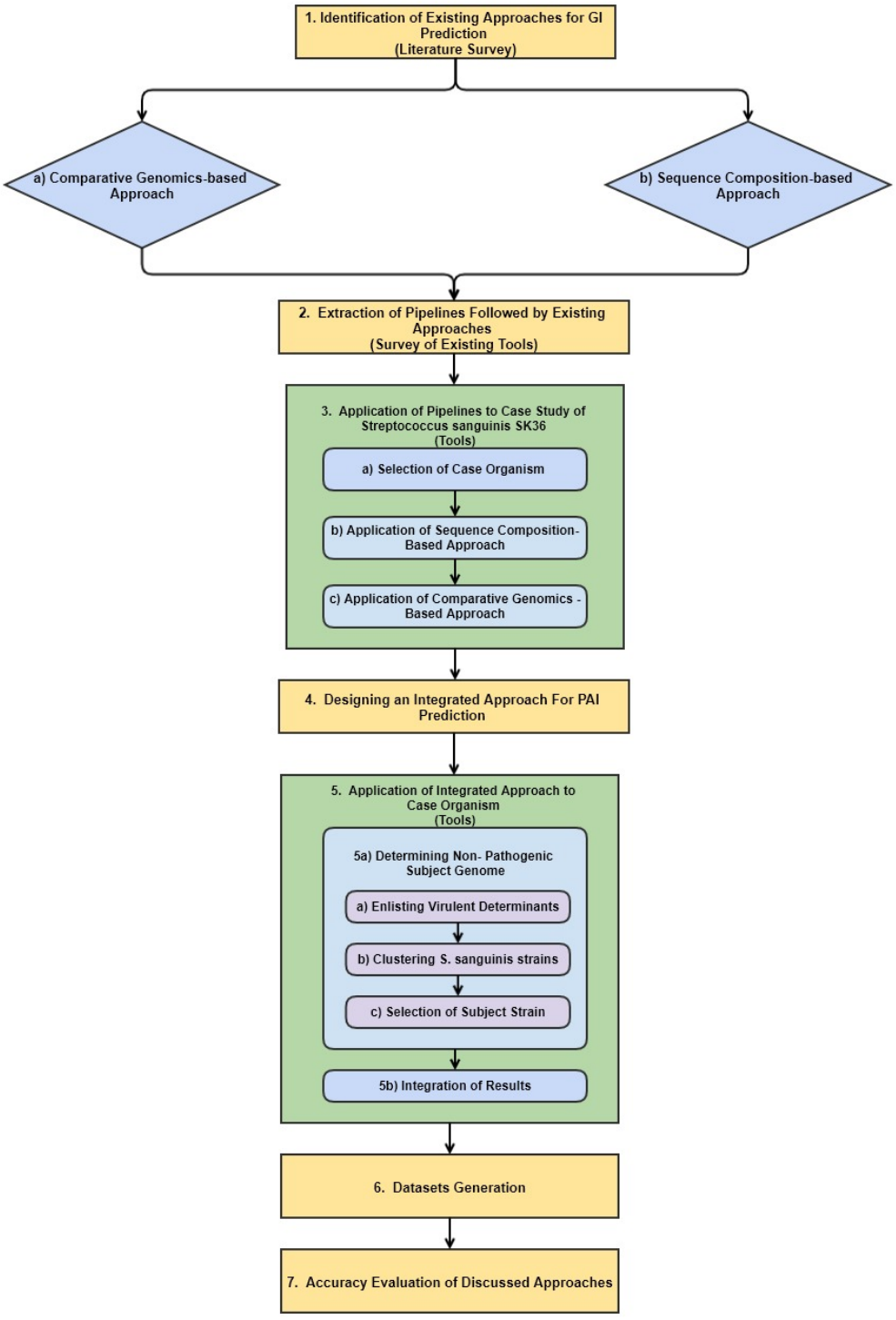

FIGURE 3.1: Methodology pipeline for designing an integrated approach to predict PAIs

Chapter 4

Results and Discussions

This chapter covers the results obtained and discussion in context of aims and objectives of the study. The study aims at evaluating the existing approaches for the identification of GIs and proposing an integrated approach for the identification of the PAIs in special reference to the case of *Streptococcus sanguinis* SK36. The integrated approach is proposed to minimize the inherited limitations of the sequence composition as well as comparative genomics-based approaches while supporting the idea via the case of an opportunistic pathogen with no identified PAIs in the literature. Set objectives included identification of the existing approaches for GIs identification, their analysis and extraction of the methodology pipelines, applying these pipelines to the case study while evaluating their accuracy and finally heading towards integrated approach.

4.1 Identification of Computational Approaches for Prediction of GIs/PAIs

The literature survey revealed that number of tools and techniques are available for identifying GIs (summarized in table 2.1 in literature review section). It was found out that existing tools usually exploit two most indicative features of horizontal origin of GIs; the distinctive sequence composition and the sporadic phylogenetic distribution. Based on these two features, the prediction approaches

fall into two broad categories; sequence composition-based approach and comparative genomics-based approach [2],[4],[12],[18],[15],[17],[23],[27],[52].It was further found out that in-spite of the wide availability of the computational resources known for detecting GIs, very limited are available for the identification of PAIs in particular [52].

4.2 Extraction of Pipelines from Existing Tools

In order to understand the pipeline followed by the sequence composition as well as comparative genomics-based tools, different tools were analyzed and the implemented approaches were extracted. The extracted pipeline for sequence composition-based approach is seven step method. Step 1 deals with the calculation of G+C content deviation of different genes of query genome while keeping another genome as subject. Step 2 deals with codon usage deviations, differentiating two genomic regions of diverse origin. Step 3 detects presence of transposases in the query genome. Step 4 searches for virulence determinants in the query genome. Step 5 looks for the unique genomic regions in query genome such that such regions are absent from same genus or related species using BLAST. Step 6 detects presence of tRNA genes which act as insertion sites for PAIs. Whereas, step 8 deals with analysis of the results acquired from all previous steps while taking decision on GIs and PAIs identification. Extracted pipeline is shown in Figure 4.1.

Contrary to that, pipeline extracted for comparative genomics tool is simpler. It requires acquisition of whole genome sequences from group of closely related organisms as step no 1. Step 2 requires performing whole genome multiple alignments such that the unique and conserved regions are detectable. In the last step, such unique regions that are present in query genome while being absent in other related genome are identified as putative PAIs and are subjected to further analysis. Extracted pipeline for comparative genomics-based approach is shown in Figure 4.2.

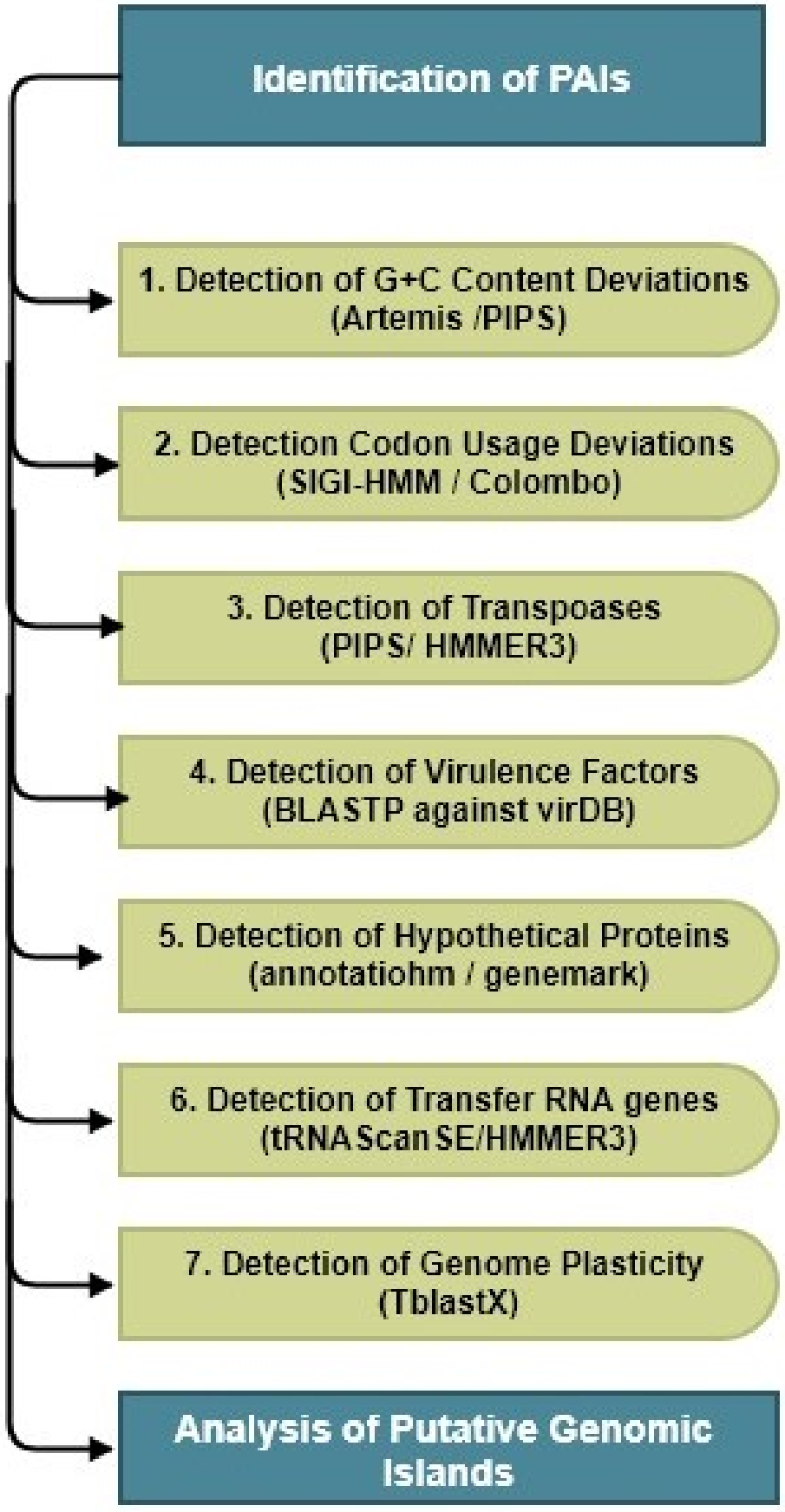

FIGURE 4.1: Manual pipeline for sequence composition-based approach of PAIs identification

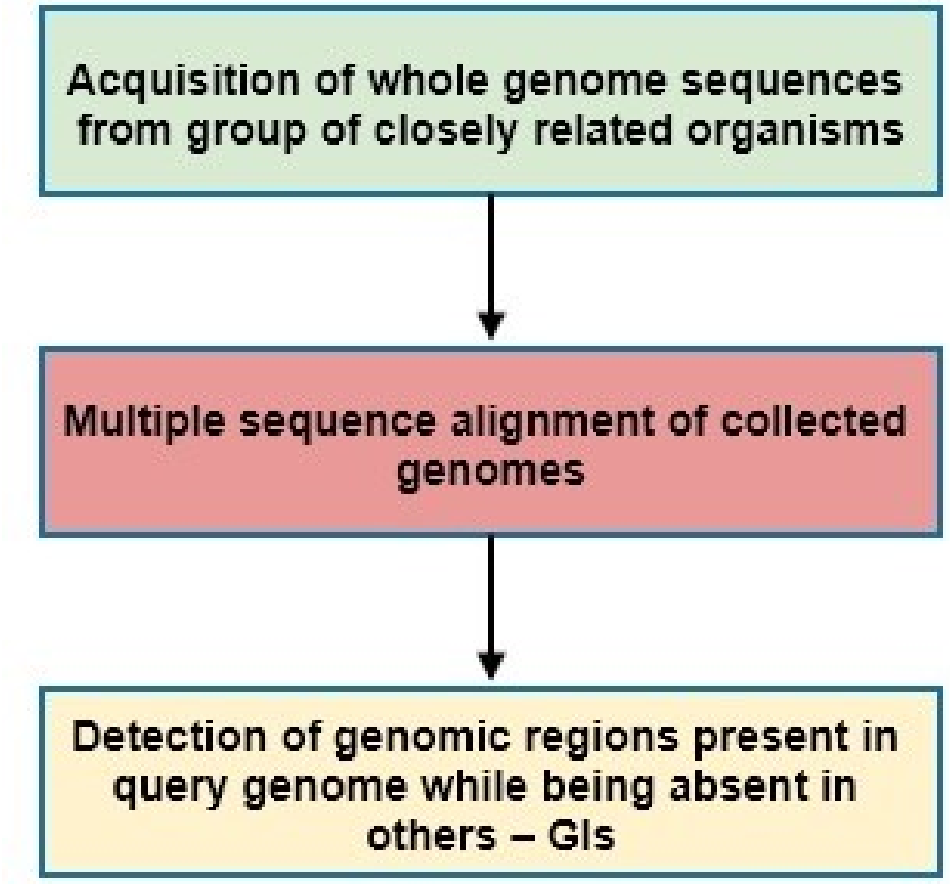

FIGURE 4.2: Extracted pipeline followed by comparative genomics-based approach

4.3 Accuracy Analysis of Existing Approaches

Streptococcus sanguinis SK36 has a circular genome comprising of 2,388,435 bp whereas, the G+C content of the genome has been mentioned to be 43.4% which is higher as compared to other gen. *streptococcal* species. The genome encodes 2,274 proteins along with 61tRNA and four rRNA [53]. Certain LGT events have also been reported in the genome via PAIDB. The circular genome map of the organism is shown in Figure 4.3.

Streptococcus sanguinis being the member of oral microflora experiences significant fluctuations in the environmental factors such as pH level, oxygen concentration or osmolarity especially when growing in dental plaque. When this bacterium enters blood stream thorough a minor cut or a wound, it encounters even greater

shift in the external environment i.e. bloodstream environment. This study has hypothesized that the adaptability and evolution of *streptococci* in response to the ever-changing environments within the human body might have been interceded via the LGT events resulting in GIs transfers. GIs in bacteria can configure various advantages to the bacterium depending upon the genes they harbor, whereas focus of this study is on the virulence factors harboring GIs- the PAIs [38]. Therefore, horizontally transferred GIs and particularly PAIs in the genomes of *S. sanguinis* SK36 were predicted using the most widely cited approaches one by one.

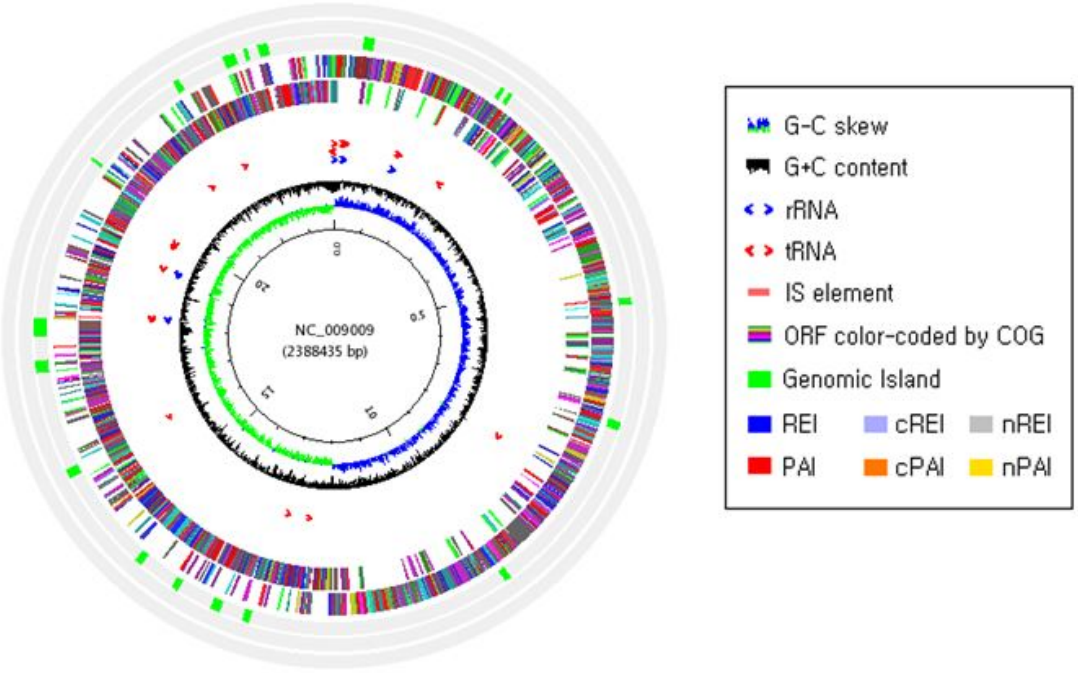

FIGURE 4.3: Circular genome map of the *Streptococcus sanguinis* SK36

Starting from the outside, the tracks show (i) position of genomic islands; (ii & iii) ORF on positive and negative strands; (iv) tRNA; (v) rRNA; (vi) G+C content; (vii) GC skew.

4.3.1 Predicting PAI by Sequence Composition-Based Approach

For the prediction of PAIs in *S. sanguinis* SK36 by sequence composition-based approach, three different non-pathogenic strains of *S. sanguinis* were used as

subject genomes and different results were obtained. Retrieval of different results each time the subject genome is changed indicates that contrary to what this approach claims (being independent of related genome), it significantly depends upon the subject genome.

Using GIPSY, the total G+C content of *S. sanguinis* reveals to be 1036586.0 with the genome size of 2388435.0 bp whereas G+C content in percentage determines to be 43.4%. However,the genome statistics of the subject genomes is summarized in Table 4.1 and Figure 4.4.

TABLE 4.1: Summary of results observed by keeping NCTC11086, NCTC7863 and NCTC3168 as subject genome

Genomic Features	**Results with NCTC11086**	**Results with NCTC7863**	**Results with NCTC3168**
G+C content Deviation	248 genes	248 genes	248 genes
Codon Usage Deviation	242 genes	242 genes	242 genes
No. of Transposases	17 genes	17 genes	17 genes
No. of tRNA	62 genes	62 genes	62 genes
No. of Virulence Genes	717 genes	717 genes	717 genes
No. of putative GI	7	10	21
No. of Putative PAI	1	2	7

Using the sequence composition-based approach, 32 GIs and 9 PAIs were predicted in total. Out of these 32 GIs, 4 GIs (GI1, GI2, GI3 and GI7) and 1 PAI (PAI1) remained conserved while keeping NCTC11086 and NCTC7863 as subject genome (Table 4.2). Whereas, 2 GIs were found conserved in SK36 when NCTC7863 and NCTC3168 were kept as subject genome.

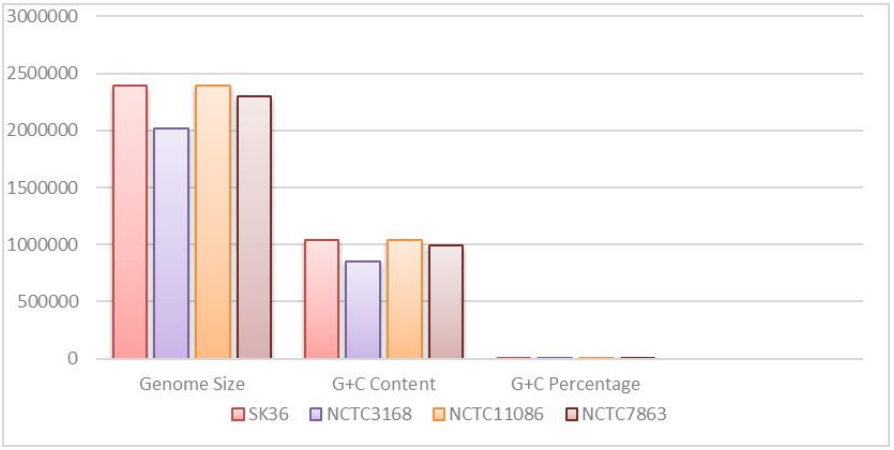

FIGURE 4.4: Genome statistics of query and subject genomes

TABLE 4.2: Summary of predicted GIs and PAIs in the genome of *S. sanguinis* SK36 by following sequence composition-based approach

Genomic Island	**Size (bp)**	**With NCTC11086**	**With NCTC7863**	**With NCTC3168**
Putative GI1	6671	*	*	
Putative GI2	11196	*	*	
Putative GI3	8805	*	*	
Putative GI4	8174	#		
Putative GI5	9365	#		
Putative GI6	7405	#		
Putative GI7	8496	*	*	
Putative GI8	10717		#	
Putative GI9	12677		*	*
Putative GI10	5647		#	

Putative GI11	14041		*	*
Putative GI12	9115		#	
Putative GI13	6627		#	
Putative GI14	6864			#
Putative GI15	10556			#
Putative GI16	14792			#
Putative GI17	16658			#
Putative GI18	8257			#
Putative GI19	7662			#
Putative GI20	9872			#
Putative GI21	5596			#
Putative GI22	14305			#
Putative GI23	15498			#
Putative GI24	11335			#
Putative GI25	8805			#
Putative GI26	8538			#

Putative GI27	10133			#
Putative GI28	13886			#
Putative GI29	17251			#
Putative GI30	19309			#
Putative GI31	15756			#
Putative GI32	7430			#
Putative PAI1	10106	*	*	
Putative PAI2	11043		#	
Putative PAI3	13227			#
Putative PAI4	18207			#
Putative PAI5	32944			#
Putative PAI7	6374			#
Putative PAI8	32944			#
Putative PAI9	12011			#

Conserved GIs and PAIs identified by using different subject genome are marked with '*' whereas subject specific are marked with #.

4.3.1.1 Calculating G+C Content Deviations

G+C content deviations for the query genome were calculated. G+C content deviation is a significant genomic signature as it distinguishes the PAIs from the host genome [18]. On average G+C content ranges from 25-75% for bacterial species whereas it is reduced to 40-60% in case of pathogenic species. As PAIs are laterally transferred, they retain base composition of their donor species, which accounts for the variation in G+C content. However, reason for such variation remains unknown [19]. With all the three subject genomes, SK36 showed similar deviations which validated the results. Out of total 2270 genes, SK36 shows no deviation for 2022 genes whereas 248 genes with deviated G+C contents have been observed. Out of the total 248 genes, 201 genes reached the lower limit of set value for standard deviation whereas 47 reaches the higher limit threshold. Reference values for upper and lower limit were set by using standard deviation of 1.5 for both query and subject genomes and were defined as 50.56 and 36.23 respectively.

4.3.1.2 Codon Usage Analysis

Codon usage was determined for SK36 via integrated SIGI/HMM approach. As each genome prefers different codon usage and the PAI and host genome have heterogenous origin, the codon usage for the two genomes vary. Therefore, this genomic signature could be exploited as reliable parameter to detect PAIs [16]. For calculating codon usage deviations, the sensitivity parameter was set to 0.95 standard value i.e. highest possible standard value [20]. Again, for all three subject genomes, SK36 show similar results where 2022 genes showed normal codon usage and 248 showed deviations indicating the possibility of LGT events.

4.3.1.3 Detection of Transposase Genes

After the codon usage analysis, transposase genes were searched in the genome. Transposases belongs to the category of functional mobility genes and are required for the excision or insertion of DNA regions into the genome [17]. Transposases were searched via HMMER3 tool which searches the genome profile

against transposase database. The e-value used in the analysis was set to standard 1E-04 [20]. During the prediction, 17 transposases genes were found in SK36 with sequence reported threshold of E-value less than 0.0001. Presence of transposases in SK36 supports the concept of LGT events in the genome.

4.3.1.4 Detection of Virulence Genes

Detection of virulence genes in the genome is critival for identifying PAIs. PAIs in pathogenic bacteria encodes virulence genes that greatly differentiate them from non-pathogenic bacteria. These virulence genes fall into multiple functional groups such as adherence factors, siderophores, exotoxins, invasion genes and type III and IV secretion systems [17], [18]. Therefore, detection of virulence genes can be exploited as clue for detecting PAIs [18].Virulence determinants were detected based on the protein similarity searches performed by blastp algorithm against mVIRdb [20]. With all three subject genomes, total 717 virulence factors were determined in the SK36. Standard value used for analysis was set to 0.000001.

4.3.1.5 Detection of Unique Regions

Unique genomic regions in a query genome are indicative of LGT events. These regions were determined in comparison to subject genomes by performing reciprocal BLASTs between CDSs of the two genomes. Approach followed to predict LGT events and inferring gene synteny includes predicting commonly shared orthologous genes [20]. In view of this, 301 unique genes were detected in SK36 that do not have orthologs in NCTC3168. Whereas, 214 and 220 unique genes were detected in comparison to NCTC7863 and NCTC11086 respectively. Presence of these unique genomic regions further validated the occurrence of HGT events.

4.3.1.6 Detection of tRNA Genes

Transfer RNAs are known to be the landmarks for the integration of exogenous DNA. Extrachromosomal genetic elements usually carry tRNA or part of them

and is therefore suggestive of extrachromosomal elements integration into the host genome by homologous recombination between tRNA genes of the chromosome and their extrachromosomal counterparts [17], [19]. Presence of tRNA genes was detected via integrated HMMER3 tool which perform query genome searching against database of bacterial tRNA genes named tRNAdb. Standard value was set to 0.0001 during the analysis [20]. This analysis revealed 62 tRNAs in the query genome which includes 1 broken tRNA which is indicative of LGT event.

4.3.1.7 Summarizing Results

At the end, results from all previous steps (1-6) were summarized and analyzed in order to determine the PAIs. One PAI was detected in SK36 by keeping NCTC11086 as subject genome. Whereas, 2 and 7 PAIs were identified when NCTC7863 and NCTC3168 were kept as subject genome respectively. This analysis also yielded some GIs, details of which are summarized in Table 4.3 to 4.5.

TABLE 4.3: GIs and PAIs observed by selecting NCTC11086 as subject genome

Putative GIs and PAIs	**G+C Dev.**	**Codon Usage**	**Virulence Factors**	**Hyp. Proteins**	**Gene Comp.**
Putative GI 1	57%	42%	0%	85%	SSA_0228-SSA_0234
Putative GI 2	38%	76%	7%	76%	SSA_0553-SSA_0565
Putative GI 3	54%	100%	0%	100%	SSA_1284-SSA_1296
Putative GI 4	75%	100%	0%	83%	SSA_1327-SSA_1338
Putative PAI 1	0%	20%	50%	30%	SSA_1359-SSA_1368
Putative GI 5	50%	58%	16%	41%	SSA_2288-SSA_2295
Putative GI 6	33%	26%	13%	26%	SSA_2288-SSA_2295
Putative GI 7	50%	87%	25%	50%	SSA_2288-SSA_2295

TABLE 4.4: GIs and PAIs observed by selecting NCTC7863 as subject genome

Putative GIs and PAIs	**G+C Dev.**	**Codon Usage**	**Virulence Factors**	**Hyp. Proteins**	**Gene Comp.**
Putative GI 1	62%	50%	12%	75%	SSA_0228-SSA_0235
Putative GI 2	38%	76%	7%	76%	SSA_0553-SSA_0565
Putative PAI 1	28%	14%	42%	28%	SSA_1099-SSA_1105
Putative GI 3	14%	28%	14%	14%	SSA_1143-SSA_1149
Putative GI 4	30%	100%	0%	90%	SSA_1246-SSA_1255
Putative GI 5	54%	100%	0%	100%	SSA_1284-SSA_1296
Putative PAI 2	0%	20%	50%	30%	SSA_1359-SSA_1368
Putative GI 6	85%	100%	0%	100%	SSA_1387-SSA_1393
Putative GI 7	46%	60%	13%	40%	SSA_1812-SSA_1821
Putative GI 8	45%	45%	18%	36%	SSA_2025-SSA_2032
Putative GI 9	16%	50%	8%	58%	SSA_2269-SSA_2276
Putative GI 10	50%	87%	25%	50%	SSA_2288-SSA_2295

TABLE 4.5: GIs and PAIs observed by selecting NCTC3168 as subject genome

Putative GIs and PAIs	**G+C Dev.**	**Codon Usage**	**Virulence Factors**	**Hyp. Proteins**	**Gene Comp.**
Putative GI1	40%	20%	10%	20%	SSA_0044-SSA_0051
Putative PAI 1	12%	0%	50%	50%	SSA_0201-SSA_0207
Putative GI 2	55%	44%	22%	66%	SSA_0227-SSA_0235
Putative PAI 2	6%	0%	50%	50%	SSA_0393-SSA_0410
Putative GI 3	31%	62%	18%	62%	SSA_0553-SSA_0568
Putative GI 4	4%	18%	27%	27%	SSA_0915-SSA_0934
Putative PAI 3	44%	66%	55%	66%	SSA_0947-SSA_0957
Putative GI 5	28%	42%	28%	57%	SSA_1166-SSA_1172
Putative GI 6	33%	66%	0%	50%	SSA_1229-SSA_1234
Putative GI 7	30%	100%	0%	90%	SSA_1246-SSA_1255
Putative GI 8	40%	90%	0%	90%	SSA_1286-SSA_1297
Putative GI 9	33%	0%	16%	50%	SSA_1312-SSA_1317
Putative GI 10	60%	80%	6%	66%	SSA_1326-SSA_1340
Putative GI 11	26%	34%	8%	78%	SSA_1378-SSA_1400
Putative GI 12	28%	21%	7%	21%	SSA_1435-SSA_1449

Putative GI 13	30%	40%	30%	40%	SSA_1472-SSA_1481
Putative GI 14	66%	66%	16%	66%	SSA_1594-SSA_1599
Putative GI 15 (islet)	0%	62%	12%	37%	SSA_1630-SSA_1638
Putative PAI 4	5%	7%	36%	52%	SSA_1643-SSA_1683
Putative PAI 5	36%	63%	36%	63%	SSA_1750-SSA_1760
Putative GI 16	46%	60%	13%	40%	SSA_1812-SSA_1821
Putative GI 17	50%	37%	25%	62%	SSA_1881-SSA_1891
Putative GI 18	0%	26%	26%	20%	SSA_1982-SSA_1994
Putative GI 19	29%	35%7%	17%	41%	SSA_2020-SSA_2032
Putative GI 20	20%	20%	26%	40%	SSA_2121-SSA_2135
Putative PAI 6	37%	25%	37%	50%	SSA_2147-SSA_2153
Putative PAI 7	66%	66%	50%	66%	SSA_2247-SSA_2252
Putative GI 21	42%	28%	14%	28%	SSA_2284-SSA_2290

4.3.2 Predicting PAIs by Comparative Genomics-Based Approach

Using the comparative genomics-based approach, 10 GIs were predicted, out of which, 4 were identified exploiting manual approach, while, 6 were found pre-computed in the database. Afterwards, all predicted GIs were checked manually for three most important genomic signatures i.e. hypothetical proteins, virulence genes and homolog of virulence genes in order to validate the findings. For the classification of GIs as PAIs, presence of virulence gene or its homology was exploited. A GI was considered as PAI if it contained virulence gene or its homolog and vice versa. Out of the total 10 predicted GIs, 3 GIs (GI2, GI7 and GI8) contain virulence genes and were identified as PAIs with strong prediction score whereas, 3 GIs (GI3, GI5 and GI10) were found to contain a homolog of virulence gene and were identified as PAIs with relatively weak prediction score. All these 6 GIs were considered as "False Negatives" with respect to PAIs as they are actually the PAIs that are negatively identified. Furthermore, three GIs (GI1, GI6 and GI8) were truly identified as GIs as they do not contain any virulence factor and thus were considered as "True negative" with respect to PAIs. Whereas, one GI (GI4) did not contain any hypothetical protein or virulence factor, thus, was not considered as PAI or GI at all, as presence of hypothetical proteins is one of the crucial markers of an LGT event. Therefore, GI4 was found to be "False Positive" prediction as this region of the genome and does not seem to have arisen in result of an LGT event. Summary of the GIs obtained by this approach is given in Table 4.6.

TABLE 4.6: Summary of GIs predicted in the genome of *Streptococcus sanguinis* SK36 by following comparative genomics approach

Putative GIs and PAIs	**Start**	**Stop**	**Size**	**Pred. Model**	**hyp.**	**vir. genes**	**Vir. genes homo.**
Putative GI1	160,317	165,657	5,340	Manual	*		
Putative GI2	1,116,985	1,122,526	5,541	Manual		*	

Putative GI3	1,171,147	1,179,963	8,816	Manual	*		*
Putative GI4	2,084,670	2,090,408	5,738	Manual			
Putative GI5	145,552	150,100	4,548	DB prediction	*		*
Putative GI6	1,037,302	1,041,509	4,207	DB prediction	*		
Putative GI7	1,115,016	1,120,426	5,410	DB prediction	*	*	
Putative GI8	1,315,090	1,320,601	5,511	DB prediction	*		
Putative GI9	2,298,588	2,305,638	7,050	DB prediction	*	*	
Putative GI10	2,305,761	2,312,880	7,119	DB prediction	*		*

4.4 Designing an Integrated Approach for PAIs Prediction

Importance of GI prediction cannot be denied as it serves as the primary step for bacterial genome characterization. Due to the growing interest of the researchers in the study of GI inferred characteristics, bioinformatics approaches for its prediction have been formulated at rapid pace. Development of new and efficient computational methods with improved prediction accuracies still remains to be the hot research area. Different existing tools uses different GI features for their identification. As a result, each of the most accurate methods have high precision but low recall, leading to variations in results predicted by different tools.

A literature survey shows different researches that have predicted GIs using different bioinformatics tools predicting GIs in different bacterial species. Some of these studies are summarized below:

In a study by Ali *et al*; (2012), PAIs have been identified in *Campylobacter fetus* subspecies by exploiting sequence composition-based approach using PIPS tool. The organisms included *C.fetus* subsp. *Venerealis NCTC10354T* (Cfv) and *C.fetus* subsp. 82-40 (Cff). As PIPS require a most closely related non-pathogenic organism in order to predict PAIs, *Campylobacter hominis* was chosen in the study as subject genome. Further validations were done manually by using tools like ACT and BLAST Ring Image Generator (BRIG). Results yielded 12 PAIs in Cfv and 10 in Cff [54].

Whereas, Soares *et al*; (2016), exploited sequence composition-based approach using GIPSY for the identification of PAIs in *Escherichia coli* CFT073. This study has used *E. coli* K12 as subject non-pathogenic genome and have predicted 23 putative PAIs and 9 GIs. To our knowledge, GIPSY is the only software that exploits maximum number of genomic signatures and is the most accurate sequence-composition based tool. In support to its highest accuracy, Soares *et al*; found 11 additional PAIs in *E. coli* CFT073 which were not reported previously in literature but were revealed when detected using GIPSY [20].

Zheng *et al*; (2017), conducted comparative genomics analysis on 14 strains of *Streptococcus gordonii* and 5 strains of *Streptococcus sanguinis* to predict GIs. This study found 13 putative GIs in *S. sanguinis* and *S. gordonii* collectively, out of which, 6 GIs were possessed uniquely by *S. gordonii* and 5 GIs were identified in *S. sanguinis*, whereas, two GIs were revealed to be conserved in both organisms. This study used 14 strains of *S. gordonii* (PV40, Blackburn, Channon, FSS2, FSS3, FSS8, M5, M99, MB666, MW10, PK488, SK12, SK120 and SK184) and 5 strains of *S. sanguinis* (NCTC 7863, FSS4, FSS9, MB451 and PJM8) [38].

Guo *et al*; (2017) uesd an integrated approach to identify GIs in *Burkholderia cenocepacia* AU 10. The adopted integrated methodology exploited four features of sequence composition-based approach along with comparative genomics approach. The features subset of sequence composition-based approach included frequency of dinucleotides, G+C content deviation, codon usage and presence of tRNA. Eight strains of *B. cenocepacia* used as comparative genome set included AU 1054, J2315, H2424, HI11, MC0-3, DDS 22E-1, DWS 37E-2 and K56-2. This study resulted in

identification of 14 putative GIs on chromosome I and 7 GIs on chromosome II of *B. cenocepacia* [55].

Klein *et al*; (2018) used G+C content as a basis to identify GI and PAIs in *Vibrio parahaemolyticus* TS-8-11-4, *Vibrio vulnificus* WR-2-BW and *Vibrio diabolicus* JBS-8-11-1. This study exploited TUBIC (Tiajin University Bioinformatics Center) to determine GIs and successfully identified 1 PAI in *V. parahaemolyticus* TS 8-11-4, 2 PAIs in *V. vulnificus* WR-2-BW and 1 GI in *Vibrio diabolicus* referred to as fitness island [51].

Filho *et al*; (2018) used the sequence based approach to identify GIs in *Escherichia coli* CFT073 using multiple sequence composition-based tools which included GIPSY, Alien Hunter, IslandViewer, PredictBias, and Zisland Explorer. This study identified 16 GIs in *E. coli* CFT07. Out of these 16 GIs, 8 were predicted to be GIs whereas, rest of 8 were identified as PAIs [27].

In the light of the above survey, it is observed that most of the GI studies utilize one of the two existing approaches i.e. either sequence composition or comparative genomics-based approach in order to predict GIs. This reflects that in-spite of the limitations with both the approaches mentioned in literature review (section 2.3), researchers continue to make use of any one of these approaches resulting in high rates of false results (low accuracy and precision) while identifying GIs [2]. There is also a possibility that these studies might not have depicted the complete picture of GIs in studied genomes due to the hindrance caused by limitations inherited by both approaches.

Very few studies included in the survey have used integrated approach proposing to minimize the challenges faced by the existing pipelines [44]. The integrated approach has been implemented by utilizing IslandViewer4 (http://www.pathogenomi cs.sfu.ca/islandviewer/) which is the only tool that incorporates both the approaches side by side. This integrated approach also faces certain chellanges that are discussed in detail in section 3.4.The integrated approach followed by IslandViewer4 (http://www.pathogenomics.sfu.ca/ islandviewer/) is depicted in Figure 4.5.

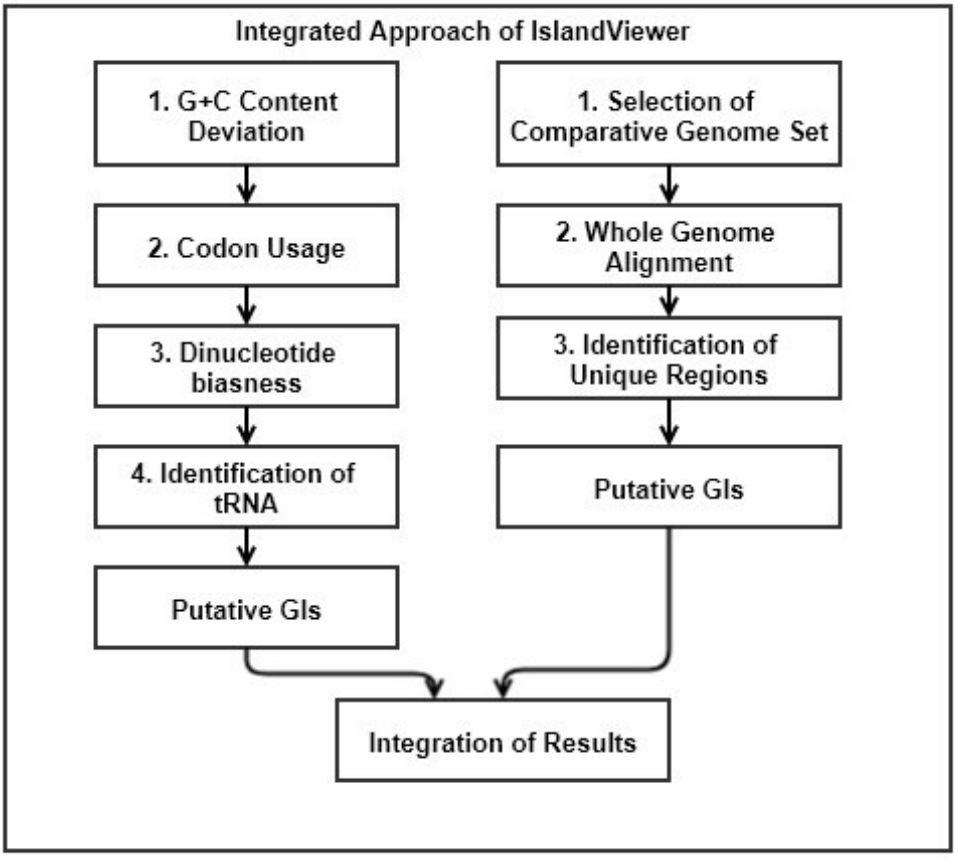

FIGURE 4.5: Integrated approach followed by IslandViewer4

Keeping in view the limitations of the existing approaches, this study has proposed an integrated approach that broadens up the genomic features subset. These features include presence of transposases, virulent genes, hypothetical proteins and unique regions identification in addition to G+C content deviations, codon usage, and presence of tRNA along with the utility of comparative genomics approach in parallel. The approach is then followed by the integration of results acquired by both approaches, preceded by manual validation to classify GIs as PAIs. Proposed integrated approach is shown in Figure 4.6.

With such integrated approach, more stringent criteria is applied for a region to be classified as GI that will minimize probability of False positive and negative results. With this approach, for a region to be qualified as a GI, it must possess 3-4 mentioned genomic signatures as it is not possible most of the time to possess all the genomic features [16].

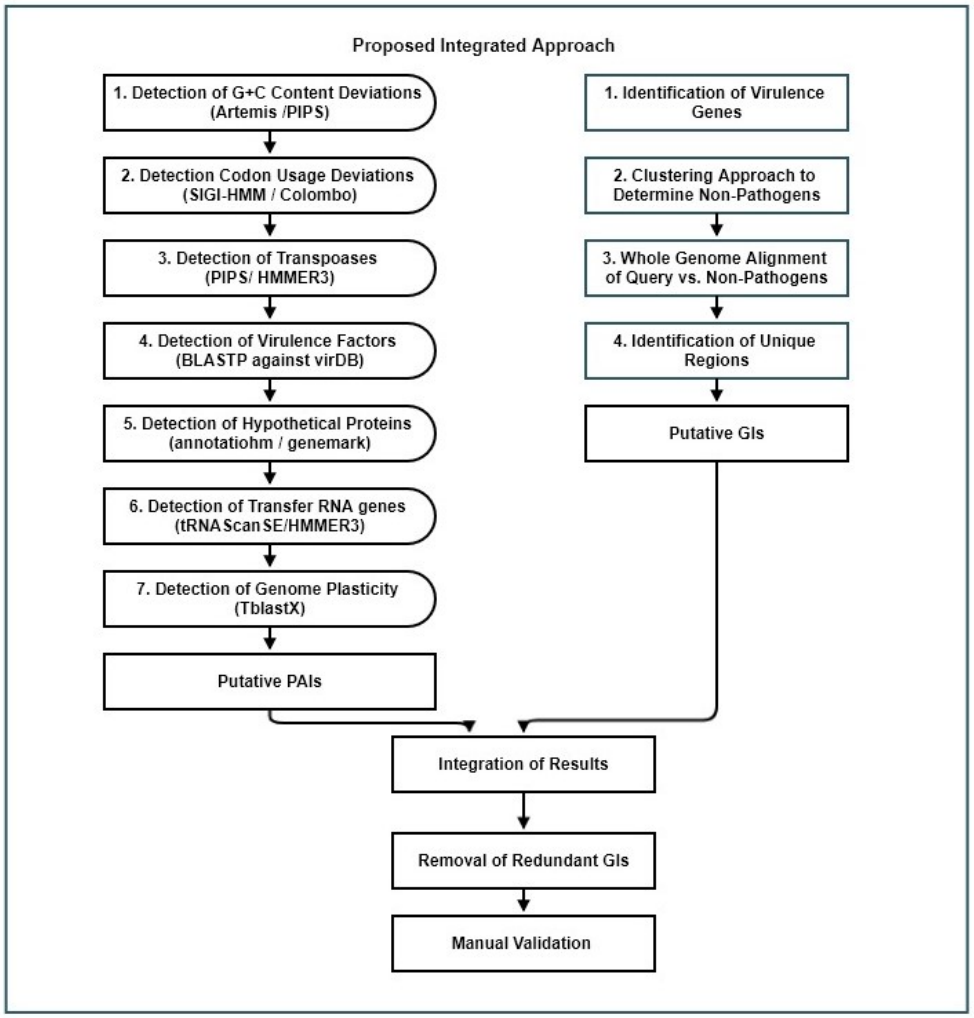

FIGURE 4.6: Integrated approach proposed by this study

4.5 Application of Integrated Approach to Case of *Streptococcus sanguinis* SK36

For the application of sequence composition as well as comparative genomics-based approach, importance of selection of reference genome has already been discussed in section 3.3. For a specie like *S. sanguinis* that has unknown basis of pathogenecity,determination of non-pathogenic strain is proposed by k-means clustering method preceded to determination of virulence factors.

4.5.1 Virulence Factors of *Streptococcus sanguinis* SK36

45 virulent factors associated with *S. sanguinis* SK36 were retrieved from VFDB (summerized in Table 4.7). These virulence factors majorly belong to the group of adherence proteins, enzymes, immune evasion, manganese uptake and proteases.

Adherence related factors include antigens, cell surface hydrophobicity proteins, collagen binding proteins, laminin binding proteins ,fibronectin binding protein, serine-rich surface glycoprotein, sortase A, glycosyltransferases, plasmin receptor and rslrA islets. Among these factors, cell surface proteins mediates attachment to the host cell and is required for initial stages of the IE [43] whereas, fibronectin binding proteins aids in binding to fibronectin which exist in the extracellular matrix of most tissues in hosts [50]. Collagen binding proteins are identified as mediators of IE and includes CpbA of *S. sanguinis*. CbpA is recognized as a mediator of platelet aggregation in vitro [39], [43]. Laminin-binding proteins aids in attachment to human-laminin and is significant for bacterial colonization [53]. Serine- rich glycoproteins mediates the adhesion to platelets [ref2] whereas, sortase A aids bacterium in invading human immune system [55]. Extracellular production of glucan polymers has also been linked to IE infectivity. The extracellular glucans are synthesized by bacterially encoded glucosyltransferase (Gtf) enzymes from a sucrose substrate and enhance colonization in the development of IE. Studies suggest that glucan production enhance streptococcal survival post-phagocytosis putatively mediating adherence to vegetation-like matrices and includes gtfD [34], [39]. Plasmin receptors also known as Streptococcal surface dehydrogenase are marked essential for evasion from neutrophils [56]. rlrA islets are the transcriptional regulators which mediates proper temporal expression of virulence genes during infection [57]. Enzyme related virulence factor includes *Streptococcal* enolase. Enolase is found abundantly on the surface of *streptococcal* groups and shows great affinity to bind plasmin that plays crucial role in host defense system [58]. In addition to that, 25 capsule genes were also identified as virulence genes in *S.sanguinis*. These genes provides resistance to complement deposition and masks cell wall-associated complement from being recognized by the complement receptors on phagocytes [59]. In addition to these, manganese uptake genes like solC also

play their role in virulence casuing endocarditis. sloC is required for the expression of Lral operon which is responsible for encoding components of ATP-binding cassettes [60]. Proteases like C3-degradation protease, igA1 protease and trigger factor also constitutes the set of virulence factors in SK36.

TABLE 4.7: IE virulence genes catalog along with the patho-comparison within five *sanguinis* strains

S. No	**VF Class**	**Locus tag**	**NCTC 11085**	**NCTC 11086**	**NCTC 7863**	**NCTC 10904**	**NCTC 3168**
	Adherence						
1	Antigen I/II	SSA_0956 SSA_0303	* *	* *	* *	* *	
2	Cell surface hydrophobicity proteins	SSA_0904 SSA_0905 SSA_0906	* * *	* * *	* * *	* * *	
3	Collagen binding proteins	SSA_1663	*	*	*	*	
4	Fibronectin binding proteins	SSA_0907	*	*	*	*	
5	Laminin-binding protein	SSA_1990	*	*	*	*	
6	Serine-rich surface glycoproteins	SSA_0829	*	*	*	*	
7	Sortase A	SSA_1219	*	*	*	*	
8	Streptococcal glucosyl-transferases	SSA_0613	*	*	*	*	
9	Streptococcal plasmin receptor	SSA_2108	*	*	*	*	

10	rlrA islet	SSA_1632 SSA_1633 SSA_1634 SSA_1631	* * * *				
	Enzymes						
11	Streptococcal enolase	SSA_0886	*	*	*	*	
	Immune evasion						
12	Capsule	SSA_1511 SSA_0858 SSA_1411 SSA_2217 SSA_1510 SSA_1519	* * * * * *	* * * * * *	* * * * *	* * * * *	*
		SSA_2223 SSA_1410 SSA_1518 SSA_1517	* * * *	* *	* *	* * * *	
	Manganese uptake						
13	SSA_0206	ssaB	*	*		*	
	Protease						
14	C3-degrading protease	SSA_0331	*	*	*	*	
15	IgA1 protease	SSA_1106	*	*	*	*	
16	Trigger factor	SSA_1998	*	*	*	*	*

4.5.2 Clustering of *Streptococcus sanguinis* Strains

Pre-enlisted virulent factors were scanned in the genome of the strains for the first round of clustering. Results indicated the presence of most of the virulent determinants in 4 out of 5 *sanguinis* strains and absence of majority of the factors from 1 strain. Strains that possess virulent genes include NCTC11085, NCTC11086, NCTC7863, and NCTC10904, whereas, NCTC3168 was found to possess only 2 virulent factors. Therefore, based on the acquired results 4 strains were clustered as pathogenic and 1 as non-pathogenic.

Pathogenic cluster was further divided into the sub-clusters of severe and moderate pathogens based on the distance of these strains from the centroid strain i.e. SK36. 2 strains of the pathogenic cluster showed distance of 5 with the centroid and hence clustered as severe pathogens. These strains include NCTC11085 and NCTC10904 and possess 40 out of total 45 virulent genes of SK36 which makes them hyper-virulent. Whereas, distance of the rest of the 2 pathogenic strains in reference to the centroid was ¿8 which indicates absence of atleast 10 virulence determinant that makes these strains moderate pathogens.

Results revealed absence of rlrA islets in most of the strains. These islets include surface protein, FimA, heme-utilization adhesion exo-proteins and sortase C which are involved in the initial binding to the blood vessels. These genes were found missing in NCTC11086, NCTC7863, NCTC10904 and NCTC3168 strains. Thus, these strains have reduced ability to cause IE as studies have indicated strains with mutant FimA reduces the ability of *S.sanguinis* to cause IE [44].

On the second highest frequency, strains lacked Cps9H, zn-porter lipoprotein, teichoic acid transporter, rgpE, capD, gtfP, Cps9G and Cpsla. These genes were found in any 2 of the 5 strains. These genes majorly involves glucosyltransferases or the transporter proteins.

Rest of the genes were present in all four strains i.e. NCTC11085, NCTC11086, NCTC7863 and NCTC10904. However, tig and glucosyltransferase SSA_1511 were the only virulent factors found to be possessed by NCTC3168.

As a result of clustering approach, NCTC11085 and NCTC10904 were clustered as "severe pathogens", NCTC11086 and NCTC7863 were clustered as "moderate

pathogens" whereas, NCTC3168 was determined as "non-pathogen". Resulted clustering is shown in Figure 4.7.

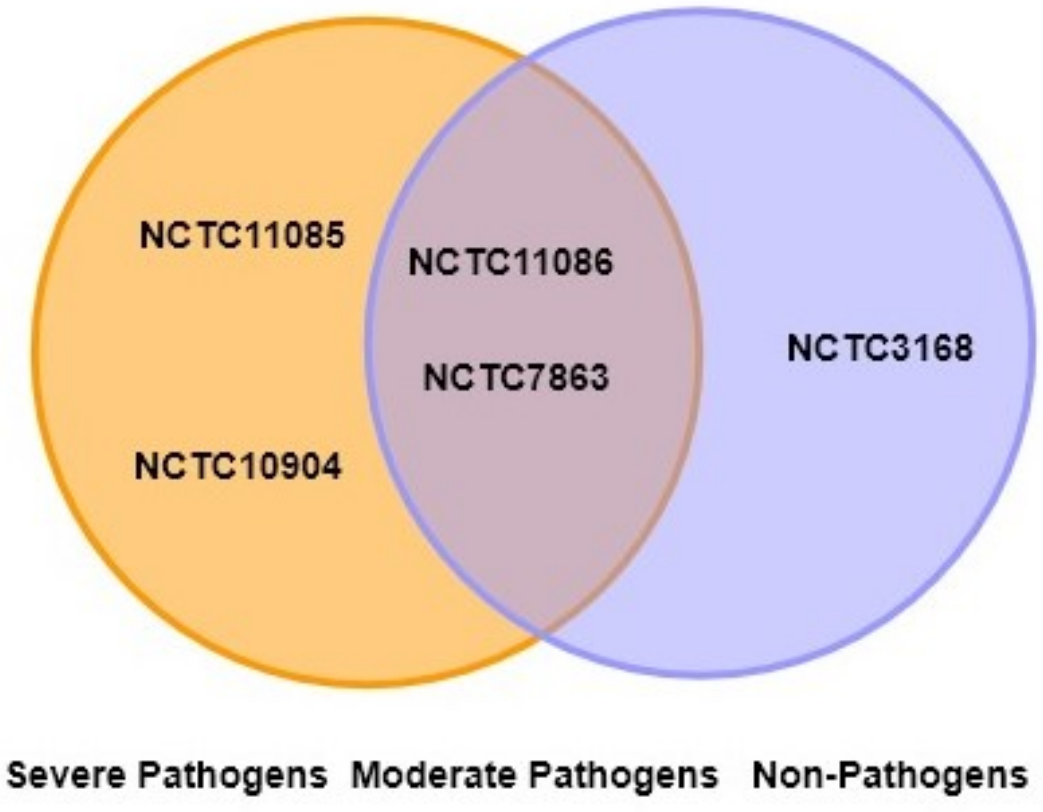

FIGURE 4.7: Classification of *S. sanguinis* strains based on their pathogenic association with IE. Pathogenic cluster includes strains that possess IE virulent genes whereas non-pathogenic strains do not.

4.5.3 Validation of Clustering and Selection of Reference Genome

With the use of k-mean clustering approach, out of total 5 strains of *S. sanguinis*, only NCTC3168 was determined as non-pathogenic. To validate the results of clustering approach, all 5 strains were compared with the query strain SK36 using BLAST pairwise alignment. The results of pairwise alignment shows NCTC3168 as to be 90% similar with the query strain, whereas rest of the four strains are yielded to be more than 90%. Distance tree of the results in Figure 4.8 show that NCTC7863 and NCTC11085 are clustered close together based on their nucleotide

homology and because of their greater similarity with the genome. This cluster is then close together with NCTC10904 and NCTC11086. On the whole all these four strains are closer in distance to the query strain SK36 which supports the clustering results that have clustered all these four strains as pathogenic. Whereas, NCTC3168 is shown to be at maximum distance from the query strain, reflecting its least nucleotide homology with SK36 and hence supports the selection of this strain as reasonable subject/reference strain for the prediction of GI based on sequence composition-based approach.

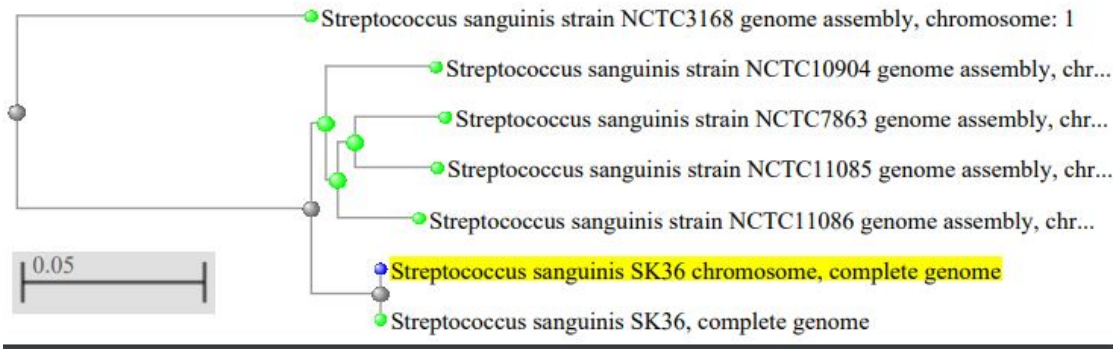

FIGURE 4.8: Distance tree of selected *sanguinis* strains

When the nucleotide sequences of all 5 strains were compared with reference to SK36 using BLAST, NCTC10904 show 94% similarity with SK36, NCTC7863, NCTC11085 and NCTC11086 show 96% homology, whereas, NCTC3168 show 90% homology which again validates the clustering results and our selection of reference genome. In Figure 4.9 genome map comparison between the chosen strains of SK36 is shown representing gene organization with genome synteny breaks referred to as GIs. In the light of the above results and validation support, NCTC3168 was chosen as the subject genome for the identification of GIs exploiting sequence composition approach. This finding could also be used in another way, depending upon the objectives of the study. Such non-pathogenic cluster could be used as a comparative genome set as a whole to predict GIs in the query genome. Using such un-ambiguous comparative genome set will obviously minimize the false results and will be time efficient as it will prevent brute force efforts.

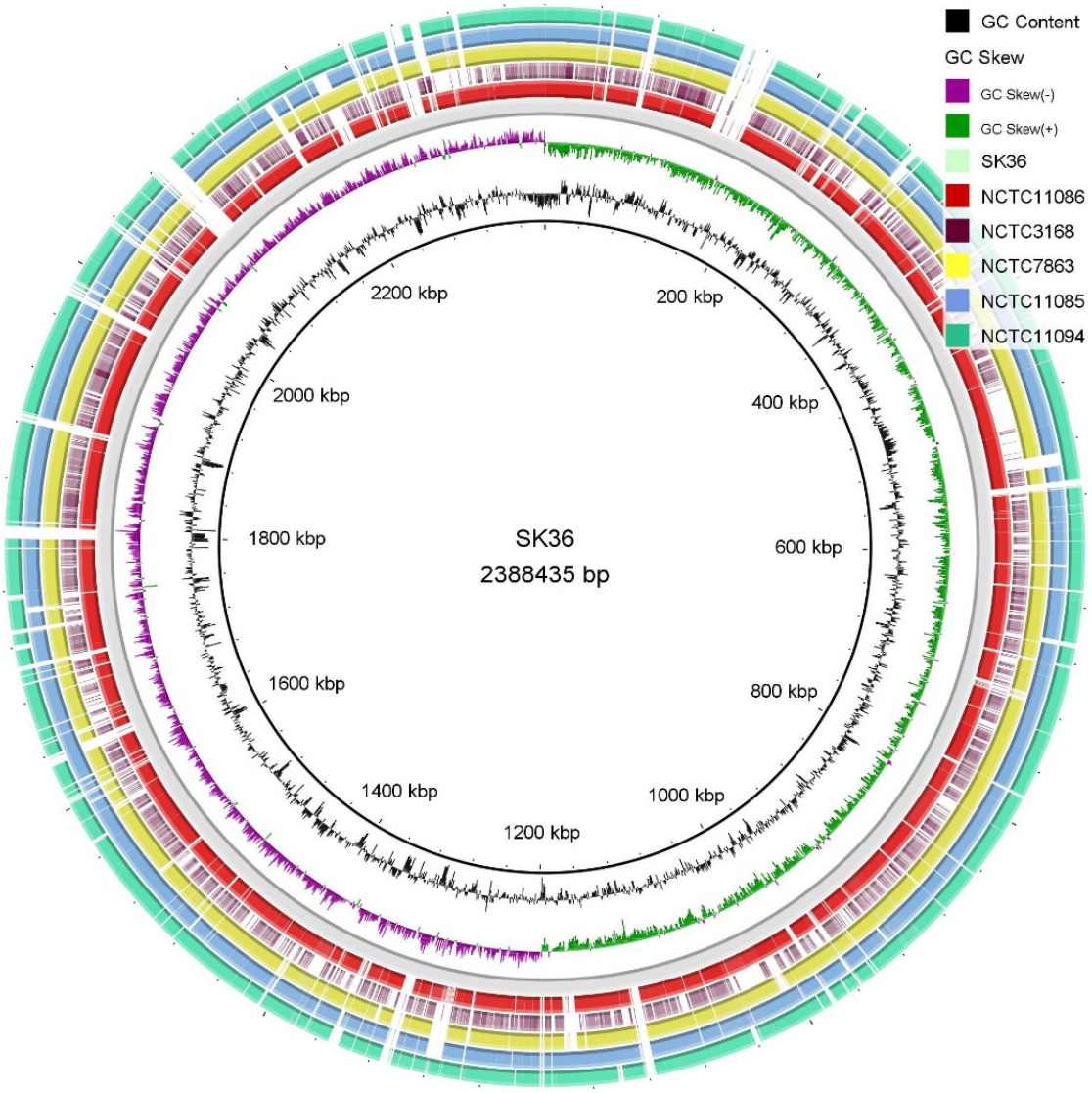

FIGURE 4.9: Genome map comparison of *S. sanguinis* strains *S. sanguinis* strains plotted using SK36 as reference, representing GIs in SK36

4.5.4 Integration of Results

For the acquisition of final GIs/PAIs, results obtained from both the approaches were integrated. For the results from sequence composition-based approach, as NCTC3168 was chosen as a reference genome, GIs/PAIs from Table 4.5 were taken into account whereas, Table 4.3 and Table 4.4 were dropped. It is also to be noted here that PAIs identified by keeping NCTC3168 has strong prediction scores as compared to the PAIs predicted by using other two strains. This observation again validates the selection of NCTC3168 as reference genome. Along with these, results of comparative genomics-based approach from Table

4.6 were selected for integration purpose and a new dataset was constructed. The newly constructed dataset now reflects integrated results. Examining this dataset, surprisingly reveals, no GI has remained conserved while being predicted by both approaches, which supports the use of integrated approach. This variation of results from both approaches suggests that use of any one of the approaches for GI/PAIs prediction is not valid in every case.

4.6 Prediction's Accuracy Evaluation of Under Study Approaches

For evaluating the accuracy of sequence composition, comparative genomics as well as the proposed integrated approach, datasets were created by determining whether the GIs and PAIs predicted by the existing approaches represents true negative, false negative, true positive or false positive results.

Approach adopted to determine negative or positive result is discussed in methodology section 3.3.4. When sequence composition-based pipeline was applied to the case organism SK36 with reference to all three randomly selected subject genomes 32 GIs and 10 PAIs were predicted in total. Out of these results, 8 PAIs turns out to be true positive, as these PAIs have truly been predicted as PAIs because of the presence of either the virulent gene or its homolog. Whereas, only 2 PAIs turns out to be false positively predicted, as those PAIs were actually GIs but have been predicted as PAIs when they did not contain any virulence gene. On the other hand,14 GIs were predicted accurately and turns out to be true negative in the dataset whereas, 18 GIs were actually the PAIs but have been predicted as GIs thus turns out to be false negative entries in Table 4.8. With this dataset, accuracy for the sequence composition-based approach turns out to be 52%. Whereas, after the selection of NCTC3168 as most closely related non-pathogenic subject genome, accuracy of sequence composition-based approach increases to 53.5%.

TABLE 4.8: Positive and Negative dataset construction of predicted GIs and PAIs for sequence composition-based approach

True Positive	**True Negative**	**False Negative**	**False Positive**
PAI1	GI1	GI5	PAI10
PAI2	GI2	GI6	PAI3
PAI4	GI3	GI7	
PAI5	GI4	GI8	
PAI6	GI10	GI9	
PAI7	GI14	GI11	
PAI8	GI16	GI12	
PAI9	GI21	GI13	
	GI22	GI15	
	GI25	GI17	
	GI26	GI18	
	GI27	GI19	
	GI28	GI20	
	GI32	GI23	
		GI24	
		GI29	
		GI30	
		GI31	

Accuracy of sequence composition-based approach = 22/42 = 0.52 =52%

While during the application of comparative genomics approach, out of total 10 predicted GIs, 4 turns out to be true positive as these GIs were truly predicted whereas, rest of the 6 GIs were actually the PAIs but were predicted as GIs, thus counted as false positive entries of Table 4.9. Whereas, no true negative entries could have been generated. Accuracy, when computed for the comparative genomics-based approach turned out to be 40%.

TABLE 4.9: Positive and negative dataset construction of predicted GIs for comparative genomics-based approach

True Negative	**False Negative**
GI33	GI34
GI36	GI35
GI38	GI37
GI40	GI39
	GI41
	GI42

Accuracy of comparative genomics-based approach = 4+0/10= 0.4 = 40%

With the proposed integrated approach, dataset (shown in Table 4.10) is constructed integrating the resultant GIs from Table 4.5 and Table 4.6. Among this dataset, out of total 31 GIs and 7 PAIs, 6 PAIs and 5 GIs were truly predicted, hence true positive entries, 7 GIs were truly identified as not being PAIs and therefore, true negative entries whereas, 12 GIs were falsely predicted as GIs when they were PAIs in actual, hence false positive results. Accuracy of integrated approach when calculated, turns out to be 47% which is better then the accuracy of comparative genomics based-approach but less then that of sequence composition-based approach. It is to be noted here, that all of the predicted GI/PAIs are the putative ones and are not confirmed. Based on the presence of virulence factors (mentioned in Table 4.7) in GIs/PAIs, they were again evaluated to be either true positive or true negative islands. This re-evaluation revealed 19 GIs to be true negative as they do not contain any of the virulence factor (from Table 4.7), 4 PAIs (PAI4, PAI6,PAI7,PAI9) were found to be true positive because of the presence of the respective virulent determinants (SSA_0206 in PAI4, SSA_0956 in PAI6, SSA_1663 in PAI7 and SSA_2217 in PAI9), 1 GI (GI26) was found false negative as it contained 4 virulent genes (SSA_1631,SSA_1632, SSA_1633, SSA_1634 constituting rlrA islet), whereas, 2 PAIs (PAI8 and PAI10) were found to be false positive because of the absence of any known virulence factor. Re-evaluating the accuracy based on

confirmed and putative GIs/PAIs, it increases to 71% which is comparatively better than the accuracy of both the approaches. It concludes that designed integrated approach has the tendency to identify confirmed GIs and PAIs with 71% accuracy even for the pathogens that do not have complete information of pathogenesis.

TABLE 4.10: Dataset constructed for proposed integrated approach

True Negative	**False Negative**	**True Positive**
GI14	GI9	PAI4
GI16	GI11	PAI5
GI21	GI15	PAI6
GI22	GI17	PAI7
GI25	GI18	PAI8
GI26	GI19	PAI9
GI27	GI20	
GI28	GI23	
GI32	GI24	
GI33	GI29	
GI38	GI30	
GI40	GI31	
	GI34	
	GI35	
	GI37	
	GI39	
	GI41	
	GI42	

Accuracy of proposed integrated approach = 18/38= 0.47=47%

Chapter 5

Conclusion and Future Recommendations

Existing approaches for the prediction of GIs are grouped into two major categories i.e. sequence composition-based approach and comparative genomics-based approach. Both of these approaches inherit certain limitations which effects the accuracy of predicting GIs. Keeping in view the challenges faced by these approaches, this study has proposed an integrated approach which could minimize the short comings of the existing approaches. The proposed integrated approach suggests certain modifications in both of the existing approaches and then integrating the results. In sequence composition-based approach, broadening up of genomic set of features is proposed with the aim to increase the stringency on decision making, minimizing the probability of false predictions. In comparative genomics approach, based on the fact that for an organism with related strains of unknown pathogenicity, k-means clustering approach is suggested to cluster pathogenic and non-pathogenic strains. This clustering is based on the identified virulence genes found in organism. The selection of comparative genome set or the subject genome based on such clustering method could provide more accurate selection of comparative genomes rather than assumption or brute force selections as this approach is greatly influenced by the selected comparison genome set. The generated non-pathogenic cluster could also be used as a comparative genome set as a whole to predict GIs in the query genome. Using such un-ambiguous comparative genome set will

obviously minimize the false results and will be time efficient as it will prevent brute force efforts. Such novel integrated approach would help to highly minimize the challenges faced by the existing GI prediction approaches, minimizing the false prediction results and solves the problem of predicting GIs/PAIs for organisms that do not have a complete picture of pathogenicity.

Bibliography

[1] S. De, C. Soares, L. De Castro Oliveira, A. Kumar Jaiswal, V. Azevedo, and S. J. Gould, "Genomic Islands: An overview of current software tools and future improvements- Living in the 'Age of Bacteria,' " J. Integr. Bioinform., vol. 13, no. 1, pp. 82-89, 2016.

[2] B. Lu and H. W. Leong, "Computational methods for predicting genomic islands in microbial genomes," CSBJ., vol. 14, pp. 200–206, 2016.

[3] O'CONNELL, MP" Genetic Transfer in Prokaryotes: Transformation, Transduction, and Conjugation", Advanced molecular genetics, vol. 2, pp. 1-6,2012.

[4] M. Gavel and I. Langille, "Computational Prediction and Characterization of Genomic Islands- Insights into Bacterial Pathogenicity by B . Sc . University of New Brunswick , 2004 B . Cs University Of New Brunswick , 2004 Thesis Submitted In Partial Fulfillment of The Requirements" F,2009.

[5] B. Foxman, "Molecular Tools and Infectious Disease Epidemiology," Mol. Tools Infect. Dis. Epidemiol.,ISBN 978-0-12-374133-2, 2012.

[6] M. Munoz-Lopez and J. Garcia-Perez, "DNA Transposons: Nature and Applications in Genomics," Curr. Genomics., vol. 11, no. 2, pp. 115–128, 2010.

[7] M. Vizváryová and D. Valková, "Transposons – the useful genetic tools General features of transposons," Biologia., vol. 593, no. 1981, pp. 309–318, 2004.

[8] M. R. Gillings, "Integrons: Past, Present, and Future," Microbiol. Mol. Biol. Rev., Microbiology and Molecular Biology Reviews., vol. 78, no. 2, pp. 257–277, 2014.

[9] D. Mazel, "Integrons: Agents of bacterial evolution," Nat. Rev. Microbiol., vol. 4, no. 8, pp. 608–620, 2006.

[10] J. Mahillon and M. Chandler, "Insertion sequences," Microbioliology and Molecular Biology Review., vol. 62, no. 3, p. 725–74., 1998.

[11] P. Siguier, E. Gourbeyre, and M. Chandler, "Bacterial insertion sequences: Their genomic impact and diversity," FEMS Microbiol. Rev., vol. 38, no. 5, pp. 865–891, 2014.

[12] M. G. I. Langille, W. W. L. Hsiao, and F. S. L. Brinkman, "Evaluation of genomic island predictors using a comparative genomics approach," BMC Bioinformatics., vol. 10, no.1, pp. 1–10, 2008.

[13] D. Che, H. Wang, J. Fazekas, and B. Chen, "An Accurate Genomic Island Prediction Method for Sequenced Bacterial and Archaeal Genomes," Proteomics & Bioinformatics., vol. 7, no. 8, pp. 214–221, 2014.

[14] U. Dobrindt, B. Hochhut, U. Hentschel, and J. Hacker, "Genomic islands in pathogenic and environmental microorganisms," Nat. Rev. Microbiol., vol. 2, no. 5, pp. 414–424, 2004.

[15] M. Juhas, J. R. Van Der Meer, M. Gaillard, R. M. Harding, D. W. Hood, and D. W. Crook, "Genomic islands: Tools of bacterial horizontal gene transfer and evolution," FEMS Microbiol. Rev., vol. 33, no. 2, pp. 376–393, 2009.

[16] M. G. I. Langille, W. W. L. Hsiao, and F. S. L. Brinkman, "Detecting genomic islands using bioinformatics approaches," Nat. Rev. Microbiol., vol. 8, no. 5, pp. 373–382, 2010.

[17] O. Gal-mor and B. B. Finlay, "Microreview Pathogenicity islands- a molecular toolbox for bacterial,"Cellular Microbiology., Vol. 8, no. 11, pp.1707–1719,2006.

[18] D. Che, M. S. Hasan, and B. Chen, "Identifying Pathogenicity Islands in Bacterial Pathogenomics Using Computational Approaches," Pathogens., vol. 3, no. 1, pp. 36–56, 2014.

[19] J. B. Kaper, "Pathogenicity Islands and Evolution of Microbes," Annu. Rev. Microbiol., Vol. 54, no.1, pp.641-679, 2000.

[20] S. C. Soares et al., "GIPSy: Genomic island prediction software," J. Biotechnol., vol. 232, pp. 2–11, 2016.

[21] M. S. Hasan, Q. Liu, H. Wang, J. Fazekas, B. Chen, and D. Che, "GIST: Genomic island suite of tools for predicting genomic islands," Bioinformation., vol. 8, no. 4, pp. 203–205, 2012.

[22] V. A. C. Abreu et al., "PIPS: Pathogenicity Island Prediction Software," PLoS One., vol. 7, no. 2, pp. e30848, 2012.

[23] C. Bertelli, K. E. Tilley, and F. S. L. Brinkman, "Microbial genomic island discovery , visualization and analysis," Briefings in Bioinformatics., Vol.42 , pp. 1–14, 2018.

[24] S. Podell and T. Gaasterland, "DarkHorse: A method for genome-wide prediction of horizontal gene transfer," Genome Biol., vol. 8, no. 2, 2007.

[25] D. Che and H. Wang, "GIV: A Tool for Genomic Islands Visualization," Bioinformation., vol. 9, no. 17, pp. 879–882, 2013.

[26] C. Bertelli et al., "IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets," Nucleic Acids Res., vol. 45, no. W1, pp. W30–W35, 2017.

[27] A. C. da Silva Filho et al., "Comparative Analysis of Genomic Island Prediction Tools," Front. Genet., vol. 9, no. December, pp. 1–15, 2018.

[28] S. Paik, S. Das, J. C. Noe, and T. Kitten, "Identification of Virulence Determinants for Endocarditis in Streptococcus sanguinis by Signature-Tagged Mutagenesis," Infection and Immunity., vol. 73, no. 9, pp. 6064–6074, 2005.

[29] X. Ge, T. Kitten, Z. Chen, S. P. Lee, C. L. Munro, and P. Xu, "Identification of Streptococcus sanguinis genes required for biofilm formation and examination of their role in endocarditis virulence," Infect. Immun., vol. 76, no. 6, pp. 2551–2559, 2008.

[30] K. Lix, L. Tronstad, and Olsen I, "Systemic Diseases Caused By Oral Infection," Clin. Microbiol. Rev., vol. 13, no. 4, pp. 547–58, 2000.

[31] C. H. Cabell, E. Abrutyn, and A. W. Karchmer, "Bacterial endocarditis: the disease, treatment, and prevention," Circulation., vol. 107, no. 20, pp. e185-e187, 2003.

[32] D.Tilley and S.W. Kerrigan, "Platelet-Bacterial Interactions in the Pathogenesis of Infective Endocarditis Part-I The Streptococcus," Recent Advances in Infective Endocarditis, Recent Advances in Infective Endocarditis., ISBN: 978-953-51-1169-6, 2013.

[33] T. L. Holland, L. M.Baddour, A.S. Bayer, B. Hoen, J.M.Miro, and V.G. Fowler Jr,"Infective Endocarditis," Nat Rev Dis Primers., Vol. 2, pp.263-296, 2017.

[34] C. T. Shun, S. Y. Lu, C. Y. Yeh, C. P. Chiang, J. S. Chia, and J. Y. Chen, "Glucosyltransferases of viridans streptococci are modulins of interleukin-6 induction in infective endocarditis," Infect. Immun., vol. 73, no. 6, pp. 3261–3270, 2005.

[35] J. R. McDonald, "Acute infective endocarditis.," Infect. Dis. Clin. North Am., vol. 23, no. 3, pp. 643–664, Sep. 2009.

[36] M. Tariq, M. Alam, G. Munir, M. A. Khan, and R. A. Smego, "Infective endocarditis: A five-year experience at a tertiary care hospital in Pakistan," Int. J. Infect. Dis., vol. 8, no. 3, pp. 163–170, 2004.

[37] T. Kitten, C. L. Munro, N. Q. Zollar, S. P. Lee, and R. D. Patel, "Oral streptococcal bacteremia in hospitalized patients: Taxonomic identification and clinical characterization," J. Clin. Microbiol., vol. 50, no. 3, pp. 1039–1042, 2012.

[38] W. Zheng, M. F. Tan, L. A. Old, I. C. Paterson, N. S. Jakubovics, and S. W. Choo, "Distinct Biological Potential of Streptococcus gordonii and Streptococcus sanguinis Revealed by Comparative Genome Analysis," Sci. Rep., vol. 7, no. 1, pp. 29-49, 2017.

[39] S. Paik, L. Senty, S. Das, J. C. Noe, and C. L. Munro, "Identification of Virulence Determinants for Endocarditis in Streptococcus sanguinis by Signature-Tagged Mutagenesis," Infection and Immunity., vol. 73, no. 9, pp. 6064–6074, 2005.

[40] S. Y. Kim, S. Joo, J. Yi, and E. Kim, "A Case of Streptococcus gallolyticus subsp . gallolyticus Infective Endocarditis with Colon Cancer, Identification by 16S Ribosomal DNA Sequencing," Korean J Lab Med., Vol. 30,no.2, pp. 160–165, 2010.

[41] A. Yeşilkaya, Ö. K. Azap, B. Pirat, B. Gültekin, and H. Arslan, "A rare cause of endocarditis: Streptococcus pyogenes," Balkan Med. J., vol. 29, no. 3, pp. 331–333, 2012.

[42] P. Xu and X. Ge, "Essential Genes Identification in Streptococcus Sanguinis and Comparison among Streptococci," Iconceptpress.Com., Vol., 2018.

[43] L. Turner, "Identification of Virulence Determinants for Streptococcus sanguinis Infective Endocarditis," Virginia Commonwealth University., Vol.3, 2008.

[44] D. Burnette-Curley et al., "FimA, a major virulence factor associated with Streptococcus parasanguinis endocarditis," Infect. Immun., vol. 63, no. 12, pp. 4669–4674, 1995.

[45] T. Tatusova, S. Ciufo, B. Fedorov, K. O'Neill, and I. Tolstoy, "RefSeq microbial genomes database: new representation and annotation strategy," Nucleic Acids Res., vol. 43, no. 7, pp. 3872, 2015.

[46] N. Krislock and H. Wolkowicz, "Euclidean Distance Matrices and Applications," Handb. Semidefinite, Conic Polynomial Optim., vol. 5971, no. 2013, pp. 879–914, 2012.

[47] K. Nakano, R. Nomura, and T. Ooshima, "Streptococcus mutans and cardiovascular diseases," JDSR., vol. 44, no. 1, pp. 29–37, 2008.

[48] Bernard, H Russell and Killworth, Peter and Kronenfeld, David and Sailer, Lee, "The problem of informant accuracy: The validity of retrospective data", vol. 13, no. 1, pp. 495-517, 1984.

[49] D. Vallenet et al., "MicroScope: a platform for microbial genome annotation and comparative genomics," Database., Vol. 2009, 2009.

[50] A. E. Darling, B. Mau, and N. T. Perna, "Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement," PLoS One., vol. 5, no. 6, 2010.

[51] S. Klein, S. Pipes, and C. R. Lovell, "Occurrence and significance of pathogenicity and fitness islands in environmental vibrios," AMB Express., vol. 8, no. 1, pp. 1-77, 2018.

[52] D. Oliveira Alvarenga, L. M. Moreira, M. Chandler, and A. M. Varani, "A practical guide for comparative genomics of mobile genetic elements in prokaryotic genomes," Methods Mol. Biol., vol. 1704, pp. 213–242, 2018.

[53] P. Xu et al., "Genome of the opportunistic pathogen Streptococcus sanguinis," J. Bacteriol., vol. 189, no. 8, pp. 3166–3175, 2007.

[54] A. Ali et al., "Campylobacter fetus subspecies: Comparative genomics and prediction of potential virulence targets," Gene., vol. 508, no. 2, pp. 145–156, 2012.

[55] F. Guo, L. Xiong, K. Zhang, C. Dong, F. Zhang, and P. C. Y. Woo, "Identification and analysis of genomic islands in Burkholderia cenocepacia AU 1054 with emphasis on pathogenicity islands," BMC Microbiology., Vol. 17, pp. 1–10, 2017.

[56] B. Spellerberg et al., "Lmb, a protein with similarities to the LraI adhesin family, mediates attachment of Streptococcus agalactiae to human laminin," Infect. Immun., vol. 67, no. 2, pp. 871–878, 1999.

[57] C. Plummer, H. Wu, S. W. Kerrigan, G. Meade, D. Cox, and C. W. I. Douglas, "A serine-rich glycoprotein of Streptococcus sanguis mediates adhesion to platelets via GPIb," Br. J. Haematol., vol. 129, no. 1, pp. 101–109, 2005.

[58] M. Yamaguchi, Y. Terao, T. Ogawa, T. Takahashi, S. Hamada, and S. Kawabata, "Role of Streptococcus sanguinis sortase A in bacterial colonization," Microbes Infect., vol. 8, no. 12–13, pp. 2791–2796, 2006.

[59] Y. Terao, M. Yamaguchi, S. Hamada, and S. Kawabata, "Multifunctional glyceraldehyde-3-phosphate dehydrogenase of Streptococcus pyogenes is essential for evasion from neutrophils," J. Biol. Chem., vol. 281, no. 20, pp. 14215–14223, 2006.

[60] D. L. Hava, C. J. Hemsley, and A. Camilli, "Transcriptional regulation in the Streptococcus pneumoniae rlrA pathogenicity islet by RlrA," J. Bacteriol., vol. 185, no. 2, pp. 413–421, 2003.

[61] V. Pancholi and V. a Fischetti, "Protein Chemistry And Structure alpha Enolase, a Novel Strong Plasmin (ogen) Binding Protein on the Surface of Pathogenic Streptococci Enolase, a Novel Strong Plasmin (ogen) Binding Protein on the Surface of Pathogenic Streptococci," Journal of Biological Chemistry., vol. 273, no. 23, pp. 14503–14515, 1998.

[62] J. O. Kim and J. N. Weiser, "Association of Intrastrain Phase Variation in Quantity of Capsular Polysaccharide and Teichoic Acid with the Virulence of Streptococcus pneumoniae ," J. Infect. Dis., vol. 177, no. 2, pp. 368–377, 2009.

[63] T. Kitten, C. L. Munro, S. M. Michalek, and F. L. Macrina, "Genetic characterization of a Streptococcus mutans LraI family operon and role in virulence," Infect. Immun., vol. 68, no. 8, pp. 4441–4451, 2000.