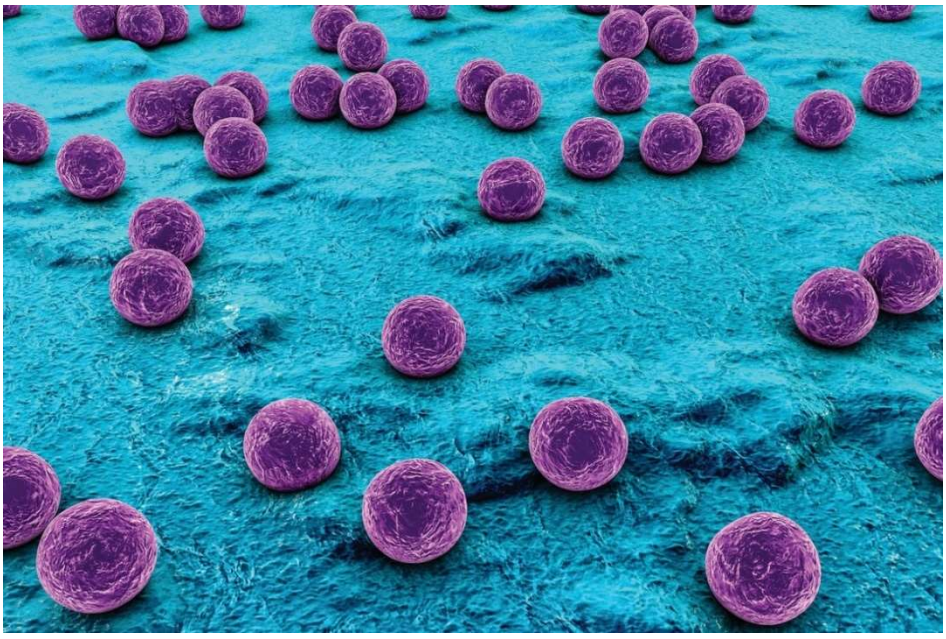
ASSIGNMENT 2

DETERMINATION OF PATHOGENICITY ISLANDS IN *STAPHYLOCOCCUS AUREUS*

Instituto Politécnico de Setúbal – Escola Superior de Tecnologia do Barreiro

Laboratórios em Bioinformática



Licenciatura em Bioinformática

Francisco Pina Martins

Alberto Júnior

Janeiro 2023

Alexandre Duarte n. °202000198

Diogo Cabrita n. °202000212

Guilherme Sá n. °202000201

Index

INDEX OF FIGURES

INDEX OF TABLES

Introduction

Staphylococcus aureus (*Staphylococcus Aureus Infections - Infections*, n.d.) is one of the most dangerous bacteria of all, among the most common *staphylococcal* bacteria. These bacteria are gram-positive and they have a spherical form that's where the name comes, from being coccus-shaped bacteria, and it´s known because of the many causes in skin infections, heart infections, and bone infections and it can also cause pneumonia in humans and animals.

S. aureus is able to cause such a wide range of diseases due in part to the presence of pathogenicity islands in its genome, which encode virulence factors, toxins, and other molecules that can enable this bacterium to colonize and infect the host tissues and other organs in the way to evade the host system.

This bacterium is frequently found on the skin and on the mucous membranes (*Mucous Membrane - MeSH - NCBI*, n.d.) of healthy individuals, and it causes infections when its able to enter the body of individuals through simple bruises in the skin and when it's transmitted from animal to animal.

These bacteria are very resistant to many antibiotics, and it able to create resistance to new antibiotics through the acquisition of antibiotic resistance genes, this method is created through horizontal gene transfer (Burmeister, 2015), making the treatment of infections caused by *S. aureus* increasingly difficult and challenging to all the ones individuals that got infected.

To prevent and treat infections caused by *S. aureus*, it is important to practice good hygiene and to use antibiotics responsibly because when the treatment it done on a high dose routine the host system becomes more vulnerable to these bacteria to evolve, and create resistance in some way, to the antibiotics. In addition, researchers are working to develop new antibiotics and other therapies that can effectively target this pathogen in way to develop a treatment that can cure the infected individuals. (*ChatGPT*, 2022)

That's where the Pathogenicity islands come to work, because of the help that they can bring on the context to be used, on develop new methods to make ways to treat and prevent from infections like the *S. aureus* that are resistant bacteria, making hard to cure it.

Pathogenicity islands are regions within an organism's genome that are associated with the presence or expression of pathogenicity traits (HAMADA et al., 2015). These regions can include genes or gene clusters (Elizondo et al., 2009) that encode virulence factors, toxins, or other molecules that contribute to the ability of the organism to cause a disease. Pathogenicity islands can also contain regulatory elements (*Gene Regulatory Elements, Major Drivers of Human Disease | Annual Review of Genomics and Human Genetics*, n.d.) that control the expression of these genes, allowing the organism to modulate its pathogenic potential in response to changes in the environment of the host.

Pathogenicity islands are often found in the genomes of bacteria and other microorganisms like, fungi, protozoa, and plants. They can be difficult to identify and characterize, as they are often located in non-coding regions (*What Is Noncoding DNA?*, n.d.) of the genome and may have evolved to evade detection by host immune systems. However, understanding the role of pathogenicity islands in the way they cause disease on organisms, can help to inform

the development of new strategies for preventing and treating infections, as mentioned above.

These islands are genetic regions found in certain bacteria, including *Staphylococcus aureus*, that contain virulence factors, or genes that contribute to the bacteria's ability to cause disease. These islands are often located on mobile genetic elements such as plasmids (*Plasmid*, 2022) or transposons (*Transposons | Learn Science at Scitable*, n.d.), which can be easily transferred between different strains of bacteria.

Methods for determining the presence of pathogenicity islands in *S. aureus* include PCR amplification of known virulence genes (Thomas & Wigneshweraraj, 2015), comparative genomics (Sivashankari & Shanmughavel, 2007) to identify regions of the genome that are unique to pathogenicity strains, and bioinformatics analysis of genome sequences.

This are the 3 most common ways that are used to identify this region:

- **PCR** is a technique used to amplify specific regions of DNA, and it can be used to detect the presence of virulence genes known to be located on pathogenicity islands.

- **Comparative genomics** is the study of the genetic content and organization of different strains of *S. aureus*, and it can be used to identify regions of the genome that are unique to pathogenicity strains.

- **Bioinformatics analysis** of genome sequences can also be used to identify regions of the genome that are unique to pathogenicity strains, and to identify potential virulence genes located on these islands.

Some strains of *S. aureus* are considered less virulent and do not contain pathogenicity islands. Identifying the presence of pathogenicity islands in *S. aureus* is important for understanding the mechanisms of virulence and for developing strategies to control the spread of pathogenicity strains. In this way, the bacteria can be easily transferred between different strains, allowing for the rapid spread of virulence factors among *S. aureus* populations. It is important to note that pathogenicity islands are not present in all S. aureus strains.

Now that we approached about our topics, the main objective is to create a paper based on a problem that involves our bacteria. This project was based on one of the options provided by the professors. Which was tilted as "Determination of pathogenicity islands in *Staphylococcus aureus*".

Objectives

The goal of this project is to mainly determinate PAIs of *Staphylococcus aureus* by using analyses and various tools to find all the topics and characteristics of PAIs.

Mentioned before the PAI´s are genetic regions found in certain bacteria that contribute to the bacteria's ability to cause diseases and the detection of this pathogenicity islands is important for understanding the mechanisms of virulence facts and for developing strategies to control the spread of pathogenic strains of bacteria, to determinate the presence and characteristics of pathogenicity islands in *S. aureus* we are going to use the methods Comparative genomics, PCR and Bioinformatics analysis.

Using these methods, it´s possible to identify and characterize in each genome, the amount of potential the variables to be a PAI candidate in *S. aureus* and to understand the role of this regions, the impact on virulence spread and the factors that they can contribute.

Materials & Methods

For this, we selected eight sequences (strains) and the reference genome used was the strain CR14-035. The principal software that we used was GIPSy.

For a more accurate analysis, we used other programs/databases like Phaster and Patric to reach a more concrete conclusion about the islands. The materials are divided in 4 topics.

1. Data Extraction

Extracting sequences from the NCBI database is a common task in the field of bioinformatics. In order to automate the process, we used Python v3.7.12 as our programming language to write the scripts, which offers various libraries, such as Biopython

This library provides functions for accessing, parsing and manipulating biological data, including sequences in the GenBank format.

In order to do that we had to give/retrieve the GenBank records, using the Entrez module of Biopython, which provides access to the NCBI's Entrez databases, including GenBank.

2. Phage analysis

Phages are ubiquitous elements in the bacterial world and play a significant role in shaping the genetic and functional diversity of their hosts.

To analyse them, we used Phaster's (PHage Search Tool Enhanced by Reciprocal best BLAST hit) online API (*PHASTER*, n.d.), which is a highly efficient and reliable computational tool that is used to identify and annotate prophages and phage remnants in bacterial genomes.

The algorithm leverages the rapid and sensitive comparison of the input genomes to a comprehensive database of known phages via the BLASTN algorithm.

We ran Phaster with the accession number of each strain but you can also run it with the sequence file (genbank or fasta) or pasting a nucleotide sequence (raw or FASTA format).

3. Functional comparison

In this method we compared the proteins present in different bacterial genomes from the 7 strains with the reference genome that we selected. With this comparison it's possible to identify differences in the functional capabilities of these genomes.

In order to compare them we used Patric database or Bacterial Virulence and Antimicrobial Resistance (BV-BRC) (*Bacterial and Viral Bioinformatics Resource Center | BV-BRC*, n.d.), which

is an information system designed to support research on bacterial and viral infectious diseases that ends up being a valuable resource for the analysis of pathogenicity islands (PAIs) in bacterial genomes.

With this system we performed a proteome comparison (*Proteome Comparison Service | BV-BRC*, n.d.) that does a protein sequence-based genome comparison using bidirectional BLASTP.

4. Pathogenicity islands

The main purpose of this article is to identify pathogenicity islands in our strains, for that we used **GIPSy** (Genome Island Prediction System) (Soares, n.d.) that is a computational tool designed to identify **pathogenicity islands** (PAIs) in bacterial genomes.

By determining them, we can gain a deeper understanding of the genetic and functional factors that contribute to the pathogenicity of bacteria and the evolution of pathogenic genomes. (*ChatGPT*, 2022)

GIPSy has 8 steps and the final one it's where we get the respective PAIs from the strains that we selected.

Step 1: Input of the genome files

Here is where you provide the files as input for the query (strain) and subject (reference) genomes to create the additional files. We used genbank files, but you can also use EMBL's.

Step 2: G+C content analyses

Often, PAIs are associated with the G+C deviation, as they are usually acquired from different sources. Detecting regions of the genome with a significantly different G+C content than the rest of the genome may indicate possible PAIs. We used the standard value which is "1.5". (Soares et al., 2016)

Step 3: Codon usage analyses

Colombo/SIGIHMM (Signature Integrated Prediction of Islands by Hidden Markov Models) calculates the codon usage deviation. It's a sensitive tool, that GIPSy uses for the identification of GEIs (genomic islands) in microbial genomes and to the prediction of PAIs in bacterial genomes. (Waack et al., 2006)

You can set the sensitivity parameter, from 0.5 to 0.95, we decided to use the standard value, which is the highest.

Step 4: Transposase genes

The prediction of these genes is performed using the software HMMER3. It searches in the ".faa" file of the query genome (generated in the first step) for hidden patterns using a transposase database. Setting the E-value we used the standard value "0.0001". (Soares et al., 2016)

Step 5: Search for virulence factors

Using blastp algorithm GIPSy can predict many factors being one of them the virulence factors.

We want to identify the PAIs, so we searched for these factors with the E-value given as default (0.000001). If you are searching for Resistance Islands, run the analyses for antibiotic resistance genes, and so on.

Step 6: Reciprocal blast

Here GIPSy identifies similarities and differences in gene content between the strain and the reference genomes (query and subject).

This a crucial step in the investigation of the genomic basis of virulence in bacterial pathogens, as it provides valuable information for the identification of regions in the query genome that are associated with pathogenicity and virulence (*ChatGPT*, 2022). We used the standard E-value (0.000001).

Step 7: tRNA genes

Here, GIPSY uses HMMer program to predict tRNA genes in the query genome. This software serves to identify homologous protein or nucleotide sequences, and to perform sequence alignments ('HMMER', 2022). The E-value used was the default (0.0001).

Step 8: Island prediction

Last but not least, the last step where all the data from the previous steps is used to predict the chosen genomic islands of our organism *(S. aureus*).

By choosing the "Pathogenicity Islands" option we can identify, not only the PAIs and the GEIs (genomic islands), but also the score of each one of them as well in order to understand some details like, which island has bigger virulence.

The scripts that were created for the analysis above are available in this GitLab repository.

Results

It is intended to obtain and carry out a good analysis and determine the PAIs of *S. aureus* and compare them to different regions of the genome, to try and find, looking for patterns and functions that are associated with this island.

With several searches were found some databases with information that can compare with the results obtained from other programs.

1. G+C Deviation

G+C deviation, also known as GC skew, is a measure of the difference between the observed G+C content (GC%), and the expected G+C content (GC%), in a given region of a genome. The G+C content of a genome is the percentage of nucleotides that are either guanine (G), or cytosine (C) in the genome.The expected G+C content is based on the overall G+C content of the genome.

GC skew is often used as a measure of compositional asymmetry in the genome and it can be calculated using the following formula:

$$\textbf{GC skew} = (G-C) / (G+C)$$

GC skew is often used as a feature in the prediction of genomic islands, as genomic islands tend to have a different G+C content than the rest of the genome, this difference in G+C content is due to the fact that many genomic islands are acquired through horizontal gene transfer and, therefore not subject to the same selective pressures as the rest of the genome. GC Skew have features to predict genomic islands, It will identify regions of the genome where the GC skew deviates significantly from the overall GC skew of the genome, these regions are considered as possible genomic islands.It is worth noting that GC skew is not always a reliable indicator of genomic islands, as some genomes have naturally high GC skew.

2. Codon usage deviation

Codon usage deviation is another feature that is often used in the prediction of genomic islands. Codon usage deviation refers to the difference between the observed usage of a particular codon, and its expected usage based on the overall codon usage of the genome.

Codon usage bias refers to the non-random usage of synonymous codons in a genome and it is influenced by factors such as GC content. A deviation in codon usage bias is a phenomenon that occurs when a particular codon is used more or less frequently than expected, based on the overall codon usage of the genome.

A genomic island is expected to have different codon usage bias than the rest of the genome, as it is often acquired through horizontal gene transfer, being often not subject to the same selective pressures as the rest of the genome. Codon usage deviation as a feature for prediction of genomic islands, it will identify regions of the genome where the codon usage, deviates significantly from the overall codon usage of the genome and in this way, these regions are considered as possible genomic islands but the Codon usage bias can vary among genomes, and some genomes have naturally high codon usage deviation.

3. Virulance Factors

Virulence factors are molecules or genetic elements that allow a pathogen to cause disease in a host, this can include enzymes, toxins, and adhesins. Virulence factors are often encoded by genes that are located on mobile genetic elements such as plasmids, transposons, or genomic islands. These mobile genetic elements can be easily transferred between bacteria, allowing them to spread virulence factors to other strains.

Some software that predict genomic islands, also look for the presence of known virulence factors within the genomic islands, using databases such as VFDB (Virulence Factor Database), which contains information about virulence factors from a wide range of bacterial pathogens, that contain information about virulence factors.

When identifies a genomic island that contains known virulence factors, it is considered as a strong indication that the island is a pathogenicity island, which is a genomic island that contains genes that are important for the pathogenicity of the organism like the other methods, the presence of virulence factors in a genomic island does not necessarily mean that the island is responsible for virulence.

4. Hypothetical proteins

Hypothetical proteins are proteins that have no known function, or are not annotated in databases such as UniProt or GenBank. These proteins are predicted based on the sequence information obtained, from genome sequencing.

In the context, it was used GYPSy program, hypothetical proteins may and will be, identified within the genome of the organism being studied but again it is important to note that the prediction of hypothetical proteins is not a conclusive indication of their actual existence or function.

Despite their unknown function, hypothetical proteins are valuable targets for further study as they have the potential to uncover biological processes, or contribute to our understanding of cellular function.

5. Hypothetical proteins

Prediction score is a value or a category assigned by a genomic island prediction software to indicate the confidence level of the prediction. The score is based on the combination of different features that are used in the prediction, such as G+C deviation, codon usage deviation, and presence of known virulence factors or hypothetical proteins. The prediction score can be represented as a numerical value or as a category such as "strong", "normal", "weak" or "not available (NA)".

The meaning of these categories can vary depending on the software used but on this case:

A "strong" prediction score indicates that the software is highly confident that the region is a genomic island. This is usually assigned to regions that have a high G+C deviation, a high codon usage deviation and contain known virulence factors or hypothetical proteins.

A "normal" prediction score indicates that the software is moderately confident that the region is a genomic island, assigned to regions that have a moderate G+C deviation, and a moderate codon usage deviation.

A "weak" prediction score indicates that the software is less confident that the region is a genomic island, assigned to regions that have a low G+C deviation and a low codon usage deviation.

A "not available (NA)" prediction score indicates that the software was not able to make a prediction for that region. This can happen when the region is too small, or when the region is not annotated. The results of the prediction scores are not absolute, and they should be interpreted with some caution, this can be affected by the quality of the input data and the parameters used in the software.

In this part it was created table´s for each genome so it could be identified the amount of potencial of this variables to be candidates of beeing PAIs.

In the next part, it was created a table for each genome for a easier view of the variables and understanding, in way to, analyse the differences of the values on each pathogenicity island.

	GENOME	ISLAND 1	ISLAND 2
G+C	11%	31%	27%
Codon Usage Deviation	6%	68%	86%
Virulence Factors	35%	37%	58%
Hypothetical proteins	19%	31%	27%
Prediction Score	N/A	Strong	Strong

Table 1 - PAIs results of strain GV51

	GENOME	ISLAND 2	ISLAND 3
G+C	12%	15%	28%
Codon Usage Deviation	6%	92%	84%
Virulence Factors	35%	38%	60%
Hypothetical proteins	17%	46%	40%
Prediction Score	N/A	Strong	Strong

Table 2 - PAIs results of strain TW20

	GENOME	ISLAND 3	ISLAND 4
G+C	12%	24%	26%
Codon Usage Deviation	6%	86%	13%
Virulence Factors	35%	58%	73%
Hypothetical proteins	26%	55%	78%
Prediction Score	N/A	Strong	Strong

Table 3 - PAIs results of strain Z172

	GENOME	ISLAND 1
G+C	12%	23%
Codon Usage Deviation	0%	0%
Virulence Factors	36%	88%
Hypothetical proteins	24%	84%
Prediction Score	N/A	Strong

Table 4 - PAIs results of strain KG-03

	GENOME	ISLAND 1	ISLAND 3	ISLAND 7
G+C	12%	9%	32%	42%
Codon Usage Deviation	6%	100%	84%	57%
Virulence Factors	35%	45%	60%	57%
Hypothetical proteins	20%	45%	32%	28%
Prediction Score	N/A	Strong	Strong	Strong

Table 5 - PAIs results of strain Be62

	GENOME	ISLAND 1	ISLAND 2
G+C	12%	23%	9%
Codon Usage Deviation	0%	0%	0%
Virulence Factors	32%	34%	45%
Hypothetical proteins	22%	40%	36%
Prediction Score	N/A	Weak	Weak

Table 6 - PAIs results of strain M92

	GENOME	ISLAND 1	ISLAND 3	ISLAND 8	ISLAND 5
G+C	11%	29%	31%	42%	16%
Codon Usage Deviation	6%	52%	86%	57%	2%
Virulence Factors	35%	26%	58%	57%	36%
Hypothetical proteins	20%	58%	31%	28%	30%
Prediction Score	N/A	Strong	Strong	Strong	Weak

Table 7 - PAIs results of strain Gv88

PHASTER

In this phage analysis we selected the strain M92 that has the most intact regions, that is, has more intact content and the odds of having phages in those regions are very high.

In table 8 it shows some details like the number of the region, length, position, if it's intact, incomplete or questionable and some specific phage-related keyword(s) found in protein name(s) in the region, that is, in sum it's the region's function(s) in the genome.

Table 9 it's more about how many proteins these regions have and how are they distributed in phage, hypothetical and bacterial proteins.

Region	Length	Completeness(score)	Position	Keywords
1	21.2Kb	Intact(95)	909119-930364	NA
2	8.3Kb	Incomplete(50)	1355091-1363443	Head, Transposase
3	5.8Kb	Incomplete(50)	1902950-1908835	Transposase, Portal
4	28.6Kb	Incomplete(40)	1909761-1938418	Protease, Integrase
5	63.3Kb	Intact(150)	1989214-2052583	Lysin, Tail, Head, Capsid, Portal, Terminase, Integrase
6	51.4Kb	Intact(130)	2101483-2152901	Transposase, Lysin, Tail, Head, Capsid, Portal, Terminase, Integrase
7	129.2Kb	Intact(125)	2178385-2307664	Recombinase, Transposase, Lysis
8	10.5Kb	Incomplete(40)	3023199-3033792	Transposase, Tail

Table 8 - Region length, completeness, position and specific keywords

Region	Total Proteins	Phage P.	Hypothetical P.	Pha. + Hypo. %	Bacterial P.
1	23	19	1	86.9%	3
2	20	9	10	95%	1
3	14	5	6	78.5%	3
4	27	14	5	70.3%	8
5	67	67	0	100%	0
6	80	74	3	96.2%	3
7	149	133	6	93.2%	10
8	13	7	2	69.2%	4

Table 9 - Total proteins and types of each region

The next graphic (figure 1) represents all the genome of our strain, and each region marked as intact (green) and incomplete (red).

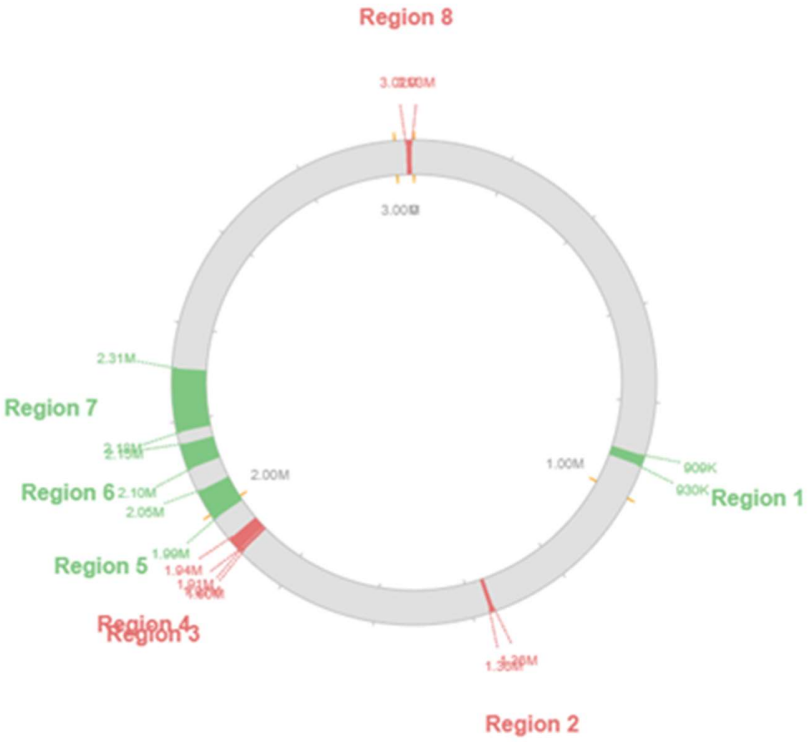


Figure 1 - Circular graphic of the regions in Staphylococcus aureus strain M92 chromosome

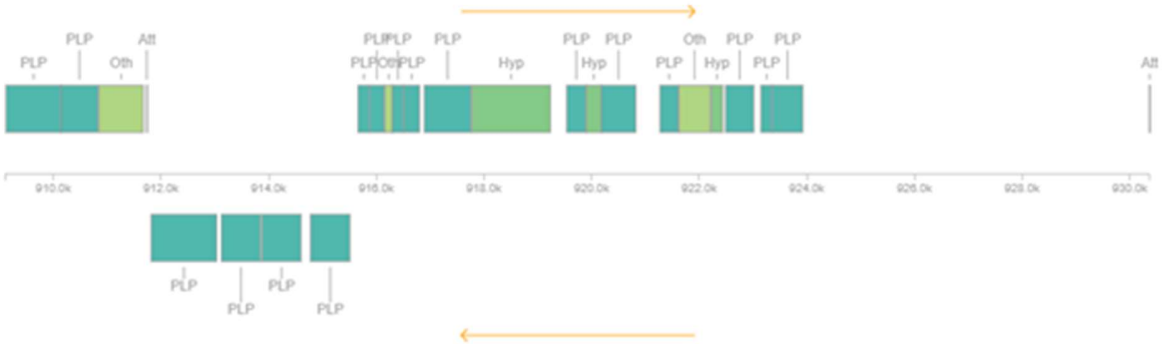


Figure 2 - Proteins in region 1 of Staphylococcus aureus strain M92 chromosome

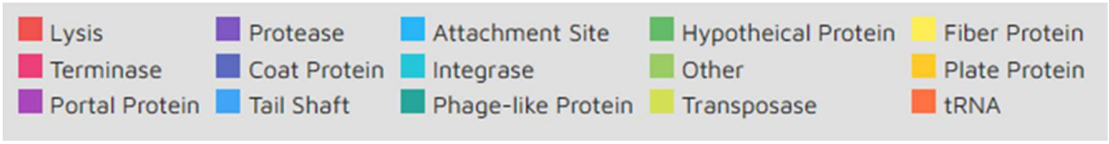


Figure 3 - List of protein types identified by phaster

With the figures 2 and 3 we can see how many proteins are in the region 1 and what are they defined as.

PATRIC(BV-BRC)

The graphic below it's the final output of the proteome comparison that we did in BV-BRC (*Proteome Comparison Service | BV-BRC*, n.d.). It shows the comparison of all the 8 strains selected and the CR14-035 as the reference genome.

The "Percent protein sequence identity" refers to the similarity of proteins, rated by colors.

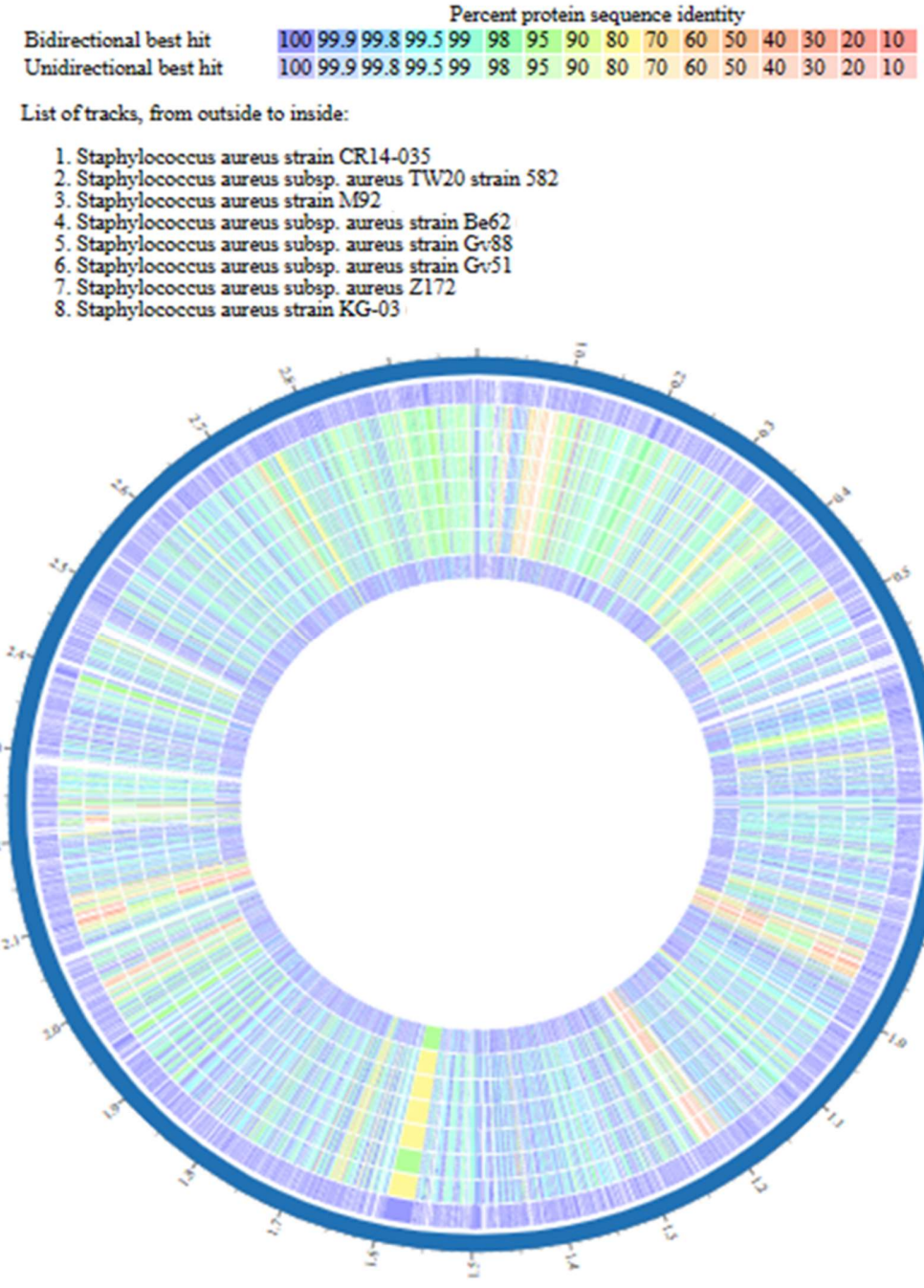


Figure 4 - Proteome comparison´s final output

Discussion

In the discussion of the results, we aim to consider the PAIs that have strong indices, or can be applied for the PAIs. Normal islands are not considered in the study because they present standard values, in almost all of them.

Candidates for PAIs with similar values are only considered as a case study, the comparison only takes place on islands with different variations. And of course analyse the results of the Phaster and Patric.

So, in this way we are going to do analysis of each variable in each genome, through the tables developed in the earlier topic, being able to carry out and determinate possible PAIs. Firstly, let's start discussing the GIPSy results.

GIPSy

Gv51 (Table 1):

- It is possible to complete that in both islands the GC skew value, increased significantly, which can maybe indicate that these regions are strong candidates for pathogenicity islands.

- The "Codon Usage Deviation" uses the same method as the GC skew, however this one has a special focus on non-random zones "synonymous codons" in a genome. Factors such as GC content, tRNA abundance and translation efficiency, in both islands these values are much higher than the initial ones but island n°2 has a higher value.

- "Virulence factors" uses databases such as VFDB (Virulence Factor Database), which this database, contains information about virulence factors from a wide range of bacterial pathogens, or other similar databases that contain information about virulence factors. Once detected, they become potential elements to be considered PAIs. In this case, island n°2 has a higher value, meaning that can be a greater capacity of adaptation, in phase to the immune system.

- About "Hypothetical proteins" in this case both islands have high values in comparison to the genome so this can mean that they are evidence of hypothetical proteins.

In a general balance, the island n°2 has a Codon usage and a virulence factor higher then island n°1 but overall, it is possible to understand that both islands present results are very similar in each variable, this possibly means, that both of this islands have a great potential to infect and adapt on the host, but the island n°2 possibly has a higher chance of being a pathogenicity island. About the prediction score all of the island presents results of strong candidate to PAIs.

TW20 (Table 2):

- With the results obtained, it´s possible to identify that the island n°2 does not present a great variation of GC skew, when compared to the genome, however, in turn island n°3 presents a high variation.

- The "Codon Usage Deviation", in both islands have a large and constant growth, which indicates that these islands, are present possible functional zones, not yet discovered in terms of their functionality.

- The "Virulence factors", on island n°2 show a small variation, in phase of the great growth in island n°3, this means that island n°3 has a greater capacity of adaptation, phase to the immune system of the host, that is, the capacity of inflammation cells and the ability of the host to infect.

- About "Hypothetical proteins", similar to the before, both islands have very high values in comparison to the genome, in such way that they are evidence of hypothetical proteins, which can claim the formation of additional proteins.

In a general balance, island n°3 presents results of virulent factors superior to island n°2, which may mean that island n°3 has a greater capacity for infection and adaptation, in phase to island n°2, besides both islands are very similar. About the prediction score all of these islands, present results of strong candidate to PAIs.

Z172 (Table 3):

- With the results obtained, it´s possible to identify that the island n°3 does not present a great variation of GC skew, when compared to the genome, however, in turn island n°4 presents a high variation.

- The "Codon Usage Deviation", on island n°4 show a small variation and in phase of the great growth in island n°3, which indicates that these islands, are present possible functional zones, not yet discovered in terms of their functionality.

- The "Virulence factors", on both islands got a growth in overall however still bigger on island n°4, this means that island n°4 has a greater capacity of adaptation, phase to the immune system of the host, that is, the capacity of inflammation cells and the ability of the host to infect.

- About "Hypothetical proteins on both islands got a growth in overall however still bigger on island n°4 in comparison to the genome, in such way that they are evidence of hypothetical proteins, which can claim the formation of additional proteins.

In a general balance, island n°4 presents results of virulent factors superior to island n°3, which may mean that island n°4 has a greater capacity for infection and adaptation, in phase to island n°3, besides both islands are very similar. About the prediction score all of these islands, present results of strong candidate to PAIs.

KG-03 (Table 4):

- With the results obtained, it´s possible to identify that the island n°1 does not present a great variation of GC skew, when compared to the genome.
- The “Codon Usage Deviation” don’t have any difference compared with genome.
- The "Virulence factors”, this island doubles the value of virulence, this means, this phase to the immune system of the host, that is, the capacity of inflammation cells and the ability of the host to infect.
- About "Hypothetical proteins", in this category got very high value in comparison to the genome, in such way that they are evidence of hypothetical proteins, which can claim the formation of additional proteins.

In a general balance, island n°1 presents high results of virulent factors, which may mean probably had a greater capacity for infection and adaptation. About the prediction score all of these island, present results of strong candidate to PAIs.

Be62 (Table 5):

- With the results obtained, it´s possible to identify that the island n°1 does not present a great variation of GC skew, when compared to the genome, however, in turn island n°3 and 7 both presents a high variation.
- The "Codon Usage Deviation", in both islands have a large and constant growth, which indicates that these islands, are present possible functional zones, not yet discovered in terms of their functionality.
- The "Virulence factors", on island n°2 show a small variation, in phase of the great growth in island n°3, this means that island n°3 has a greater capacity of adaptation, phase to the immune system of the host, that is, the capacity of inflammation cells and the ability of the host to infect.
- About "Hypothetical proteins", both islands have very high values in comparison to the genome, in such way that they are evidence of hypothetical proteins, which can claim the formation of additional proteins.

In a general balance, island n°3 presents results of virulent factors superior to island n°2, which may mean that island n°3 has a greater capacity for infection and adaptation, in phase to island n°2, besides both islands are very similar. About the prediction score all of the island, present results of strong candidate to PAIs.

M92 (Table 6):

- With the results obtained, it´s possible to identify that the island n°2 does not present a great variation of GC skew, when compared to the genome, however, in turn island n°1 presents a small growth variation.

- The “Codon Usage Deviation” don’t have any difference compared with genome.

- The "Virulence factors", on both islands got a growth in overall however still small growth, phase to the immune system of the host, that is, the capacity of inflammation cells and the ability of the host to infect.

- About "Hypothetical proteins on both islands got a growth in overall however still small in comparison to the genome, in such way that they are evidence of hypothetical proteins, which can claim the formation of additional proteins.

In a general balance, island n°2 presents results of virulent factors superior to island n°1, which may mean that island n°2 has a greater capacity for infection and adaptation, in phase to island n°1, besides both islands are very similar. About the prediction score, island n°1 and n°2 present results of weak candidate to PAIs.

Gv88 (Table 7):

- With the results obtained, it´s possible to identify that the island n°1 and 5 does not present a great variation of GC skew otherwise, island n°3 and 8 have a big growth when compared to the genome.

- The “Codon Usage Deviation” don’t have any difference compared with genome for island n°2 and a big growth for the rest of islands.

- The "Virulence factors", all islands got a big growth expect the number 5, phase to the immune system of the host, that is, the capacity of inflammation cells and the ability of the host to infect.

- About "Hypothetical proteins on overall islands got a growth, however still small in comparison to the genome, in such way that they are evidence of hypothetical proteins, which can claim the formation of additional proteins.

In a general balance, island n°5 presents results of weak candidate to be considered a pathogenicity island, which may mean the rest of the islands have a stronger possibility or indicators to be a PAI.

PHASTER

This genome has 32.81% GC content, that represents the highly structured, and functional regions of this strain.

If the region's total score greater than 90 Phaster marks it as intact, if between 70 to 90, it is marked as questionable, and if is less than 70, it is marked as incomplete.
Region 5 it's the only one that all the proteins are phage hit proteins, so it has a 100% in the "Pha. + Hypo. %" column (Table 9).

Analysing the other selected strains, we discovered that the fact that the first region of this genome (M92) not having any keywords it's because there are pathogenicity island proteins, which for example, we found in the second region of TW20's genome too.

From our analysis the last two proteins that are identified as OTH (Other) (figure 2 and 3), are the pathogenicity islands proteins, which were mentioned before.

Patric (BV-BRC)

From the circular graphic (Figure 4), we can say that the first and last rings, which are the reference genome (CR14-035) and the strain KG-03, have approximately 100% identical in all their genomes.

We decided to analyse 2 different separated regions. The one that is in between of the 0.9 mark and 1.0 with a green block, and the other one it's near to the 1.6 mark (the big yellow/green block) (Figure 4).

Fun fact about the strains Gv88 and Be62 is that, in this region they have the same proteins and their length too. Not to mention that, when using phaster, the phage graphics were pretty much the same as well.

The similarity from the other strains and these two are different either it's because, all the other strains have a gap of 1 protein that it's present in the strains Gv88, Be62 and CR14-035, or it's because of the protein's length.
These factors may be the cause that made the similarity of the strains decrease drastically.

The second region that we analysed, we noticed that each one of those blocks are a specific protein called "surface anchored protein".

This means that they are covalently or non-covalently attached to the membrane surface of cells. These proteins play critical roles in various biological processes, including cell adhesion, signalling, and transport of molecules across the membrane.

They often contain hydrophobic regions that interact with the lipid bilayer of the cell membrane, anchoring them in place. Some examples of surface anchored proteins include integrins, transmembrane receptors, and adhesion molecules.

These proteins are important targets for drugs, as they are involved in a variety of diseases including cancer, inflammation, and neurological disorders. (*ChatGPT*, 2022)

Still in figure 4, we can also see that in that specific region two of our strains (M92 and KG-03) have the highest PPSI (Percent Protein Sequence Identity), that is around the 95% of identity, except for the other strains that have ± 80%.

With this project it was possible to understand what PAIs are, and how they can influence a better expression of viruses or bacteria in the immune system of almost every animal in the world, in this case, the amount of virulence factors that PAIs can bring from *S. aureus*.

About the analysis part, it was possible to analyse how these regions could be considered possible PAIs of certain organisms, as well as all related variables or contexts for the process of identification of this islands.

It was also possible to take more conclusions regarding the pathogenicity islands, for instance, values of PAIs that are lower than the genome's, in most cases, are weak candidates. In turn, the higher the PAIs values, the higher possibilities to be a pathogenicity island, meaning that this is a general analysis that is constant in our project.

Contributions

Tasks from each group member:

- Alexandre Duarte: Scripts, análise do Phaster e Patric

- Diogo Cabrita: Estruturação e desenvolvimento do relatório

- Guilherme Sá: Análise GIPSY e referências

References

Bacterial and Viral Bioinformatics Resource Center | BV-BRC. (n.d.). Retrieved 2 February 2023, from https://www.bv-brc.org/

Burmeister, A. R. (2015). Horizontal Gene Transfer. *Evolution, Medicine, and Public Health*, *2015*(1), 193–194. https://doi.org/10.1093/emph/eov018

ChatGPT: Optimizing Language Models for Dialogue. (2022, November 30). OpenAI. https://openai.com/blog/chatgpt/

Elizondo, L. I., Jafar-Nejad, P., Clewing, J. M., & Boerkoel, C. F. (2009). Gene Clusters, Molecular Evolution and Disease: A Speculation. *Current Genomics*, *10*(1), 64–75. https://doi.org/10.2174/138920209787581271

Gene Regulatory Elements, Major Drivers of Human Disease | Annual Review of Genomics and Human Genetics. (n.d.). Retrieved 2 February 2023, from https://www.annualreviews.org/doi/10.1146/annurev-genom-091416-035537

HAMADA, S., KAWABATA, S., & NAKAGAWA, I. (2015). Molecular and genomic characterization of pathogenic traits of group A Streptococcus pyogenes. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, *91*(10), 539–559. https://doi.org/10.2183/pjab.91.539

HMMER. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=HMMER&oldid=1090926305

Mucous Membrane—MeSH - NCBI. (n.d.). Retrieved 2 February 2023, from https://www.ncbi.nlm.nih.gov/mesh?Db=mesh&Cmd=DetailsSearch&Term=%22Mucous+Membrane%22%5BMeSH+Terms%5D

PHASTER. (n.d.). Retrieved 2 February 2023, from https://phaster.ca/

Plasmid. (2022, September 14). Genome.Gov. https://www.genome.gov/genetics-glossary/Plasmid

Proteome Comparison Service | BV-BRC. (n.d.). Retrieved 2 February 2023, from https://www.bv-brc.org/app/SeqComparison

Sivashankari, S., & Shanmughavel, P. (2007). Comparative genomics—A perspective. *Bioinformation*, *1*(9), 376–378.

Soares, S. C. (n.d.). *Index of /download/gipsy*. Retrieved 2 February 2023, from https://www.bioinformatics.org/download/gipsy/

Soares, S. C., Geyik, H., Ramos, R. T. J., de Sá, P. H. C. G., Barbosa, E. G. V., Baumbach, J., Figueiredo, H. C. P., Miyoshi, A., Tauch, A., Silva, A., & Azevedo, V. (2016). GIPSy: Genomic island prediction software. *Journal of Biotechnology*, *232*, 2–11. https://doi.org/10.1016/j.jbiotec.2015.09.008

Staphylococcus aureus Infections—Infections. (n.d.). MSD Manual Consumer Version. Retrieved 2 February 2023, from https://www.msdmanuals.com/en-pt/home/infections/bacterial-infections-gram-positive-bacteria/staphylococcus-aureus-infections

Thomas, M. S., & Wigneshweraraj, S. (2015). Regulation of virulence gene expression. *Virulence*, *5*(8), 832–834. https://doi.org/10.1080/21505594.2014.995573

Transposons | Learn Science at Scitable. (n.d.). Retrieved 2 February 2023, from http://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-518

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., Surovcik, K., Meinicke, P., & Merkl, R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*, *7*, 142. https://doi.org/10.1186/1471-2105-7-142

What is noncoding DNA?: MedlinePlus Genetics. (n.d.). Retrieved 2 February 2023, from https://medlineplus.gov/genetics/understanding/basics/noncodingdna/