Comparative and Regulatory Genomics

Kristen Michelle Nader
Student Number: `r0771801`
Staphylococcus aureus and *Streptococcus pyogenes*

1 Introduction

Staphylococcus aureus is a gram-positive facultative bacterium that has properties of being both commensal and opportunistic in nature.[7]. Infections can range from skin tissue infections to sepsis and endocarditis.[7]. According to the Center of Disease Control and prevention, there are many strains of staphylococcus aureus. The four most popular are listed below[1]:

1. methicillin-resistant Staphylococcus aureus (MRSA) : more difficult to treat due to its developed resistance to common antibiotics
2. methicillin-susceptible Staphylococcus aureus (MSSA)
3. vancomycin-intermediate Staphylococcus aureus (VISA)
4. vancomycin-resistant Staphylococcus aureus (VRSA)

In addition to its ability to develop resistance to antibiotics, it also able to evade the innate immune system through "resistance to antimicrobial peptides (AMPs) and killing by phagocytic leukocytes"[5]!

Streptococcus pyogenes is an aero-tolerant(does not require oxygen and produce energy source through fermentation) Gram-positive coccus with a Lancefield group A antigen and hence is often know as Group A streptococcus(GAS). Infections caused by *S.pyogenes* can take on a wide variaty of clinical manifestations including tissue infections and pharyngitis and even pneumonia and necrotizing fasciitis.[11] Necrotizing fasciitis is known as the flesh eating disease. From this, *S.pyogenes* has been called the flesh-eating bacteria.

Much research, including my own bachelors thesis, is centered on finding similarities between the genomes and proteomes of these two gram-positive bacteria in order to identify targets for therapeutics. In fact, an article published in Nature 2015 on the identification of surface antigens in Streptococcus pyogenes did so to enhance therapeutic methods[10]. This short report aims to find best-bidirectional hits, taking into consideration the potential in-paralogs and the identification of conserved functional regions.

2 Methods

The fasta files for the complete protein was downloaded from NCBI using these websites[[2],[3]] for *Staphylococcus aureus* and *Streptococcus pyogenes* respectively.Before Blast can be run on the high performance cluster, a database of the sequences must be made. In order to determine the presence of in-paralogs, the proteomes were merged into a single fasta file and a database was made for it. The individual proteomes were then "blasted" against the merged database and a python function was developed to determine the presence on in-paralogs. A python script ,attached in the supplementary "best_bidirectional_hits.py", was developed in order to parse the Blastp output and determine Best Bidirectional Hits.

```
makeblastdb -in merged_db.faa -dbtype prot
blastp -query GCF_000006785.2_ASM678v2_protein.faa -db merged_db.faa -out mergedDB_v2query.txt
blastp -query GCF_000013425.1_ASM1342v1_protein.faa -db merged_db.faa -out mergedDB_v1query.
    txt
```

To check orthology using the tree based -operational definition , three sequences were chosen that were previously found to be co-orthologs. These sequences were "blasted" to retrieve 25 homologs from different species

and a multiple sequence alignment was performed using ClustalW[4] on the HPC. The command is provided below.

```
clustalw2 -infile=seqdump.txt
clustalw2 -infile=seqdump.aln -BOOTSTRAP=1000
```

The tree constructed was done using bootstrapping where n=1000. This tree was further visualized using iTol where branch length and bootstrap values can be visualized. A species tree was developed using the 25 species found in ordder to determine speciation and duplication events. Using the same multiple sequence alignment(for the proteins not the 16S rRNA), a python script was developed to identify functional regions using both shannon entropy and simpsons diversity. Pfam and SMART databases were used to understand the domain architecture of the protein of interest.

An attempt towards the identification of regulatory regions was done using phastCons web browser phastWeb and visualzied using the UCSC browser. The input for phastWeb was the phylogenetic tree and mutliple sequence alignment, of the DNA encoding the proteins explored earlier,using ClustalW[4].

3 Results

Figure 1 displays all images accumulated during the workflow. There were 2767 protein coding genes in *Staphylococcus aureus*' genome and 1703 protein coding genes in *Streptococcus pyogenes*' genome.This was done by searching the proteome for the number of ">", an identifier for the beginning of a sequence. Using Best-Bidirectional Hits, 846 orthologs were found. Examples of co-orthologs, taking into account in-paralogs, were extracted and BLAST was used to find homologs in 25 other species. As a result, 26 sequences were used for a multiple sequence alignment using ClustalW[4] with n=1000 bootstrap iterations. The resultant bootstrap tree can be visualized ,using iTol,in figure 1A. Other programs could have been used such as Seaview, that uses Clustal[4] for the multiple sequence alignment and tree construction using bootstrapping, and Figtree for visualization. A species tree was constructed using the 25 species found. 16S rRNA sequences for each species was found using NCBI taxonomy browser. These sequences were aligned and the respective tree was constructed using clustal[4] with n=1000 bootstrap iterations. The protein sequences multiple sequence alignment was used to identify conserved functional groups. The issue faced at this point was that the sequences were more similar to same species and this introduced many gaps in the MSA. The sequences from the same genus were more similar than from different. This clustered sequences from *Staphylococcus* together (having high sequence similarity) and *Streptococcus* together(with high sequence similarity). However, when aligned together, they had relatively poorly conserved regions and only short regions were found which can be visualized from Figure 1D,E. Searching SMART for the protein domain architecture, the domain encoding a Biotin carboxylase C-terminal domain appeared most important from this query. This domain is a component of a multi-component enzyme involved in fatty acid synthesis. Active site residues are generally reported to be within this domain.

4 Discussion

4.1 Compute orthologs using Best Bidirectional Hits

Best Bi-directional hits can be described as an operational definition of orthology. BBH are genes that are most similar to each other in two separate species. However, there are some complications when defining orthologs and this is the presence of possible paralogs. Paralogs are genes that were created by a duplication event within a genome. In this case, the 2 sequences will be in-paralogs where the 2 sequences are co-ortholos to the same sequence in the 2nd species. If a best-bidirectional hit is done, only one sequence will be found to be most similar to the sequence in the other species. In order to account for this, the proteomes are merged and a separate database is constructed. Each proteome is then individually blasted against this merged database. Using this method, e-values can be compared. They could not be compared before because e-values are dependent on the size of the database. A python function was created to search for the in-paralogs and results are saved in a dictionary data structure. In this way, the top hit will be the most similar sequence either in the same species (in-paralogs) or in a different species (orthologs). The dictionaries are then compared for BBH. As a result, some sequences will have a BBH in the same species and others in the other species. There is also the interesting case where sequence 1's best hit is sequence 2 in the same species-paralogs. An important limitation

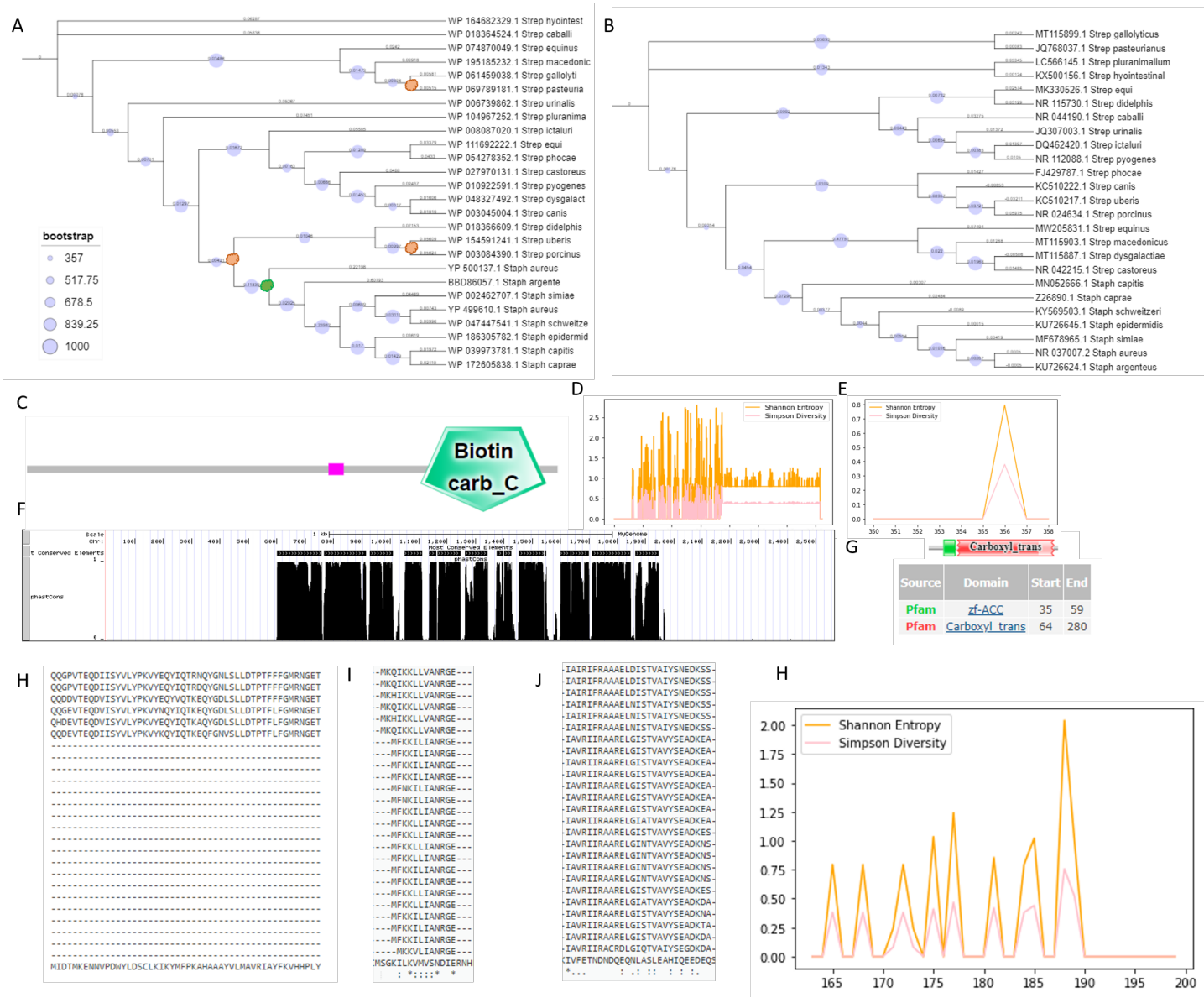

Figure 1: **Workflow and resultant images for all questions including the bonus. (A)**: Visualization of orthologs using iTOl. Bootstrap set to 1000 iterations. Branch length and bootstrap values shown. Green circle indicates a duplication event and orange circles indicate a speciation event. **(B)**: 16S rRNA tree visualized using iTol, bootstrap set to 1000 iterations. Branch length and bootstap values on the tree. Multiple sequence alignment for each tree done using clustalW[4]. **(C)**: Identification of functional regions using SMART database. **(D)**: Shannon entropy(orange) and Simpson diversity(pink) plot for the multiple sequence alignment. **(E)**: Closer look at conserved regions. **(F)** : Output of PhastCons web-server visualized using UCSC browser. **(G)**: Pfam results for Acetyl-coenzyme A carboxylase in *Streptococcus pyogenes*.**(H)**: Multiple sequence alignment displaying aligned of Staphylococcus sequences and gaps introduced in Streptococcus sequences. **(I)**: An example of a short sequence that aligned **(J)**: Example of a large sequence that aligned. **(H)**: Shannon Entropy(orange) and Simpson Diversity(pink) of region highlighted in Image J.

is the BBH can only detect a 1-1 relationship and the relationships between sequence are much more complex. This being said, there are many complications that arise. Firstly and most important in the context of this report are the presence of paralogs in a species. Paralogs are genes/sequences that arose due to a duplication within a genome. However, these gens are then more similar to each other than any other gene/sequence in another species. These sequences will be in-paralogs and both will be orthologous to the BBH. Now, as stated earlier, the concept of a BBH is only able to capture a 1-1 relationship. Therefore, the BBH would only find one sequence and not both in paralogs. There are other complications that arise such as gene loss, multi-domain proteins and horizontal gene transfer which brings in the concepts of xenology(two genes homologous if they have a common ancestor by lateral gene transfer) and pseudoparology(HGT that results in a pseudo-duplication event). In the species assigned, there are and protein coding genes and orthologs.

4.2 Check orthology with a phylogenetic tree. Pick an example of orthologs of co-orthology

The sequences chosen to continue the analysis: YP_500137.1(SA) , YP_499610.1(SA) and WP_010922591.1(SP). These proteins are acetyl-CoA carboxylases . Blast was used to retrieve homologs from 25 different species. These sequences were aligned and the tree was constructed using clustalw and boostrapping. The larger the bootstrap value, the more evidence for that the particular event. There are 3 definitions of orthologs that should be highlighted in this section:

1. Evolutionary definition : genes that descend from the same ancestral gene but separated by a speciation event.
2. Operational definition : Network based : Using cluster of orthologouys groups, observe similarity between gennes and find those that are most similar to eahc other
3. Operational definition: Phylogeny based: rather than finding best hits, reconstruct phylogeny.

Here we explore the concept of orthology using the phylogeny operational definition. This can be done using the species overlap rule or using the species tree reconciliation.

1. Species overlap rule: a node represents a duplication event if the branches have overlapping set of species
2. Species tree reconciliation where both species and gene tree are used to define duplication and speciation events. If the node cannot be mapped from the gene tree to the species tree, it represents a species tree.

Using these methods, we care able to annotate the tree and identify potential speciation and duplication events.Because we are unable to map the node labeled in green to the species tree , by species tree reconciliation, this node represents a duplication event. We can also use the species overlap rule. In this case, there is no need for the species tree and this method aims to deal with an inaccuracy that may arise in the gene tree. Looking at Figure 1A, there is an overlap in species where Staphylococcus aureus appears twice–indicating a duplication event at this node. Other events represent speciation events.

4.3 Identify Functional regions

Due to events of random mutations, elements in a sequence are subject to modifications. Regions that are deemed important for the function of the protein and thus have important functions will be conserved. This can be determined using a multiple sequence alignment where some positions will be more conserved than others. These regions are important due to the fact that they are under selective pressure. Any variations may change the function of the protein or result in inactivity. There are four methods of measuring conservation : Variability, Shannon Entropy, Simpson Diversity and physicochemical properties. A python script was developed to read the multiple sequence alignment and calculate both shannon entropy and simpson diversity. Both measures were used in order to visualize the nice property that simpson diversity measure has- being bound by 0 and 1 allowing a more concrete definition of a variable and conserved region. Regions/positions that are less variable(having only a few different amino acids in that position) are given a lower score-meaning less diversity and more conservation. Figure 1D and E reveal the results of this script. Something that should be pointed out is the difference in the positions that will be aligned as pointed out in the previous section. Therefore , regions where there is a gap is assumed by the script to be similar even though it is indeed a gap and not an amino acid. This being said, the region that shows a "straight line" at the end of figure 1 D is not a conserved region but a large gap. It is not given a score of 0 because there are sequences from the Staphylococcus genus that are aligned very nicely1H. Under further investigation of the sequence alignment, a threshold of 0.5 for the Simpson Diversity was assumed. Positions that scored below 0.5 are assumed to be conserved whereas those above are not. Figures 1 F and H display sites where the simpson score for many positions score below 0.5– around 0.25 and are therefore considered to be conserved. Remark: The x-axis does not indicate the position in the sequence but the position in the multiple sequence alignment. The Pfam and SMART database was investigated to determine regions that were previously determined to be important for the functioning of this protein[[6],[8]]. The sequence of the Biotin carb_C domain is around 336-442aa according to SMART in S.aureus. This corresponds roughly to position 550-650 of the multiple sequence alignment. We see from 2 that the positions scores are below 0.5 and by this standard are considered as conserved. Then, any drastic modifications to this region would change the effect of the protein or the ability of the protein to participate in fatty acid synthesis.

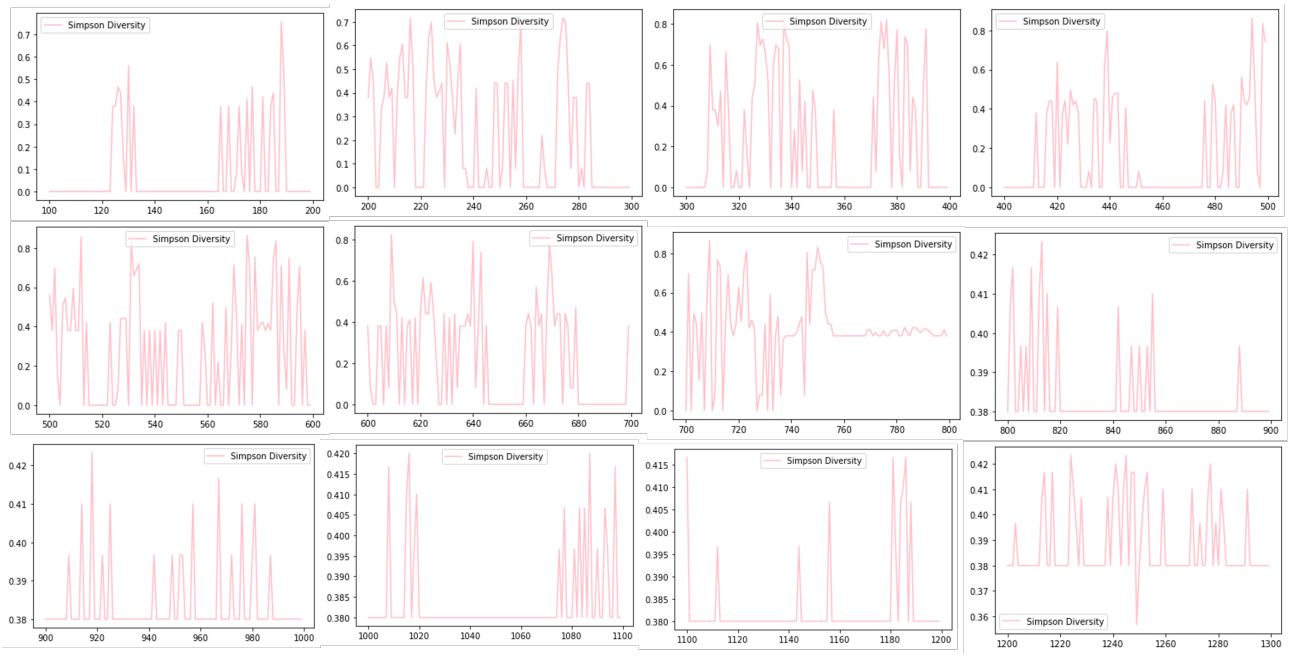

Figure 2: Simpson Diversity plots for intervals of 100

4.4 Conservation of promotor regions

In order to determine regulatory regions, the DNA sequence for each of the proteins was extracted using EnsemblBacteria. These sequences were aligned using CustalW and the resultant alignments was used as input to phastWeb: a web interface of phastCons[9] for determining conserved elements. This tool uses a Hiddden Markov Model to classify segments of the sequence into conserved or non-conserved states. Figure 1F displays the output of phastWeb which represents the conserved regions in the DNA sequences. Black bars represent conserved regions in the DNA. In order to determine conservation in these regions, a multiple sequence alignment can be done.

References

[1] Staphylococcus aureus in healthcare settings-cdc. https://www.cdc.gov/hai/organisms/staph.html.

[2] Streptococcus aureus. https://www.ncbi.nlm.nih.gov/genome/?term=Staphylococcus%20aureus[Organism]&cmd=DetailsSearch. Accessed: 2020-12-1.

[3] Streptococcus pyogenes. https://www.ncbi.nlm.nih.gov/genome?term=gyrA%5BAll%20Fields%5D%20AND%20%22Streptococcus%20pyogenes%22%5BOrganism%5D&cmd=DetailsSearch. Accessed: 2020-12-1.

[4] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research 31*, 13 (07 2003), 3497–3500.

[5] DeLeo, F. R., Diep, B. A., and Otto, M. Host defense and pathogenesis in staphylococcus aureus infections. *Infectious Disease Clinics of North America 23*, 1 (2009), 17 – 34. Staphylococcal Infections.

[6] El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C., and Finn, R. D. The Pfam protein families database in 2019. *Nucleic Acids Research 47*, D1 (10 2018), D427–D432.

[7] Jenkins, A., Diep, B. A., Mai, T. T., Vo, N. H., Warrener, P., Suzich, J., Stover, C. K., and Sellman, B. R. Differential expression and roles of staphylococcus aureus virulence determinants during colonization and disease. *mBio 6*, 1 (2015).

[8] Letunic, I., and Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research 46*, D1 (10 2017), D493–D496.

[9] Ramani, R., Krumholz, K., Huang, Y.-F., and Siepel, A. PhastWeb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastCons and phyloP. *Bioinformatics 35*, 13 (11 2018), 2320–2322.

[10] Reglinski, M., Gierula, M., Lynskey, N. N., Edwards, R. J., and Sriskandan, S. Identification of the Streptococcus pyogenes surface antigens recognised by pooled human immunoglobulin. *Scientific Reports 5*, August (2015), 1–9.

[11] Walker, M. J., Barnett, T. C., McArthur, J. D., Cole, J. N., Gillen, C. M., Henningham, A., Sriprakash, K. S., Sanderson-Smith, M. L., and Nizet, V. Disease manifestations and pathogenic mechanisms of group a streptococcus. *Clinical Microbiology Reviews 27*, 2 (2014), 264–301.