Spring 2022

Functional Facets of Intergenic Hairpin Structures in Genus Caulobacter

Geetha Saarunya Clarke

Functional facets of intergenic hairpin structures in genus Caulobacter

by

Geetha Saarunya Clarke

Bachelor of Technology
SRM University, 2008

Master of Sciences
University of South Carolina, 2010

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Biological Sciences

College of Arts and Sciences

University of South Carolina

2022

Accepted by:

Bert Ely, Major Professor

Jeff Dudycha, Committee Member

Jason Stewart, Committee Member

Jijun Tang, Committee Member

Joe Quattro, Committee Member

Tracey L. Weldon, Vice Provost and Dean of the Graduate School

ABSTRACT

DNA replication, recombination and repairs maintain bacterial genome stability. But these processes may also induce genome rearrangements leading to inter and intra chromosomal structural variations. Genus Caulobacter undergoes extensive genome rearrangements. Genomic studies in bacteria usually focus on the codingregions, but there is important information present in the intergenic DNA spaces inaddition to the regulatory elements involved in transcription. Recently, Ely published a new model for recombination in genus Caulobacter with simultaneousloss and gain of genes resulting from preferential recombination at non- homologous regions flanked by regions of homology. In my dissertation, I observedand catalogued hairpin structures at known sites of recombination in both closely and distantly related species to Caulobacter crescentus strain NA1000. To automate the process of identifying conserved base patterns in long sequences inbacterial genomes, I developed an unsupervised machine-learning pipeline using agglomerative clustering. These analyses have identified the presence of sequences capable of forming hairpins at the previously identified recombination hotspots. When additional Caulobacter genomes were examined, an increase in phylogenetic distance led to a decrease in the number of hairpins matching the model organism Caulobacter crescentus NA1000, with most of the differences seen in the loop sequence of the hairpin. I also observed that stem structures tendto remain consistent across species. We did observe changes in either the length or bases. This can be due to differences in sequence conservation as an outcomeof phylogenetic distance. The presence of these hairpin structures seemsto have been

conserved at sites of recombination suggesting that they may play role in initiating recombination by acting as substrates. It has also previously been shownthat Caulobacter crescentus uses Rho dependent termination machinery under stress. We identified some of the hairpin structures at sites of both rho dependent and independent termination in Caulobacter genus and compared it with previouslyidentified structures using ARNold for intrinsic termination and RHOTermPredict for rho-dependent termination. Our hairpin structures matched the ones identifiedwith ARNold but RHOTermPredict is designed for genomes with low GC %. The latter identified 6 times as many RUT sites as were genes, hence limiting our confirmation of Rho-independent terminators.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Bacteria are ubiquitous microbes found in very large numbers across life forms to carry out essential functions. These organisms are usually stable from one generation to next but are dynamic across evolutionary scale due to horizontal gene transfer, genome rearrangements and as outcomes of mobile DNA elements.Hence, maintaining the right balance of genome integrity with instability is essentialfor the survival of complex and dynamic life forms (1). Genome instability can arisefrom both intra-chromosomal and inter-chromosomal structural variations. While the former includes substitutions, insertions, and deletions, the latter includes inversions, translocation, duplications and transpositions. Inter-chromosomal genome variations are also called as genome rearrangements.

Genomes are dynamic. They are affected by many factors: the environment theyare found in, mutations during cell division, transposition activity of the jumping genes, to name a few. The evolution of biological diversity has occurred through these types of genetic changes, which differentiate closely related organisms fromeach other. DNA modifications in both genic and intergenic spaces can be seen atthree levels of observation: (i) point mutations leading to local sequence change.

DNA segment rearrangement by gene duplications, insertion/deletion and inversions and translocations (iii) acquisition of new DNA components via horizontal gene transfer. These measures contribute to the altered phenotypes o of bacteria. The underlying factors that lead to the evolution of genomic traits in bacteria can be seen at a multiscale level through interspecific and intraspecific comparisons. MUTATIONS: Mutations are vital for evolution. Every genetic feature

acquired by an organism is the result of a mutational variant or DNA acquired through HGT. The previously assumed effect of "neutral evolution" on intergenic DNA space is today in question. This is because regulatory elements in the codingregions of DNA are usually under constant selective forces, and intergenic DNA spaces have regulatory elements required for the functioning of the genome. It isimpossible to track deleterious mutations from both the coding and intergenic DNA space that have been eliminated from the population, and the obvious beneficial mutations in the systems are not the only representation of purifying selection. Purifying selection can also result from a combination of selective forces acting onthe DNA space. Ely et al. (3) compared three closely related Caulobacter crescentus genomes NA1000, CB1, CB2 and one more distantly related C. crescentus CB13 genome to identify potential genetic targets.

DNA segment rearrangement by gene duplications, insertion/deletion and inversions and translocations (iii) ac quisition of new DNA components via horizontal gene transfer. These measures contribute to the altered phenotypes o of bacteria. The underlying factors that lead to the evolution of genomic traits in bacteria can be seen at a multiscale level through interspecific and intraspecific comparisons. MUTATIONS: Mutations are vital for evolution. Every genetic feature acquired by an organism is the result of a mutational variant or DNA acquired through HGT. The previously assumed effect of "neutral evolution" on intergenic DNA space is today in question. This is because regulatory elements in the codingregions of DNA are usually under constant selective forces, and intergenic DNA spaces have regulatory elements required for the functioning of the genome. It is impossible to track deleterious mutations from both the coding and intergenic DNA space that have been eliminated from the population, and the obvious beneficial mutations in the systems are not the only representation of purifying selection. Purifying selection can also result from a combination of selective forces acting on the DNA space. Ely et al. (3) compared three

closely related Caulobacter crescentus genomes NA1000, CB1, CB2 and one more distantly related C. crescentus CB13 genome to identify potential genetic drivers of diversity. They showed that single base insertion-deletions. CHROMOSOMAL REARRANGEMENTS:

They are a part of the 'Divergence' concept, also known as the 'Biological designprinciple' of natural evolution. This can be seen when comparing the genomic DNAsequences of chimpanzees and humans. The DNA coding regions of both the organisms differ by 1.23% ifconsidering only point mutations, but the percentage increases to 5% after including insertions and deletions. The percentages increasefurther when duplications are included.

During the comparison of the three closely related 'Caulobacter' genomes (3), noinversion events were observed. However, when these closely related 'Caulobacter' genomes were compared to a more distantly related genome, eightinversions were observed indicating that they had occurred at a rate of one per 10-12 million generations.

HORIZONTAL GENE TRANSFER:

Horizontal gene transfer can be considered equivalent to the concept of 'Information transfer' where each of the organisms are agents of transfer, i.e., one is a transmitter that transmits information and the other as a sensor that intercepts it. These agents can also adapt, communicate and change the environment to match their requirements. In the comparisons described above, Ely et al. (3) foundthat INDEL events containing a few genes were horizontally transferred between closely related species at a frequency of 10-3 to 10-4 insertions per generations.

DRIVERS OF GENOME PLASTICITY:

Genome stability is usually maintained by DNA replication, recombination and repair. But these processes may also induce genome rearrangements and instability. Genome instability mediated by homologous or illegitimate recombination is carried out by relatedand repeated

sequences within the chromosome or specialized genetic elements like tRNA or mobile elements. Related sequences act as substrates for gene conversion and recombination between repeated sequences can lead to duplication, amplification or deletion. At the same time, recombination between inverted sequences can lead to DNAinversion (1). There are also numerous cooperating and antagonistic elements like DNA repair systems, mobile genetic elements, restriction modification systems, toxin-antitoxin systems that lead to horizontal gene transfer and gene redundancy. Image 1 below categorizes all these elements and the process by which they make bacterial genomes complex and dynamic (4). Ely (5) recently published a new recombination model where he found simultaneous gene gain and loss in genus Caulobacter resulting from preferential recombination at non- homologous regions flanked by regions of homology. It has been previously shown that that hairpin structures act as substrates to catalyze integration into the host sites (6) at recombination hotspots

In my dissertation, I show that intergenic sequences are repeated within closely and distantly related species of genus Caulobacter. I also identified and characterized hairpin structures within the intergenic sequences and provided evidence that they might be involved in both transcription termination and recombination during HGT and inversions.

CHAPTER 2

CHARACHTERIZATIONS OF LONG INTERGENIC AND HAIRPIN STRUCTURES IN CAULOBACTER CRESCENTUS

ABSTRACT:

DNA repeats within genomes are sequences with extensive similarities leading to functional overlapping or sequence recombination. Genomic studies in bacteria usually focus on the coding regions, but there is important regulatory information in the intergenicDNA spaces. This chapter focuses on the identification and functional distribution of longintergenic sequences and hairpin structures in Caulobacter crescentus. We show that many of the repeated intergenic sequences contain sequences capable of forming hairpin structures. These intergenic hairpin structures may play a role in transcription termination.However, in the recombinant CB2A strain hairpins were observed at more than 100 siteswhere recombination occurred as part of a horizontal gene transfer event.

INTRODUCTION:

Bacteria are universal and ubiquitous members of domain prokaryote that inhabit vast and varied sites including oceanic and terrestrial sub surfaces of earth,open ocean and deep portions of earth's crust, acidic hot springs, and radioactivewaste. They can also have a symbiotic or pathogenic relationship with animals. The evolution of bacteria can be a rapid process. Mutations occur not only by deletion and substitution, but also by horizontal gene transfer and genome rearrangements. There is an interesting degree of duality that a bacterial genome is exposed to constantly: maintaining a constant tradeoff between genome evolution and genome maintenance.

Genomes are dynamic. They are affected by many factors like the environment they are found in, mutations that occur during DNA replication, transposition activity, and horizontal gene transfer. DNA modifications in both genic and intergenic spaces can be seen at three level: (i) point mutations leading to local sequence change (ii) DNA segment rearrangement by gene duplications, insertion/deletion and inversions and translocations (iii) acquisition of new DNA components via horizontal gene transfer. Recent studies with Caulobacter crescentus have shown that while point mutations are relatively frequent, genome rearrangements occur less than once per thousand generations and horizontal gene transfer occurs a rate of once per 10 million generations. Bacteria are universal and ubiquitous members of domain prokaryote that inhabit vast and varied sites including oceanic and terrestrial sub surfaces of earth,open ocean and deep portions of earth's crust, acidic hot springs, and radioactivewaste. They can also have a symbiotic or pathogenic relationship with animals. The evolution of bacteria can be a rapid process. Mutations occur not only by deletion and substitution, but also by horizontal gene transfer and genome rearrangements. There is an interesting degree of duality that a bacterial genomeis exposed to constantly: maintaining a constant tradeoff between genome evolution and genome maintenance. Thus, the survival and evolution of these microbes requires a balance of maintaining genome integrity while allowing for a degree of instability. The field of bacterial genome rearrangements is generally focused on the reorganization of the coding DNA. But the bacterial intergenic DNA space,considered non-coding DNA, is a complex and dynamic system that includes critical regulatory elements. Deletions, duplications, inversions, insertions, and amplifications can disrupt genes, leading to phenotypic variation, genome evolution, and speciation. Rearrangements have also been shown to lead to the appearance of new sequences at the sites of these events. In addition, gene acquisition through horizontal gene transfer of DNA from other bacteria has been shown to radically transform bacterial pathogenicity, antibiotic resistance, and theutilization of unusual energy resources. As indicated above, a genome is comprised of both coding and non-codingregions with coding regions comprising 80 to 90% of most prokaryotic genomes. However, critical biological information is present in the intergenic

DNA space (IDS) including transcription factor binding sites for regulatory elements that impactgene expression, promoters and terminators for transcription of the adjacent genesnon-coding RNAs that regulate gene expression. Though bacterial genomes are streamlined, they contain small repeat elements whose origins and function are mostly unknown. Repeats restricted to single or closely related species are usually considered to have been acquired recently and are unlikely to affect fundamental processes. Also, most short repetitive sequences in bacteria have the potential for secondary structures that may enhance the stability of mRNA. Previously analyzed structures in 40 different bacterial genomes found non-random populations of such structures across all the genomes with most of the hairpins within the coding regions. But thehairpin structures found across intergenic regions were structurally stable found at 3'-end side of flanking CDSs.

DNA repeat regions, especially tandem repeats that are often seen in non–coding regions of eukaryotic genomes are seldom seen in prokaryotic systems. One potential explanation is that the bacteria need to streamline their DNA which may confer a selective advantage by reducing the time needed for genome replication The DNA sequence repeats (DSR) that are present in prokaryotic genomes are usually less than 400 bp and are primarily found as multiple copies in intergenic regions of the chromosome. In terms of length, they can further be divided into short DSR (<200 bp) and long DSR (>200 bp & <400 bp). The short bacterial DSR's have been classified into two broad categories. MITE (miniature inverted-repeat transposable element) and REP (repetitive extragenic palindromic sequence). There are further subclasses of the short repeat elements such as REP2-5 units, YPLA/RU2, bcr elements and CRISPR sequences. The short DSR arealso irregular and less defined with the potential to fold into stable secondary structures at both the DNA and RNA level and function as regulatory elements responsible for regulating gene expression. Long DSR's are uncommon in prokaryotes and are known to be subject to negative selection. They are usually variable in length due to DNA polymerase slippage and/or recombination.

Caulobacter is a genus of gram negative, oligotrophic Alpha proteobacterium that

undergoes asymmetrical cell division. They produce two distinct cell types: a motile swarmer cell and a sessile stalked cell. Only stalked cells are capable of replication and cellular division, and swarmer cells must undergo differentiation into stalked cells to proliferate. Division of stalked cells results in two daughter cells; a stalked cell that continues to serve as a parent cell,and a mobile flagellated swarmer cell. Due to its unique lifestyle and well- established system for genetic analysis, Caulobacter is an important model organism for studying cell cycle regulation, asymmetric cell division, and cellular differentiation

Previously, multiple copies of three distinct DSRs were identified and called CcrM-associated intergenic repeat sequences or CIR in the intergenic spaces of the C. crescentus NA1000 genome. Though these repeats seem to resemble the IRU/ERIC sequences and have also shown to have some properties of MITE elements, the Caulobacter motifs were identified by the presence of a consensus CcrM binding site. CcrM is a methyltransferase found in α- proteobacteria that methylates the 'A' residue in the nucleotide sequence 'GANTC'. In this study, we identified all long intergenic sequences the C. crescentus NA1000 genome that contained a repeated region. We identified 390 long intergenic sequences that ranged in length from 43 base pairs (bp) to 2513 bp (Supplementary Table 1). We also found that 258 of these long intergenic sequences contained sequences capable of forming hairpin structures. hairpin structures were further classified into 34 hairpin repeat families and 66hairpin structures (Supplementary Table 2).

METHODS:

We used a four-step strategy to identify repeated elements and hairpin structures (Fig. 2.1). The steps are explained in detail below: Identification of the sequences, Length determination and Characterizations.

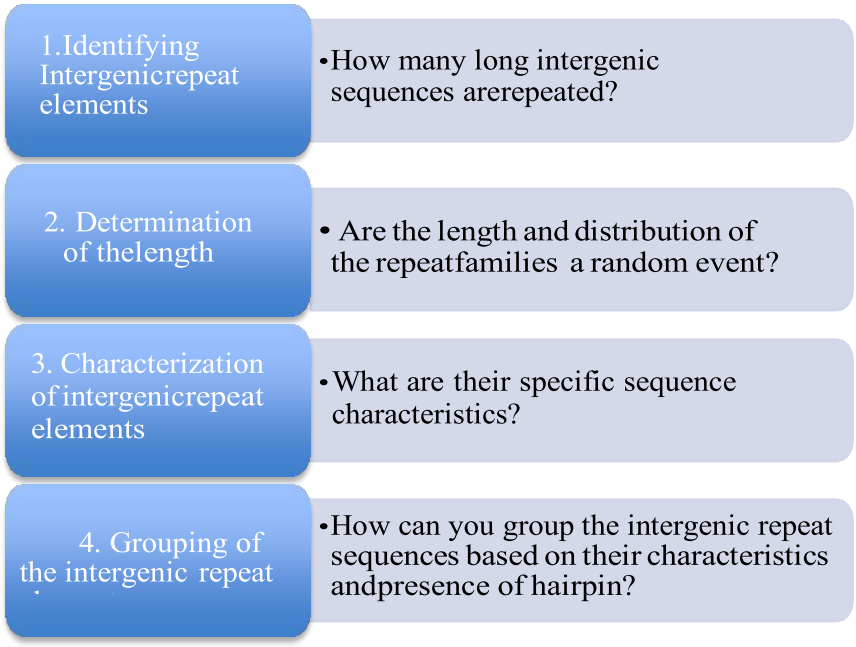

Figure 2.1 Clustering of long intergenic repeats in C.crescentus.

Step 1. Genomic Sequence and Annotation data:

The Caulobacter crescentus NA1000 genome sequence was downloaded from GenBank (accession number NC_011916.1), and the intergenic sequences were obtained using ARTEMIS software (7).

Step 2: Sequence alignment:

Since, HC is accurate and fast, it is by far the most used clustering methodfor sequence alignments. For unique N sequences in a group, N(N-1)/2 unique pairwise comparisons were made in the form of similarity. As we had sequences with variable lengths, a maximum of 50 folds and a maximum distance of 30 bases ensured stringency in computed structures. Structures with 3 to 8 base loops and stems with a minimum length of 6 bp were classified as hairpins and replicates found using Artemis software (7).

Step 3: Agglomerative hierarchical clustering and heat map:

The standard way of displaying and identifying structure within –omics datais achieved by hierarchical clustering, but the associated structural visualization ofdata and identification of

subclusters is not intuitive. Therefore, to identify subclusters, the sequences of each family were subjected to pairwise or multiple sequence alignment depending on the number of sequences and the consensus sequences were then obtained.

Step 4: Hairpin identification and distribution:

The DNA hairpins were identified using MFOLD (9) under following conditions: (i) an upper bound of 50 on computed folding. (ii) The maximum distance between paired bases of 30.

RESULTS:

All 3082 intergenic regions were blasted against each other resulting in 6849 matches. Most of the intergenic regions were unique or too small to producea significant result, but 390 intergenic regions contained repeated sequences thatwere designated as Caulobacter intergenic repeat elements (IRE) that ranged in size from 43 bp to 2513 bp (Table 1). Three of these repeated sequences had been identified previously and designated CcrM-associated intergenic repeat(CIR) sequences.

The fourth CIR family (CIR3) previously described has been re-annotated as a family of repeated mobile elements. Each of the CIR sequences was further analyzed for the presence of hairpin structures, and 258 of the 390 CIRsequences were found to have hairpin structures.

Table 2.1: Characteristics of repeated sequences

Length of the smallest sequence	43 bp
Length of the largest sequence	2513 bp
Average length	243 bp

The intergenic regions containing repeat elements were also classified according to the orientation of the adjacent genes (Table 2). Most of these intergenic regions contained both a promoter region and a terminator region. Supplementary Table 2 shows that 290 of the long

intergenic sequences have one or more of thetranscription factor binding sites that control gene expression during the Caulobacter cell cycle. Since our laboratory was one of the first to propose that DNA could form hairpin structures with important biological functions, we decidedto check the IRE for the presence of hairpin structures using the MFOLD programunder standard folding conditions, and we found that 258 of the 390 intergenic regions contained at least one sequence that could form a hairpin structure. (Table2 and Fig. 2.2).

Table 2.2 Properties of the long intergenic repeat sequences

	Distribution of the IRE	Number of repeats
1.	Total IRE with repeat regions	390
2.	Number of IRE at the end of 2 genes(double terminators)	44
3.	Number of repeats between 2 promoters(sense and antisense)	98
4.	Number of repeats between a promoterand a terminator	247
5.	Number of IRE with hairpins	258 sequences
6.	Number of IRE without hairpins	131 sequences
7.	Number of repeats near tRNA	2 sequences

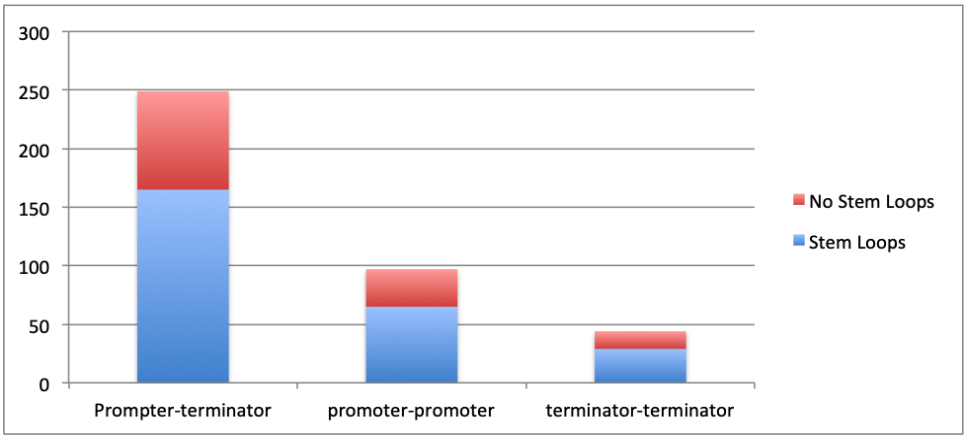

Figure 2.2 IRE distribution relative to the flanking genes

The hairpins were grouped into 34 families based on the sequences of the stems.(Table 3). The hairpin families have a minimum of three to a maximum of 32 members. In addition, 66

hairpins were present only once in the genome (Table 4). Hairpins found near 3' end of genes were found to have both or either rho- dependent or rho-independent terminators (Supplementary Table 1). We have observed 186 hairpin structures upstream downstream of the previously identified transcription stop sites in Caulobacter crescentus leading to the assumption that they may have a role to play in the bacterial transcription process. Bacterial transcription termination is an important regulatory step of gene expression. Transcription in bacteria can terminate by two different mechanisms: Rho independent or intrinsic termination and Rho dependent termination. The intrinsicterminators have a GC-rich hairpin with stretch of 6-8 uridine residues, while the Rho-dependent terminators rely on the rich and G poor nascent RNA with regularlyspaced cytosines, called a RUT site /Rho utilization site. The site of termination isusually 10-20 nt downstream to RUT site and not more than 100 bp downstream. We have identified both rho dependent and rho independent terminators (Supplementary Table 3) in the intergenic regions using the ARNOLD(10) and RHOTermPredict (11) software packages respectively. Some families of hairpins were also found to be associated with or near insertion sequences (mobile elements) and non-coding RNA. We also identified hairpins atnearly all of sites of recombination hotspots (Supplementary Table 2) where recombination and gene loss have occurred repeatedly in C. crescentus genomes(54). The rich and G poor nascent RNA with regularly spaced cytosines are called a RUT site /Rho utilization site. The site of termination is usually 10-20 nt downstream to RUTsite and not more than 100 bp downstream (40-41). We have identified both rho dependent and rho independent terminators (Supplementary Table 3) in the intergenic regions using the ARNOLD (10) and RHOTermPredict (11) software packages respectively. Some families of hairpins were also found to be associated with or near insertion sequences (mobile elements) and non-coding RNA. We also identified hairpins at nearly all of sites of recombination hotspots (Supplementary Table 2) where recombination and gene loss have occurred repeatedly in C. crescentus (54).The rich and G poor nascent RNA with regularly spaced cytosines are called a RUT site /Rho utilization site. The site of termination is usually 10-20 nt downstream to RUTsite and not more than 100 bp downstream (40-

41). Some families of hairpins were also found to be associated with or near insertion sequences (mobile elements) and non-coding RNA. We also identified hairpins atnearly all of sites of recombination hotspots (Supplementary Table 2) where recombination and gene loss have occurred repeatedly in C. crescentus. Alternatively, RNA polymerase or transcription factors could potentially recognize hairpins present on ssDNA or dsDNA extrusions. For example, it has been shown that hairpin formation involvingthe transcription start site and spacer sequence between promoter leads toregulation of transcription in Ebola virus. It has previously been shown that Caulobacter crescentus uses Rho dependent termination machinery under stress.RHOTermPredict could not identify the RUT sites and pause sites that are hairpins in Caulobacter genomes since the program was designed for low GC organisms. Caulobacter crescentus has high genome GC% and hence the program identifiedSIX times as many RUT sites as there were genes (27496/4097). Thus, the misidentified sites greatly outnumber the true termination sites. The absence of long intergenic repeats is usually attributed to the selective pressure to maintain the compactness of the genome. But it has been have shownthat long repetitions exist across bacterial prokaryotes, and they can be involved in recombination and horizontal gene transfer. When a gene is transferred horizontally, there have been two expected outcomes: (i) the transferred gene inserts itself without overwriting any gene and instead creates a new locus thus leading to an increase in genome size. (ii) or the new gene can replace an existinghomologous copy, thus preserving the total number of genes in the recipient genome. In the Caulobacter crescentus CB2A genome, 114 insertions of genetic material were horizontally transferred from the closely related NA1000 strain (54). These insertions led to a new recombination model where non-homologous regions wereflanked by regions of homology without the involvement of any mutational processand that in contrast to the two models described above, HGT usually involves thereplacement of nonhomologous genes. In this chapter, we have shown that at eachof these insertion sites, there are usually complete and sometimes incomplete hairpin structures flanked by homologous sequences. (Supplementary Table 2). These hairpins are found either in one or the other or in both the genomes at the sites where horizontal transfer

occurred. The position and identity of the hairpins are conserved across both the genomes. In most cases, the hairpins have 6 bp stems, but 4 or 5 basepair stems were also observed at a few sites. Also 8 HGT sites either had a tRNA or a transposase gene at the sites of insertion instead of ahairpin structure. Thus, that we propose that in addition to their possible role in transcription termination, the hairpin structures that we have identified in these Caulobacter genomes serve as sites that initiate recombination during HGTevents.

CHAPTER 3

IDENTIFICATION AND DISTRIBUTION OF LONG INTERGENIC SEQUENCES WITH HAIRPIN ELEMENTS IN GENUS CAULOBACTER

ABSTRACT:

Bacterial genome size variation is usually dependent on acquisition and lossof functional accessory genes. Though not common in prokaryotes, repetitive sequences are known to play important functional roles required for the maintenance of the bacteria. In the first chapter, we catalogued and characterized the presence of hairpins in repeated intergenic sequences of Caulobacter crescentus strain NA1000 about termination and recombination. In this chapter, we extend the analyses to closely and distantly related strains and species of genus Caulobacter and identify similar and dissimilar repeated intergenic sequences with hairpin structures.

INTRODUCTION:

The survival and evolution of microbes require a balance of maintaining genome integrity while allowchapter 1ing for a degree of instability. Prokaryotic genomes are usually compact to maintain selective pressure for rapid DNA replication. A genome is comprised of both coding and non-coding regions with 80 to 90% of prokaryotic genomes made of coding regions. However, critical biological information is present in the intergenic DNA space (IDS) that is necessary for the regulation of gene expression. The IDS are also sites for DNA repeat sequences. DNA repeat regions, especially the long repeats that are 26 base pairs and greater are unlikely to exist by chance alone, and therefore, they must be important to the biology of the organism.

But the essential genome is conserved. The field of bacterial genome rearrangements is extensively focused on the reorganization ofthe coding DNA. But the bacterial intergenic DNA space is a complex and dynamicsystem that includes critical regulatory elements. Rearrangements in intergenic regions have been shown to change microbial phenotypic characteristics.

In the previous chapter, I have identified long intergenic repeats withhairpins in Caulobacter crescentus strain NA1000 and hypothesized that they might be involved in transcription termination and/or homologous recombination. In this chapter, I analyze the conservation of the long intergenic repeats with hairpin structures in closely and more distantly related members of genus Caulobacter across the phylogenetic distribution.

METHODS:

Extraction of intergenic data:

Complete Genome sequences of three different Caulobacter species weredownloaded from GenBank. The intergenic sequences were extracted using the ARTEMIS software (7).

Whole genome phylogeny:

Pairwise comparisons among the sets of genomes were carried out in the TYPE(STRAIN) GENOME SERVER (12). The Tree was inferred with FastME1.1.6.1 using GBDP distances calculated from genomes and branch lengths were scaled in terms of GBDP distance formula d5. The phylogenetic treewas created using the newick format file submitted to TreeDyn (13).

Identification of repeated sequences and hairpin structures across genomes:

All the intergenic sequences were subjected to heuristic sequence matching using the local BLAST algorithm (14). Each of the intergenic sequenceswas matched with previously identified intergenic sequences in Caulobacter crescentus strain NA1000. The hairpins found within these sequences wereidentified using MFOLD (10) under the following conditions: (i) An upper bound of 50 on computed foldings. (ii)The maximum distance between paired bases of 30. Structures with 3 to 8 base loops and stems with a minimum length of 6 bpwere hairpins.

RESULTS:

The Ely laboratory has a long- established interest in Caulobacter genetics andgenome evolution. In this study, we compared four strains of closely related C.crescentus genomes with the more distantly related C. segnis TK0059 genome. The genome characteristics are shown in Table 3.1

Table 3.1: Genome characteristics of the genus Caulobacter.

Caulobacter strains	Genomesize	GC%	Number ofCDS/Protein coding sequences
NA1000 NC_011916.1	4.04MB	67.2	3886
CB1 NZ_CP023314.2	4.14 MB	67.2	3990
CB2 NZ_CP023313.2	4.12 MB	67.2	3896
CB13	4.14 MB	67.1	3140
TK0059 NZ_CP027850	4.66 MB	67.70	4201

The sequences of each genome were subjected to sequence matching using the local BLAST algorithm (9). Significant sequences that folded into hairpins were found distance from NA1000.The sequences of each genome were subjected to sequence matching using the local BLAST algorithm (9). Significant sequences that folded into hairpins were found distance from NA1000. The sequences of each genome were subjected to sequence matching using the local BLAST algorithm (9). Significant sequences that folded into hairpins were found distance from NA1000. The sequences of each genome were subjected to sequence matching using the local BLAST algorithm (9). Significant sequences that folded into hairpins were found distance from NA1000. Once the repeated sequences were identified each of the sequence was subjected to folding to identify hairpin loops. The hairpin loops were further analyzed to group them into families. The hairpin families were then subjected to cataloguing to determine the distribution of the said structures across the genomes of the genus Caulobacter. Significant sequences that folded into hairpins were found distance from NA1000. . The hairpin loops were further analyzed to group them into families.

The repeated sequences were also subjected to MFOLD (10) to identify hairpin

structures within them. Conserved hairpin stem sequences were observed in all the genomes, but some variation in the sequences of the hairpin loops wasobserved (Supplementary Table 1). We also determined the position of eachhairpin relative to the transcription promoter and terminator regions (Table 4). Inall strains, hairpins in regions that contained both a promoter and a terminator werethe most common.

DISCUSSION:

In recent years, long repeats have been shown to play an important role in the evolutionary adaptation of bacteria to environmental changes. It has previously been shown that genome rearrangements are usually observed only between distantly related genomes, but HGT events can be observed in closely related species. This work led to the identification of the new model of HGT where simultaneous gene loss and recombination in closely related strains of genus Caulobacter occur at non-homologous regions that are flanked by regions of homology. This preferential recombination model was further analyzed in Chapter1, and I found that in most cases, the sites of gene insertions are flanked by hairpin structures in one or the other genome. In this paper. we focus on the distribution of long intergenic sequences with hairpin structures and their positional organization within other Caulobacter genomes. Over 90% of all intergenic sequences were repeated between closely and more distantly related Caulobacter genomes. To better identify the presence of hairpinsin the repeated intergenic sequences, we chose a cutoff of 28 bp and higher to identify the hairpin structures. As expected, the number of hairpins matching Caulobacter strain NA1000 was reduced with increased phylogenetic distance. Due to differences in sequence conservation, hairpins corresponding to the matched repeated sequence to NA1000 were not always the same. The most frequent differences were changes in the sequence of the hairpin loop suggesting that the loop sequence may not be as important as the hairpin structure itself. Otherchanges included changes in the stem sequence that changed the length of the stem or that changed the sequence of the bases in the stem while maintain the ability to form a hairpin.

CHAPTER 4
HIERARCHICAL CLUSTERING OF LONG DNA REPEATS TO EXTRACT MEANIGFUL SEQUENCE PATTERNS.

ABSTRACT:

In the previous chapters, we have shown that most horizontal gene transferevents in genus Caulobacter occur in intergenic DNA spaces. There is simultaneous gene loss and gene gain through recombination thus maintaining thegenome integrity of these bacteria. In addition to the non-homologous regions flanked by regions of homology, we found hairpin structures at each of the recombination hotspots. In order to automate our analyses and work with larger and variable datasets of intergenic sequences, I designed this pipeline/package in R. This clustering algorithm can be used for many purposes including but not limited to RUT sites to find rho dependent terminators, transcription start site motifsto name a few.

INTRODUCTION:

Advancements in sequencing technologies have led to a deluge of geneticdata. Today data generation has surpassed data analyses, hence requiring strategies and techniques for appropriate interpretation and evaluation. DNA sequence clustering is one such approach that helps analyze the data. DNA sequence clustering relies on two complementary approaches: comparative classification and unsupervised clustering. The former approach ensures the identity of a new sequence by matching it to a curated database. Butthis method cannot be used for the analysis of novel sequences and that is whenunsupervised clustering is valuable.

The established approach of clustering involves building a multiple sequence alignment of

all sequences, followed by a pairwise distance matrix based on the alignment and finally clustering the resulting matrix (5). The clusteringalgorithm most often used in discovering hierarchy is the agglomerative bottom-upclustering. This method comes with its own set of challenges as multiple sequencealignment of large volumes of sequence data becomes computationally difficult thus giving rise to NP-hard problems. Also, when working with raw data, the clustering algorithm proceeds through a series of local improvements, making them sensitive to local maxima. And if there are no, pre-processing steps prior to agglomerative clustering, small perturbations in the data can make the structure of the constructed hierarchies brittle.

In this paper, we will look at the clustering of long intergenic DNA repeat sequences. The repeat sequences are usually found across systems. The protein-coding component of a human genome accounts for only 1.2% of the total DNA with 43% of the sequenced euchromatic portion of the genome consisting of repeated and mobile DNA elements. In bacteria, repetitive sequences account foranywhere between 5-10% of the genome. Caulobacter is a genus of gram negative, oligotrophic bacteria with a rod- like structure and asymmetrical cell division. They produce two distinct cell types:a motile swarmer cell and a sessile stalked cell. Only stalked cells are capable of replication and cellular division, and swarmer cells must undergo differentiation into stalked cells to proliferate. The division of stalked cells results in two daughtercells; a stalked cell that continues to serve as a parent cell, and a mobile flagellatedswarmer cell. Chromosome replication and cell division occurs only in the stalkedcell stage so the swarmer cell must go through a maturation process and becomea stalked cell before it can replicate. Due to its unique lifestyle and well-established system for genetic analysis, Caulobacter is an important model organism for studying cell cycle regulation, asymmetric cell division, and cellular differentiation.

In the previous chapters I identified new classes of hairpin structures that play important roles in potential transcriptional termination and recombination. In this chapter, I automate the process of identifying conserved base sequences across bacterial genomes using hierarchical agglomerative clustering. Materials and Methods: ‘Clustering’ is the process of organizing data

into disjoint classes such that: constituent members of the class have high ‘intra-cluster similarity’ and constituent members of other classes have high ‘inter-cluster’ dissimilarity. An unsupervised algorithm, clustering, does not depend on predefined classes and training examples to categorize the data objects (15). Instead, clustering groups objects based on degrees of similarity. Agglomerative clustering follows the ‘bottom-up’ approach with each object initially being a cluster by itself. At each step of the algorithm, two clusters related to each other are combined to form a larger cluster or a node. This process is iterated until all points combine to form a single node. Summarily, hierarchical agglomerative clustering is the method of combining ‘n’ small groups into a single large group where ‘n’ is the number of data points (16). Agglomerative clustering includes four common methods of linkage amongst the clusters; ‘single linkage’ based on nearest distance, ‘complete linkage’ based on farthest distance, ‘average linkage’ based on average distance a ‘wards linkage’ based on analysis of variance. In single linkage methodology, the clusters are combined due to single data points being close to each other despite many data points in each cluster being distant.In complete linkage, all data points are like each other thus making the clusters compact. Average linkage methodology generates homogenous clusters formed by arithmetic mean of all proximities between data points of one cluster with the data points of another. And finally in ward’s linkage, clusters are formed by analysisof variance between them.

Selection of the linkage type depends on the dimensions in the space that represent the characteristics upon which the data points of clusters are compared. The similarities between cluster data points can be measured by either identifying the correlation of entity scores on the dimensions by cophenetic correlation or by identifying the distance between the most similar data points. In this chapter, we will use correlation between the distance matrix andthe cophenetic distance to assess the choice of clustering linkage. To determine the stability of cluster, it is important to evaluate data representation. This assessment can be done using bootstrapping. In this chapter, we will use correlation between the distance matrix andthe cophenetic distance to assess the choice of clustering linkage. To determine the stability of cluster, it is important to evaluate data

representation. This assessment can be done using bootstrapping. Bootstrapping ensures rigorous selection of data points in a cluster and removes any unnecessary artefacts introduced as a product of the clustering algorithm.

METHODOLOGY:

Preprocessing of DNA:

The extraction and pre-processing of intergenic DNA into families is elucidated in the first two chapters. In brief, intergenic DNA sequences were downloaded from NCBI ftp site and subjected to BLAST analyses and then grouped into families. For this case study, we will use an intergenic sequence with a conserved

Multiple sequence alignment of families:

The newly grouped families are subjected to multiple sequence alignment followed by hierarchical clustering (HC). Since, HC is accurate and fast, it is by far the mostused clustering method for sequence alignments.For unique N sequences in a group, N(N-1)/2 unique pair wise comparisons are made in the form of similarity scores. In our analyses, we use CLUSTALOMEGA (18) for sequence alignment. This program administers different score-pair matrices when sequences of differing similarities are aligned and also uses seededguide trees and HMM profile-profile techniques to produce alignments amongst three or more sequences.

Agglomerative hierarchical clustering and heat map :

The standard way of displaying and identifying structure amongst –omics data is achieved by hierarchical clustering (26). But the associated structural visualization of data and the identification of subclusters is not intuitive (27-30).

Distance Matrix :

The subsequent heatmap generation depends on clustering of distance matrix of similarity scores of rows and columns of the data. This is done using the distancemeasure that determines the difference between the two data points and scaling the data using rank analysis.

In our analysis, we use Pearson's parametric correlation for the distance measure and Spearman's non parametric rank correlation for scaling the data. This is because there is variability in length and conservation of the sequences leading ton elliptical distribution of data. As there are outliers due differences in length of the sequences, Spearman's ρ limits the outlier to the value of its rank

Pearson's correlation:

This coefficient determines the strength of the linear relationship between two datapoints and is measured as follows:

$r = Cov(x,y)\ \sigma x\ \sigma y$ where

$\sigma x = \Sigma (x - \bar{x})\ 2$ is the standard deviation of x

$\sigma y = \Sigma (y - \bar{y})\ 2$ is the standard deviation of y

The correlation value usually ranges between -1 and 1

A value equal to or near 0 implies non-linear relationship

And a value closer to 1 or -1 shows a stronger linear relationship .

Spearman Rank correlation:

This correlation sorts observations and computes the degree of similarity by rank. The main advantage of using this correlation is that it is neither sensitive to outliers nor is it linked to distribution of data. The rank between two data points is measured as follows:

$\rho = Cov(rgx,rgy)$

$\sigma rgx\ \sigma rgy$ where

$\sigma rgx = \Sigma (x - \bar{x})\ 2$ is the standard deviation of rgx

$\sigma rgy = \Sigma (y - \bar{y})^2$ is the standard deviation of rgy The correlation is always between

-1 and 1

Values close to either -1 or 1 indicates strong relationship

Cluster Linkage determination:

After the calculation of distances between a pair of individual data points, we calculate the distances between the clusters by linkage determination. They are usually dependent on object-object distances and each type of linkage results in different hierarchical clustering There are four major linkage types and they are explained as follows:

Single linkage or nearest neighbor or minimal jump method:

This type of linkage measures the smallest distance between any two closest points in the two clusters of consideration and it is represented as follows:

$d(C(ij),Ck = \min\{d(Ci,Ck),d(Cj,Ck)\}$

Advantage:

This linkage gives small clusters.

Disadvantage:

This linkage produces skewed hierarchy, thus giving rise to a chaining problem.

Complete linkage or maximum jump method:

This type of linkage measures the largest distance between any two closest datapoints in the two clusters of consideration and is represented as:

$d(C(ij),Ck = \min\{d(Ci,Ck),d(Cj,Ck)\}$

Advantage:

This linkage gives small clusters

Disadvantage:

The linkage is very sensitive to noise.

All the characteristics of the hairpin structures are subjected to the sequences and their conservation. Observation similar or different to linkage will establish if the given sequence is conserved or not conserved within the genome.

Average linkage:

This method is halfway between the above two methods. This linkage takes the mean of all the data points in cluster i to cluster j.

The average distance can be defined as WPGMA or weighted pair group methodwith arithmetic mean, UPGMA or unweighted pair group method with arithmetic mean , UPGMC or unweighted pair group method centroid and WPGMC or weighted pair group method centroid.

The linkage can be represented as: $d12 = 1 \sum dij$

$|C1||C2|$

$i \in C1, i \in C2$

Advantage:

This linkage gives similar size and variance of clusters.

Disadvantage:

This linkage is not robust.

Wards linkage:

This linkage method uses analysis of variance to minimize the variance.

Advantage:

This linkage minimizes inertia and is efficient

Disadvantage:

This linkage gives rise to smaller clusters if there is high variability in data points. The choice of best clustering method is determined by calculating the copheneticcorrelation coefficient for each of the families using the distance matrix and thecophenetic distance.

The cophenetic coefficient is a linear correlation between the dissimilarities dij of each pair of observations (i j) and their corresponding cophenetic distances dcophij . The cophenetic distance is also known as the intergroup dissimilarity when observations i,j merge together initially in the same cluster.

This correlation can thus be represented as:

$$CCC(D,Z) = Cor(D,Z) = \sum i<j\ (Dij-D^{-}\ )(Zij-Z^{-}\ )$$

$$\sqrt{\sum i<j\ (Dij-D^{-})2\ \sum i<j(Zij-Z^{-}\ )2}$$

Where D = Distance matrix based on dZ = distance matrix

D^{-} = mean of Dij

Z^{-} = mean of Zij

The closer the value is to 1, closer is the appropriate linkage reflecting the data. . Inall of the families, either complete linkage or ward linkage had the highest values and accordingly, linkage method with the highest values was used for the agglomerative hierarchical clustering of each of the family. Following the above, each cluster are subjected to bootstrap evaluation.

Bootstrap evaluation of cluster:

To establish if cluster representations are meaningful, introducing plausible variation in the dataset can validate the data. We use cluster boot function from ‘fpc’ package in r to establish the stability of the cluster (19).

Cluster boot uses Jaccard coefficient to measure similarity between two clusters. Jaccard similarity between two clusters A and B is determined by the ratioof number of data points at the intersection of A and B over the number of elementsat union of A and B. The algorithm was run in the following way:

Cluster the data:

Draw a new dataset that is of the same size of the original by resampling the original dataset with replacement and then cluster the new dataset.

For every original cluster, find a similar cluster from the new cluster and compute the value. If Jaccard coefficient is less than 0.5, then the original cluster is dissolved, as it will not show up in the new clustering.

Identification of representative sequences of the clusters to get consensus:

After removal of irrelevant sequences from original families, the new data are now reclassified as clusters and are then subjected to another hierarchical clustering to identify the representative sequences. These were obtained by findingthe 'medoid' of clusters that are computed from the distance matrix. It is the clustermember with minimum pairwise distance to all the other members of the clusters.The Medoid sequences of each cluster within the new clusters are then subjectedto pairwise or multiple sequence alignment depending on the number of clusters and the consensus sequences are then obtained.

RESULTS:

A: Phylogenetic tree of blast results of previously identified repeated intergenicsequence:

We used a single repeated intergenic sequence with a stable hairpin found at the recombination site from Caulobacter crescentus strain NA1000 and used BLAST (14) to identify sequence matches to all the organisms in the order Caulobacterales. The images below show the phylogenetic distribution of the sequences and the distribution of the cluster linkage of hairpin structures across the genome of Caulobacter genus.

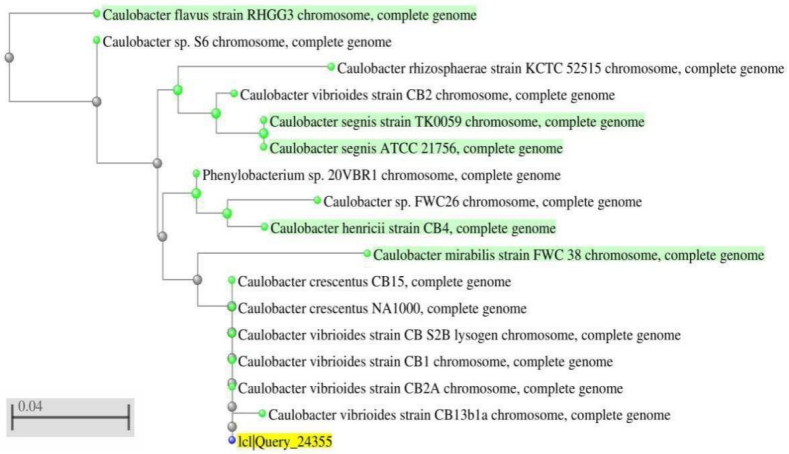

Figure 4.1: Phylogenetic tree based on agglomerative clustering and dendrograms.

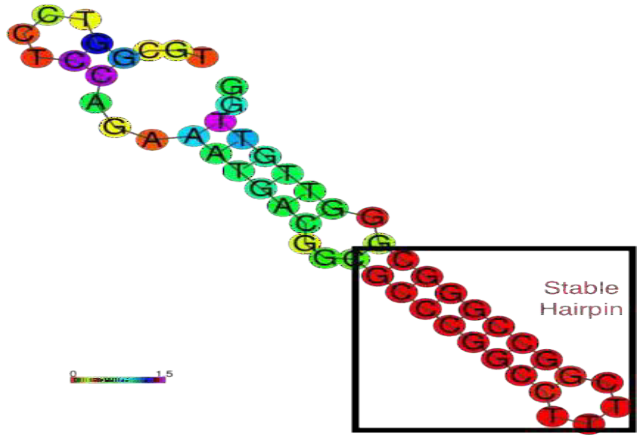

Figure 4.2: Identification of clusters based on agglomerative clustering and dendrograms

CLUSTALOMEGA (18) was used to address the differences in length and conservation of the bases in the intergenic sequences. Pearson coefficient was used to calculate the distance measure of the row matrix based on the sequence similarity scores. Spearman's nonparametric rank correlation was also used to scale the data. Z value equaling to zero or near zero shows a non-linear relationship amongst the sequence conservation and values closer to 1 or -1. This helps establish the distribution of the conserved consensus sequences within a family of the sequences and across the family of sequences.

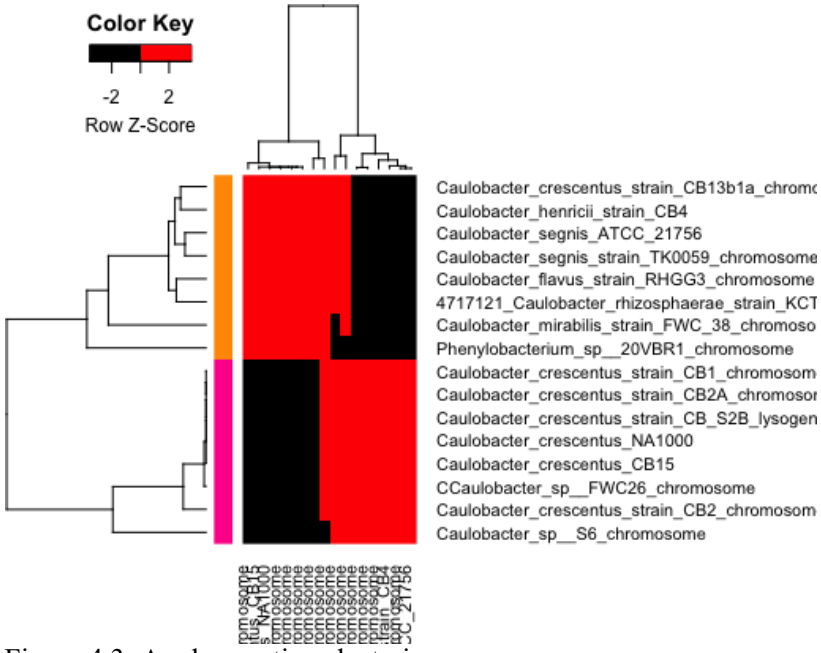

Figure 4.3: Agglomerative clustering

The above images show the similarities and differences between the clustered sequences. Agglomerative clustering helps determine the similarities and differences between consensus of the clustered sequence and quantify their distribution across the conserved consensus sequences. distribution across the conserved consensus sequences. Dendrograms give graphical representation of individual data points from the hierarchical heat maps. Figure 4.2 shows the two important clusters forming whilefigure 4.3 helps identify the individual units or sequences (in this case) corresponding to each cluster.Agglomerative clustering helps determine the similarities and differences between consensus of the clustered sequence and quantify their distribution across the conserved consensus sequences. distribution across the conserved consensus sequences. Agglomerative clustering helps determine the similarities and differences between consensus of the clustered sequence and quantify their distribution across the conserved consensus sequences. distribution across the conserved consensus sequences.

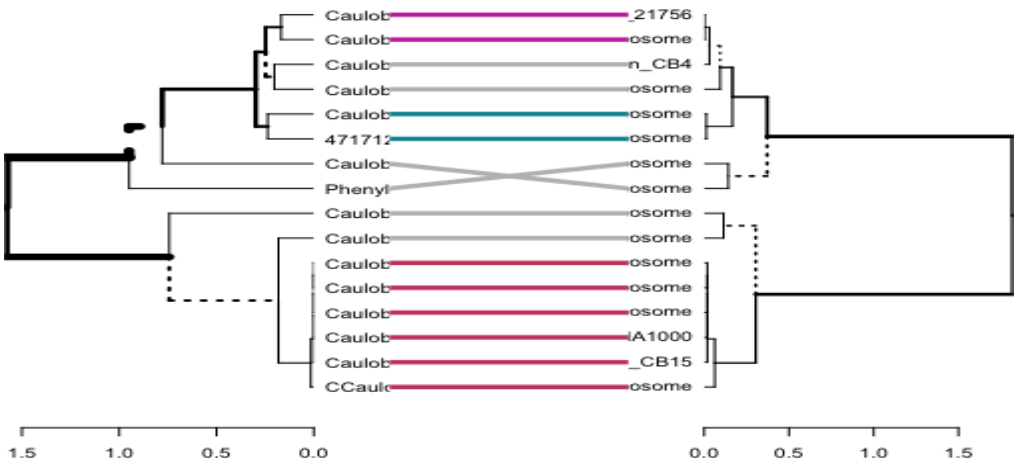

Figure 4.4: Dendrograms of clusters

Linkage determination to establish distance between clusters using multipleDendrograms:

In order to establish the appropriate distance measure between the clusters,we first identify the appropriate type of linkage that will ensure the stability of the clusters. Cophenetic distance is usually calculated to determine or identify the bestlinkage methods to Closer the value is to 1, better is the cluster linkage.

Table 4.1: Cophenetic values of linkage determination in Dendrograms

	Complete linkage	Single linkage	Average Linkage	Centroid
Complete linkage	1	1	1	0.99
Single linkage	1	1	1	1
Average linkage	1	1	1	1
Centroid	0.99	1	1	

Bootstrap evaluation of clusters:

Bootstrap evaluation is a reliable method to assess is the phylogenetic treeconstructions are statistically quantifiable (34). In this part of the analyses, we introduce plausible variation in the dataset to validate the data and establish the stability of the cluster. We use 'Jaccard's coefficient to determine similarity between the members of the cluster and dissolve or remove to the next cluster if they are deemed unstable.

Comparison of multiple sequence alignment of all the matches vs. the sequence alignment of clusters.

As seen in figure 4.5, the expected consensus sequence does not have thestable hairpin structure. In figure 4.6, Only Caulobacter crescentus strain 13b hasthe complete sequence while all the members of cluster 2 in image 7 have the same conserved sequence.

Figure 4.5: Multiple sequence alignment of the intergenic sequences

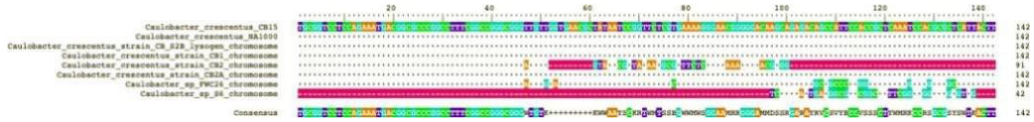

Figure 4.6: Multiple sequence alignment of sequences in cluster 1

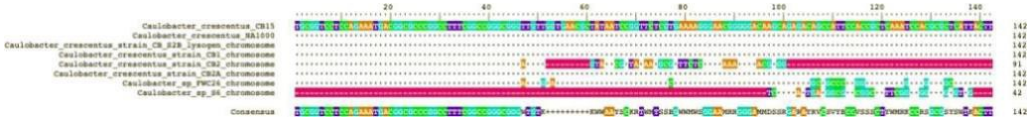

Figure 4.7: Multiple sequence alignment of sequences in cluster 2

DISCUSSION:

I used the example of intergenic sequences found at recombination hotspot in genus Caulobacter to show the analyses by the pipeline. These sequences ranged in length from 54 bp to

150 bp. Initially analyses where phylogeny was carried out using Phylogenetic tree based on the neighbor joining method of pairwise alignments, there were multiple clusters (Figure4.1). Our pipeline was able to establish two major clusters. Cluster 1 had a consensus sequence with no conserved hairpin structures. However, it was interesting to note that within the cluster 2 that had a consensus of the intergenic sequence with a stable hairpin, we identified new strains (i.e.) Caulobacter strain S6 (Isolated from rocky mountain soil), Caulobacter S2B: (A lysogenic phage derivative of NA1000 isolated in our laboratory) and Caulobacter strain FWC26 (shown by our laboratory to be a close relative of CB13). Unlike the phylogenetic tree that focuses on the slow evolution within a genome (e.g. point mutations), Agglomerative-clustering focuses on genome rearrangement events like insertion and deletion. This methodology can be used to compare sequences of variable lengths and unique sequence patterns and help glean important genetic information.

CHAPTER 5
CONCLUSIONS

The survival and evolution of these microbes requires a balance of maintaining genome integrity while allowing for a degree of instability. A genome is comprisedof both the coding and non-coding regions with 80 to 90% of prokaryotic genomes made of coding regions. However, critical biological information is present in the intergenic space in the form of promoters and terminators for transcription of the adjacent genes and binding sites for regulatory elements that impact gene expression. Hence, the intergenic DNA sequences regulate how the codingregions are expressed. Genetic acquisition through horizontal gene transfer of DNA from other bacteria has shown to radically transform bacterial pathogenicity, antibiotic resistance, andthe utilization of unusual energy resources. In contrast, studies focused on how regulatory plasticity affects bacterial evolution, are mostly overlooked. For example, in Photorhobdus, a single promoter switch changed the organism from acommensal to a pathogen. My current research focuses on identifying the role of intergenic space in horizontal gene transfer of the Caulobacter species. They are found varied habitats including fresh and saltwater systems soil and root systems.

Due to its unique lifestyle and well-established system for genetic analysis, Caulobacter is important model organism for studying cell cycle regulation, asymmetric cell division, and cellular differentiation. The gene order of closely related species of bacteria is usually conserved. But the Caulobacter genus has a higher magnitude of genome scrambling than what is seen in most other bacterial genera. Genome reorganizations are commonly due to mutations and horizontal flow of genes. The latter contributes to rearrangements through recombination where foreign genetic material is incorporated into the genome. My research has shown that most of the HGT events in this genus occur via recombination in the intergenic DNA spaces. The evolution of bacteria can be

a rapid process. They achieve this by local point mutations through insertions, deletion, and substitutions and by horizontal gene transfer, recombination and genome rearrangements. There is an interesting degree of duality that a bacterialgenome is exposed to constantly: To maintain a constant tradeoff between genome evolvability and genome maintenance through robustness. Recently, Dr. Ely published a new model for recombination in genus,Caulobacter with simultaneous loss and gain of genes resulting from preferential recombination at non-homologous regions flanked by regions of homology. In my dissertation, I observed and catalogued in hairpin structures at known sites of recombination closely and distantly related species to Caulobacter crescentus strain NA1000. I also observed that stem structures tend to remain consistent across species with ‘some changes in either the length or bases due to differences in sequence conservation due to phylogenetic distance. Therefore, the presence of these hairpin structures seems to have been conserved at sites of recombination suggesting that they may play role in initiating recombination. It is interesting to note that despite variable intergenic sequence conservation with increase phylogenetic distance, there are conserved stem bases present across different species of genus Caulobacter. Is the stability of hairpin stem a consequence of ‘equivalence classes’ of higher- level evolutionary selection at those intergenic DNA spaces? Or is it a constraint of genome organization as a function of external environment leading to regulatory conservation of these hairpin structures? Evolutionary dynamic interactions of each of these genome variations in the intergenic DNA landscape through experimental exploration will help us understand the topology of the regulatory plasticity at both interspecific and intraspecific time scales to elucidate their functional capabilities.

REFERENCES

1. Darmon, E., Leach, D.R. Bacterial genome instability. Microbiol Mol Biol Rev. 2014;78(1):1-39.
2. Noureen, M., Tada, I., Kawashima, T. et al. Rearrangement analysis of multiple bacterial genomes. BMC Bioinformatics 20, 631 (2019). https://doi.org/10.1186/s12859-019-3293-4
3. Ely B, Wilson K, Ross K, et al. Genome comparisons of wild isolates of Caulobacter crescentus reveal rates of inversion and horizontal gene transfer. Curr Microbiol. 2019;76(2):159-167.
4. Patel S. Drivers of bacterial genomes plasticity and roles they play in pathogenvirulence, persistence and drug resistance. Infection, Genetics and Evolution. 2016;45:151-164.
5. Ely B. Recombination and gene loss occur simultaneously during bacterialhorizontal gene transfer. PLoS ONE. 2020;15(1):e0227987.
6. Bikard D, Loot C, Baharoglu Z, Mazel D. Folded dna in action: hairpin formationand biological functions in prokaryotes. Microbiol Mol Biol Rev. 2010;74(4):570- 588.
7. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Carver T, Harris SR,
8. D. ARNold: a web tool for the prediction of Rho-independent transcription terminators. RNA Biol. 2011;8(1):11-3.
9. Di Salvo, M., Puccio, S., Peano, C. et al. RhoTermPredict: an algorithm for predicting Rho-dependent transcription based on Escherichiacoli, Bacillus subtilis and Salmonella enterica databases. BMCBioinformatics 20, 117 (2019).
10. Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat. Commun. 2019;10: 2182. DOI: 10.1038/s41467-019-10210-

3

11. Lefort V, Desper R, Gascuel O. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evol. 2015;32: 2798–2800.
12. BLAST local alignmentsearch tool. J Mol Biol. 1990;215(3):403-10.
13. Daxin Jiang, Chun Tang and Aidong Zhang, "Cluster analysis for geneexpression data: a survey," in IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1370-1386, Nov. 2004. doi: 10.1109/TKDE.2004.68
14. R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis 5th ed. (Prentice Hall, New Jersey, 2001).
15. Ely B. Recombination and gene loss occur simultaneously during bacterial horizontal gene transfer. PLoS ONE. 2020;15(1):e0227987.
16. Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011).Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539doi:10.1038/msb.2011.
17. Christian Hennig (2019).fpc: Flexible Procedures for Clustering. R package version 2.2-2. https://CRAN.R-project.org/package=fpc