Spring 5-2019

Understanding The Similarity And Diveristy Of The Accessory Gene Regulator Quorum Sensing Systems In The Genus Clostridium

Rotem Magal
UTHealth School of Public Health

UNDERSTANDING THE SIMILARITY AND DIVERISTY OF THE ACCESSORY GENE REGULATOR QUORUM SENSING SYSTEMS IN THE GENUS CLOSTRIDIUM

by

ROTEM MAGAL, BA

APPROVED:

CHARLES DARKOH, PH.D.

MARY ANN SMITH, PH.D.

DEDICATION

I dedicate this thesis to my wife, Anna Blum for all of her love, support and respect. To my parents, sister, and Savta, for bringing me life, showering me with unconditional support, and providing me with the ability to succeed. To my parents-in-law, for your trust in me to be the best I can be for myself and your daughter.

UNDERSTANDING THE SIMILARITY AND DIVERISTY OF THE ACCESSORY GENE REGULATOR QUORUM SENSING SYSTEMS IN THE GENUS CLOSTRIDIUM

by

ROTEM MAGAL
BA, Clark University, 2015

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
May, 2019

ACKNOWLEDGEMENTS

I would like to thank my wife for the constant support and cheer she has given me throughout this process. Dr. Darkoh, thank you for your presence and advice throughout this process. You have challenged me and made me a better version of myself and I cannot thank you enough for that. Dr. Smith, thank you for the insightful feedback and the support. The Darkoh laboratory members have been incredibly supportive and understanding and I will miss them. Joseph Hicks, thank you for always being there to answer my questions regarding phylogenetic trees. Lastly, Dr. Bahl, I appreciate the interest and support in this project.

UNDERSTANDING THE SIMILARITY AND DIVERISTY OF THE ACCESSORY GENE REGULATOR QUORUM SENSING SYSTEMS IN THE GENUS CLOSTRIDIUM

Rotem Magal, BA, MS
The University of Texas
School of Public Health, 2019

Thesis Chair: Charles Darkoh, MS, PH.D.

The Accessory Gene Regulator (Agr) quorum sensing system is a cell-cell communication system that is involved in regulating various bacterial processes such as toxin production, antibiotic production, biofilm formation, and other biomolecules. Despite the importance of the Agr system to Clostridia, the similarity and diversity of the system have been overshadowed by phylum-wide investigations of individual Agr components. To determine the variability of the Agr system within and between Clostridium species, we compared the sequences of its components within and between species using bioinformatics and phylogenetic tools. Putative Agr operons were found in over 50 Clostridia species, including undescribed components in some of the species with known operons. The Agr components were mostly similar within species and in some cases, differed between other Clostridial species. Conserved residues of unknown function were also found. The prevalence of the Agr system and the identification of common motifs in its components opens up therapeutic targets to be harnessed for the development of non-antibiotic and anti-virulence therapies for pathogenic Clostridial infections.

TABLE OF CONTENTS

List of Tables……i
List of Figures……..ii
List of Appendices……v
Introduction……..6
The Clostridium Genus and its Relevance……...6
Quorum Sensing and its Presence in Clostridium Bacteria……7
The Accessory Gene Regulator……...8
The Agr System and its Significant in Clostridia and Other Species……12
Rationale for Project……...15
Specific Aims……17
Methods……..18
Materials……18
Methods……..18
Results and Discussion……...25
The Sequence Identity of the Agr Components of *C. botulinum*……31
The Sequence Identity of the Agr Components of *C. difficile*……38
The Sequence Identity of the Agr Components of *C. sporogenes*……..45
The Sequence Identity in the Agr Components Between Clostridial Species.....50
Comparison of the AgrD Sequence Between the Five Clostridial Species.....50
Comparison of the AgrB Sequences Between the Five Clostridial Species.....52
Comparison of the AgrC Sequences Between the Five Clostridial Species.....54
Comparison of the AgrA Sequences Between the Five Clostridial Species.....58
Comparison of all of AgrD Sequences Between Clostridial Species.....60
Comparison of all of AgrB Sequences Between Clostridial Species.....64
Comparison of all of AgrC Sequences Between Clostridial Species.....68
Comparison of all of AgrA Sequences Between Clostridial Species.....73
Evolutionary Inference of the Agr Components.....78
Conclusion……81
Appendix I……..86
References……..87

LIST OF TABLES

Table 1: The Components and Arrangement of the Agr Systems in Clostridium Species……12

LIST OF FIGURES

Figure 1: Percent Sequence Identity of all the Homologs of AgrD Autoinducing Prepeptide Variants in Clostridial Species…….....26

Figure 2: Percent Sequence Identity of all the Homologs of AgrB Proteins in Clostridial Species…….27

Figure 3: Percent Sequence Identity of all the Homologs of AgrC Proteins in Clostridial Species…….28

Figure 4: Percent Sequence Identity of all the Homologs of AgrA Proteins in Clostridial Species…….29

Figure 5: (A) Comparative Analysis of the *S. aureus* AgrDI-IV and AgrD1 Sequences of *C. botulinum* Strains…….30

Figure 6: (B) Wheel diagram mimicking the putative amphipathic helix of *C. botulinum* AgrD1…….30

Figure 7: Comparative Analysis of the Sequences of *S. aureus* AgrBI-IV and AgrB1 Sequences of Strains of *C. botulinum*…….33

Figure 8: (A) Comparative Analysis of the Amino Acid Sequences of *S. aureus* AgrDI-IV and AgrD2 of *C. botulinum* Strains…….34

Figure 9: (B) Wheel diagram mimicking the putative amphipathic helix of *C. botulinum* AgrD2…….34

Figure 10: Comparative Analysis of the Amino Acid Sequences of *S. aureus* AgrBI-IV and AgrB2 of *C. botulinum* Strains…….36

Figure 11: (A) Comparative Analysis of the Amino Acid Sequences of *S. aureus* AgrDI-IV and AgrD2 of *C. difficile* Strains…….37

Figure 12: (B) Wheel diagram mimicking the putative amphipathic helix of *C. difficile* AgrD2…….37

Figure 13: Comparative Analysis of the Amino Acid Sequences of *S. aureus* AgrBI-IV and AgrB2 of *C. difficile* Strains…….39

Figure 14: Comparative Analysis of the Amino Acid Sequences of *S. aureus* AgrCI-IV and AgrC2 of *C. difficile* Strains…….42

Figure 15: Comparative Analysis of the Amino Acid Sequences of *S. aureus* AgrA and AgrA2 of *C. difficile* Strains…….43

Figure 16: (A) Comparative Analysis of the *S. aureus* AgrDI-IV and AgrD1 Sequences of *C. sporogenes* Strains……..44

Figure 17: (B) Wheel diagram mimicking the putative amphipathic helix of *C. sporogenes* AgrD1……..44

Figure 18: Comparative Analysis of the Sequences of *S. aureus* AgrBI-IV and AgrB1 Sequences of Strains of *C. sporogenes*. ……...45

Figure 19: (A) Comparative Analysis of the Staphylococcus aureus AgrDI-IV and AgrD2 Sequences of *C. sporogenes* Strains……47

Figure 20: (B) Wheel diagram mimicking the putative amphipathic helix of *C. sporogenes* AgrD2…….……47

Figure 21: Comparative Analysis of the Sequences of S. aureus AgrBI-IV and AgrB2 Sequences of Strains of *C. sporogenes*…….48

Figure 22: Comparative Analysis of the *S. aureus* AgrDI-IV and AgrD Consensus Sequences of Clostridium Species with Quorum-Sensing Agr Components…….49

Figure 23: Comparative Analysis of the S. aureus AgrBI-IV and AgrB Consensus Sequences of Clostridium Species. ……...52

Figure 24: Comparative Analysis of the S. aureus AgrCI-IV and AgrC Consensus Sequences of Clostridium Species with Quorum-Sensing Agr Components. ……..55

Figure 25: Comparative Analysis of the S. aureus AgrA and AgrA Consensus Sequences of Clostridium Species. ……...57

Figure 26: Comparative Analysis of the S. aureus AgrDI-IV and AgrD Consensus Sequences of all Clostridium Species. ……..60

Figure 27: Comparative Analysis of Regions with Relevant Similarities and Differences Between the Sequences of S. aureus AgrBI-IV and AgrB Consensus Sequences of all Clostridium Species…….66

Figure 28: Comparative Analysis of Regions with Relevant Similarities and Differences Between the Sequences of S. aureus AgrCI-IV and AgrC Consensus Sequences of all Clostridium Species. ……...69

Figure 29: Comparative Analysis of the S. aureus AgrA and AgrA Consensus sequences of all Clostridium species. ……...73

Figure 30: The Bootstrap Phylogenetic Trees of the Consensus AgrD (Right) and AgrB (Left) Sequences of Clostridium Species. ……...78

Figure 31: The Bootstrap Phylogenetic Trees of the Consensus AgrC (Left) and AgrC (Right) Sequences of Clostridium Species……...79

LIST OF APPENDICES

Appendix I: List of Species Included in Analysis……...85

INTRODUCTION

The Clostridium Genus and its Relevance

With over 300 species, the Clostridium genus is one of the largest prokaryotic genera of the Firmicutes phylum. These ancient bacteria are Gram-positive obligate anaerobes that form endospores. They are rod-shaped, fermentative bacteria that do not produce catalase. As a result of fermentation, however, they produce valuable compounds such as butyric acid, acetic acid, butanol, acetone, and large amounts of CO_2 and H_2. Colonizing almost all organic-containing anaerobic habitats, this genus of bacteria is ubiquitous. They produce enzymes that catabolize large molecules, such as proteins, lipids, cellulose, and collagen into fermentation precursors and participate in processes of biodegradation and carbon cycling (Darkoh & Asiedu, 2014).

Due to their catabolism potential, Clostridia are considered medically and biotechnologically relevant bacteria. *Clostridium botulinum* produces one of the deadliest toxins on earth (Darkoh & Asiedu, 2014) and is considered a biological warfare threat (Arnon et al., 2001). *C. difficile* causes both primary and recurrent infections. In the United States, *C. difficile* recurs at a rate of 25% after antibiotic treatment (Darkoh, DuPont, Norris, & Kaplan, 2015), costing an estimated $2.8 billion in total healthcare costs (Rodrigues, Barber, & Ananthakrishnan, 2017). Another emerging pathogen, *Clostridium sordellii*, causes myonecrosis, sepsis, and shock (Darkoh & Asiedu, 2014). In the United States, *Clostridium perfringens* was responsible for 10% of yearly food-related illnesses between 2000 and 2008 (Scallan, 2011), and ranked second among the most common foodborne diseases between 1998 and 2010 (Grass, Gould, & Mahon, 2013). There are also pathogens that could potentially affect agriculture by infecting livestock, such as *Clostridium chauvoei* (Darkoh & Asiedu, 2014) and *C. perfringens* (Yu et al., 2017).

Although the Clostridium genus segregates into these medical and non-medical areas of relevance, their phylogeny does not follow the same segregation. A Multi-locus sequencing analysis of four housekeeping genes in the Clostridium genus revealed that toxigenic and pathogenic bacteria are spread throughout the phylogenetic tree. Similarly, the genome sizes of these bacteria do not correlate with the two traits and vary from 2.55 Mb for *C. novyi* to 6.00 Mb for *C. beijerinckii*. Neither does the number of open reading frames, as some species have more than twice the number of proteins in their genomes compared to others (Udaondo, Duque, & Ramos, 2017). The variations in the genomes of Clostridia are evident, but do not correlate with the pathogenicity of the bacteria. On the other hand, there are reports that show an operon used for communication and cell regulation with similar structure and function within a few Clostridium species and other genuses. The genes of this operon are prominent in different gram-positive species and given the relevance of the Clostridium genus, are an interesting subject of comparison similar to the four housekeeping genes mentioned previously. .

Quorum sensing and its presence in Clostridium bacteria

Quorum signaling allows bacterial cells to communicate and regulate gene expression based on population density. Therefore, quorum-signaling systems allow bacteria to respond to their environment, making it an indispensable mechanism for bacterial virulence and physiology. Although there are different quorum sensing mechanisms, Gram-positive bacteria mediate their signaling process through a secreted peptide called autoinducing peptide (AIP). The three steps necessary for quorum sensing are production of the AIP, its recognition, and the response it ensues within the cell. The production of AIP happens through post-translational processing of the autoinducer pre-peptide by a peptidase, which processes the linear or cyclical AIP for secretion

extracellularly. At the extracellular membrane, a two-component sensor histidine kinase detects the AIP and autophosphorylates. The phosphoryl group is then transferred onto a response regulator within the cytoplasm that effects the regulation of the quorum signaling system. In some systems, the actual AIP is taken into the cell to interact with receptors and transcription factors. Some bacteria also have a positive feedback loop for the quorum sensing genes as the system regulates the expression of its own genes (Darkoh & Asiedu, 2014).

Clostridia, specifically, have two different mechanisms of quorum sensing, the Accessory Gene Regulator (Agr) and the LuxS systems. The LuxS system, however, has a metabolic byproduct for a signal and, therefore, is not considered a real quorum sensing system. On the other hand, the Agr system has genes encoding all four components, including the pre-peptide, the pre-peptide processing protein, the sensor histidine kinase, and the response regulator (Darkoh & Asiedu, 2014). As will be shown in the sections below, the Agr system is responsible for crucial processes within Clostridia and will be the focus of this investigation.

The Accessory gene regulator

The Agr system is a quorum signaling system widely found in Clostridia and responsible for vital functions within the bacteria. However, the system has not been as thoroughly explored in Clostridia, but it is well characterized in the Staphylococcus genus. In *Staphylococcus aureus*, for example, the Agr system regulates colonization and toxin production (Darkoh & DuPont, 2017) through its four genetic components: *agrA*, *agrC*, *agrD*, and *agrB*. The proteins AgrA and AgrC are the response regulator and sensor histidine kinase, respectively. They sense and translate the message of the cyclic-autoinducer (c-AIP), derived from the pre-peptide AgrD. *S. aureus'* AgrB is the protein that processes AgrD into an intermediate between AgrD and the fully functional c-

AIP (Darkoh & Asiedu, 2014) that will be further processed and excreted by the protein SpsB (Cisar, & Elizabeth, 2009). Once the c-AIP is sensed and the AgrA becomes phosphorylated, AgrA binds to the P2 promoter leading to expression of the Agr system proteins through positive feedback. Additionally, AgrA binds to the P3 promoter, which is responsible for the expression of genes involved in regulating toxin production and colonization (Darkoh & Asiedu, 2014). Interestingly, the Agr system in *S. aureus* has been categorized into four different groups containing variations of the Agr proteins that, nonetheless, regulate the same genes. Because of the variation within the Agr system, the individual components of *S. aureus'* Agr system have been thoroughly characterized and provide a valuable homolog for comparison with Clostridia.

The AgrA of *S. aureus*, like most response regulators, consists of two domains, a regulatory domain at its N-terminus and an effector domain at its C-terminus (Stock, Robinson, & Goudreau, 2000). The former domain is a receiver (REC) domain that enables activation and dimerization of the AgrA component following phosphorylation. The phosphoryl group binds to a conserved Asp residue in the REC domain as the ATP molecule is stabilized by its interactions between $Mg2^{+}$ ions and an Asp and a glutamine residue. Once activated, a Lys residue forms a salt bridge with the bound phosphoryl group (Gao & Stock, 2009). The same interactions occur at the ATP binding site in *S. epidermidis*, but with a second aspartate instead of the glutamate in *S. aureus* (Zhiqiang et al., 2004). The latter domain of *S. aureus'* AgrA is the effector domain and is conserved throughout different response regulators of two-component systems, including VirR of *C. perfringens* (Nikolskaya & Galperin, 2002). The C-terminus domain, termed the LytTR domain, is structured as a 10-stranded elongated β-β-β fold. Out of the loops of an edge of the domain emerge the side chains of residues H169, N201, and R223 that bind to the DNA and activate

transcription (Sidote, Barbieri, Wu, & Stock, 2009). Interestingly, AgrA is the only component conserved throughout all four groups.

AgrA receives its activating phosphoryl group from AgrC. AgrC is part of the 10HPK family and contains a sensor domain connected to histidine kinase domain by an α-helical linker (Wang, Zhao, Novick, & Muir, 2014). The sensor domain is composed of transmembrane segments in the N-terminal domain. The first and second extracellular loops between the transmembrane segments are responsible for activation and specificity, respectively (Cisar, Geisinger, Muir, & Novick, 2009). The activation translates through physical changes in the protein to allow phosphorylation of the histidine kinase (HK) domain. The HK domain contains two subdomains that work together to autophosphorylate AgrC (Wang et al., 2017). The subdomains are the helical dimerization and histidine phosphorylation (DHp) subdomain, and the catalytic ATP-binding subdomain (Cisar & Elizabeth, 2009). The autophosphorylation happens at His239, which is located within the H-box motif of the HK domain. The domain also has important residues in the N-box and the G-box motifs, both of which delineate the ATP binding pocket (Stock, Robinson, & Goudreau, 2000). The Asn339 in the N-box was mutated to Asp and AgrC activity was partially reduced, while the two glycine residues at positions 394 and 396 of the G-box lead to complete inactivation of AgrC, when mutated to Ala (Cisar et al., 2009).

Before AgrC can sense the c-AIP, AgrB has to cleave the pre-peptide. The peptidase is a unique protein, as it is not homologous to other proteins apart from AgrBs in Gram-positive bacteria (Thoendel & Horswill, 2013). Located in the membrane, the AgrB spans through to the extracellular milieu a few times, but there is a debate on the topology of the membrane (Zhang, Gray, Novick, & Ji 2002; Thoendel & Horswill, 2013). The catalytic residues of AgrB, His77 and Cys84 are more accessible to the cytoplasmic milieu and to AgrD (Qiu, Pei, Zhang, Lin, & Ji,

2005). The different AgrDs of *S. aureus* are recognized through different mechanisms as different parts of AgrB are involved in the processing of different AgrDs (George & Muir, 2007). Furthermore, the first 34 amino acids of AgrB, conserved throughout all groups, are essential for AgrD processing as mutations lead to undetectable levels of the c-AIP (Qiu et al., 2005).

The last component of the Agr system is the AgrD. The pre-peptide has three segments, including the amphipathic N-terminus that is tethered to the cytoplasmic membrane, the residues that will become the AIP, and the predominantly charged C-terminus (Kavanaugh, Thoendel, & Horswill, 2007). The segments have specific functions in the three steps that lead to the transformation of AgrD into c-AIP. The N-terminus tethers the pre-peptide close to the membrane-bound AgrB to facilitate the second cleavage step and increase the rate of AIP processing (Wang & Muir, 2016). In *S. aureus*, the amphipathic region also has the recognition site for the second cleavage, which is carried out by a more common peptidase called SpsB (Kavanaugh et al., 2007). The residues that become the AIP have a conserved Cys28, where the end of the AIP forms a thioester linkage. In some bacteria, the cAIP also has a tail composed of 1-4 residues. Both the tail and thioester linkage are necessary for activation of AgrC (Cisar & Elizabeth, 2009). Furthermore, the AIP residues also have a conserved motif of two or three hydrophobic residues that form a hydrophobic knob (Tal-Gan et al., 2013). The hydrophobic knob in addition to the thioester linkage are necessary for bioactivity of AIP (Cisar & Elizabeth, 2009). Lastly, the C-terminus segment is responsible for recognition and interactions that facilitate cleavage of the first transformational step (Cisar & Elizabeth, 2009).

The Agr system and its significance in Clostridia and other species

Clostridial species utilizes the Agr system as a key player in their pathogenesis pathways. Clostridial Agr proteins are homologous to the Agr genes of *S. aureus*. Table 1 shows the arrangement and orientation of the Agr systems in Clostridia in relation to that of *S. aureus*. Evidently, there are similar Agr components between *S. aureus* and Clostridia, but within the Clostridium genus as well. Although there are some variations between the Agr components in Clostridia, the genes for *agrB* and *agrD* are present within all Clostridium species. Most importantly, there are similarities in function between the *S. aureus* Agr system and the Agr system of Clostridium species. However, some Clostridial strains encode two Agr systems in their genomes and these are designated Agr1 and Agr2. The Agr1 locus contains only the genes required for AIP synthesis (AgrD1 and AgrD2) whereas the Agr2 locus encodes genes required for both

Table 32: The Components and Arrangement of the Agr Systems in Clostridium Species

Clostridium Species	**Agr system components**
C. acetobutylicum	*agrB1D1*, *agrB2D2* (Darkoh & Asiedu, 2014)
C. botulinum	*agrB1D1*, *agrB2D2* (Darkoh & Asiedu, 2014)
C. difficile	*agrD1B1*, *agrA2C2D2B2* (Darkoh & Asiedu, 2014; Stabler et al., 2009), *agrC3B3D3* (Hargreaves, Kropinski, & Clokie, 2014)
C. perfringens	*agrB1D1* (Gray, Hall, & Gresham, 2013)
C. sporogenes	*agrBDCA* (Darkoh, Odo, & DuPont, 2016)
S. aureus	*agrBDCA* (Darkoh & Asiedu, 2014)

AIP synthesis (AgrB2 and AgrD2) and response (AgrC2 and AgrA2). Recently, a third Agr locus was described in *C. difficile* containing *agrC3B3D3* (Hargreaves, Kropinski, & Clokie, 2014).

The Agr system in Clostridia, similar to *S. aureus*, regulates toxicity, colonization, and expression of similar target genes (Darkoh & Asiedu, 2014). Specifically, the *C. botulinum* a*grB2D2* regulates its neurotoxin production. Such regulation was determined by knocking out *agrD2, leading to* a phenotype of decreased toxin production that could be restored by complementation (Cooksley et al., 2010). The production of *C. difficile* toxin A also decreased significantly once *agrA2* was knocked out (Martin et al., 2013). Furthermore, deletion of *agrB1D1* in *C. difficile* resulted in loss of toxin production (Kök, 2015). Another Clostridium species that has toxin production regulated by *agrD1B1* is *C. perfringens*, as it only has one *agr* locus. The *agr* locus regulates toxin production in all strain types of *C. perfringens* (Chen & McClane, 2012; Darkoh & DuPont, 2017; Li, Chen, Vidal, & McClane, 2011; Ohtani et al., 2009; McClane et al., 2012). Regarding colonization, knocking out *agrA2* in *C. difficil*e significantly reduced colonization of mice (Darkoh & Asiedu, 2014; Martin et al., 2013). Another similarity between *C. perfringens* and *S. aureus* is how the Agr system regulates the expression of a regulatory RNA (rRNA) molecule. Similar to *S. aureus*, different toxinotypes of *C. perfringens* also express two proteins (VirR and VirS) that respond to the quorum signal. The VirR and VirS of *C. perfringens* are analogous to the *S. aureus* AgrA and AgrC, respectively. Furthermore, the *S. aureus* RNAIII, regulated by AgrA and AgrC, corresponds to the VR-RNA regulatory molecule in *C. perfringens*. Similarly, VirR and VirS also regulate VR-RNA, which is also involved in toxicity. Therefore, *S. aureus* and *C. perfringens* show functional similarities in their Agr systems (Ohtani, 2016) and can be considered homologous.

Apart from toxicity and mice colonization, the Agr system within Clostridia also modulates motility and sporulation. *C. difficile* moves by using flagella, which are hair like structures that propel the bacterium. Flagellar synthesis and its regulation were severely affected in *C. difficile* with a mutant *agrA2* (Martin et al., 2013). *C difficile* is the only Clostridium species proven to regulate motility through the Agr system. On the other hand, many Clostridia regulate sporulation through the Agr system. *C. acetobutylicum*'s spore formation significantly decreases after knocking out *agrA* and a*grC*. These mutants, including that of *agrB*, also exhibit a decrease in granulose and endospores formation, both direct consequences of sporulation (Steiner et al., 2012; Jabbari et al., 2013). In contrast, *C. botulinum*'s *agrB1D1* is involved in sporulation because an *agrD1* mutant could not produce spores effectively. On the other hand, *C. sporogenes* spore production depends on both *agrB1D1* and *agrB2D2* genes (Cooksley et al., 2010). A *C. perfringens* type A mutant with an inactive *agr* locus had sporulation efficiency of less than one percent. Furthermore, various gene products necessary for sporulation were mostly or completely absent in the mutant. These genes included *Spo0A* transcripts involved in sporulation initiation; enterotoxin production during sporulation; and sporulation sigma factors that initiate transcription of sporulation regulators (McClane et al., 2015). In contrast, there is no primary data in the literature proving a relationship between the *C. difficile* Agr system and sporulation. There is data, however, that shows an increase in *agrD* expression of 2.5 concurrent with expression of sporulation sigma factors (Saujet et al., 2011). Additionally, like *C. perfringens* type A, *C. difficile* expresses the *Spo0A* protein involved in sporulation regulation (Underwood et al., 2009).

Interestingly, experiments by Verbeke et al. (2017) suggested that the Agr system of *C. thermocellum* does not function as a quorum signaling system and regulates bacterial growth in specific conditions. *AgrD1* seems to be upregulated by a factor of 2.3 in the presence of the sugar

xylose in *C. thermocellum*. Furthermore, the bacterium's *agrD1* also inhibited growth in the absence of the xylose sugar. Nevertheless, the specific mechanism of growth inhibition is still unknown (Verbeke et al., 2017).

Despite the significance of the Agr system in Clostridia, our understanding of the system is limited. The Clostridium genus gets little mention in comparisons of the Agr proteins throughout the Firmicute phylum (Wuster & Babu, 2008; Peter, 2014). Although increasingly focused comparisons exist, they are limited to single components within and between specific classes of Firmicutes (Canovas et al., 2016; Darkoh et al., 2015; Ohtani et al., 2009). Unfortunately, these comparisons do not include Clostridia as a genus and analyses do not include all of the components of the Agr system. To better understand the similarity and diversity of the Agr system within Clostridia, we compared the sequences of Agr components of over 50 species through multiple sequence alignments, motif and structure-specific bioinformatics tools, and phylogenetics. This thesis addresses the differences within and between the Agr components of Clostridium species and provides potential paths of investigation on the potential of targeting the components for therapy.

Rationale for project

Although a comprehensive comparison of the Agr systems among Clostridia has not been conducted, data about the mechanisms and functions of its components between and within Clostridia suggest structural similarity. However, a similarity in structure does not rule out differences in residues, motifs, and even secondary structures. While the Agr system has similar functions between and within Clostridia, the systems' functions also vary, ranging from toxin production to sporulation. Given its different functions, understanding the Agr system will provide

different paths for pathogen treatment development. Furthermore, therapies targeting the Agr components could be more effective than current therapies such as antibiotics, as resistance is less likely to develop given that the system does not directly affect growth (Darkoh & DuPont, 2017).

The potential for targeted manipulation and modulation of the Agr system in the medical field relies on the understanding of the similarity and diversity of the Agr system. This understanding will come from a detailed analysis of the residue-specific similarities and differences between the Agr components within and between Clostridium species. The analysis compares the sequences for each Agr component throughout all species with comprehensive alignments. Thus, the analysis will orient research efforts towards amino acid motifs and domains with a robust potential of functional significance and plausible malleability. Furthermore, phylogenetic trees will show the ancestral relationship between the sequences based on the alignments. These trees will also uncover if the Agr sequences relate to a species' toxicity, a relationship that has not been explored yet. Therefore, this research expands our understanding of the function of the Agr system within Clostridia and demonstrates that the Agr system may be a good target for therapies.

SPECIFIC AIMS

The Agr system is responsible for regulating virulence and other cellular mechanisms in many Gram-positive pathogens that cause life threatening infections. In this study, the sequences of the Agr system components were compared to determine similarities and differences among them. These specific aims were:

Aim 1: To conduct a comprehensive comparative analysis of the similarities and differences between the components of the Agr system in Clostridia.

Aim 2: To use bioinformatics tools to predict structural features of the Agr components within and between Clostridial species.

Aim 3: To generate a phylogenetic tree to determine the evolutionary relationship between the components of the Agr system in the different Clostridia.

METHODS

Materials

The amino acid sequences of the four different Agr components, AgrA, AgrB, AgrC, and AgrD were analyzed. A list of the bacteria analyzed in this study are shown in Appendix I. The sequences of the Agr proteins of the listed Clostridium species were downloaded from the website of the National Center for Biotechnology Information (NCBI). Within the NCBI website is the BLASTP 2.7.1+ program (Altschul 1991), which was used to search for and downloaded the Agr protein sequences. The downloaded Agr protein sequences were also compiled with the BioEdit program (Hall, 1999). SignalP (Nielsen, 2017), Predisi (Hiller et al., 2004), and Phobius (Käll et al., 2004; Käll et al., 2007) were used to predict the quorum sensing signaling peptide cleavage sites and HeliQuest (Gautier, Douguet, Antonny, & Drin, 2008) was used to predict the helical composition for AgrDs. Furthermore, PSIPRED was used to predict secondary structure of the AgrB and AgrC sequences. All sequences were aligned using the MUSCLE (Edgar, 2004a) program. Based on the MUSCLE aligned sequences, the MEGA X (Kumar et al., 2018) program was used to estimate statistically supported maximum likelihood phylogenetic trees.

Methods

The AgrD amino acid sequence of *Clostridium difficile* 630 strain (Accession or identification number: CAJ69637.1) was used as the starting sequence and searched with the BLASTP program of NCBI. The BLASTP search parameters were set to default, except the *Max Target Sequence* parameter, which was set to output 20,000 sequences. The *Database* parameters were set to *Non-redundant protein sequences (nr)* to search through the most extensive protein sequence databases (GenBank CDS translations, RefSeq, PDB, SwissProt, PIR, PRF, excluding

those in PAT, TSA, and env_nr). The parameter for *Organism* was left blank, as there are different names for the same organism. The *Exclude* parameter was left blank to avoid excluding low value sequences and nothing was indicated in the *EntreZ Query* parameter, which aims at limiting the search to certain protein types, sequence lengths or organisms. The parameter for *Program Selection* was left as the default *blastp (protein-protein BLAST)* as it is the most general of the protein to protein search programs from BLASTP. The *Max Target Sequences* parameter, which determines the "maximum number of aligned sequences to display" (Altschul 1991), was set to 20,000. Likewise, the *Expect threshold*, which determines the cutoff E-value for the search, was left at the default value of ten. The E-value determines the statistical significance of the match of a sequence to the query sequence (lower E-values are more significant). The *Short Queries* parameter was set to default, which "automatically adjusts parameters for short input sequences" (Altschul 1991). The parameter *Word Size* does not make a significant difference for BLASTP programs as incomplete words are also matched to assess a possible alignment during the search. Therefore, *Word Size* was set to the default value of six. The *Maximum Matches in a Query Range* parameter limits the search to output a certain number of results per region of the protein. Given that the sequences of all Agr proteins only have one functional region, the *Maximum Matches in a Query Range* parameter was set to the default value of zero. The *Matrix* parameter provides options for different substitution matrices. Substitution matrices score the quality of the alignment based on alignment of pairs of residues (Altschul, 1993; Altschul, 1991; Cooksley et al., 2010; Edgar, 2004b). So, the scores of the pairs determine the composite alignment score. BLOSUM-62 was the scoring matrix chosen for the *Matrix* parameter, as it is the best scoring matrix available (Arnon et al., 2001). The parameter for *Gap Costs* determines the penalties that gap introduction has on the alignment score. The higher the gap cost, the least gaps introduced (Altschul 1991). As there

was not a high expectation for gaps, the default value *Existence: 11 Extension: 1* was used. The parameter *Compositional Adjustments* accounts for the amino acid composition of the sequences aligned. The *Compositional Score Matrix Adjustment* was present as default and the chosen option for *Compositional Adjustments* was used throughout the entire investigation, although the *Composition-based Statistics* was suggested for general use (Altschul 1991). Because the parameter *Compositional Adjustments* was used, the parameters *Filter*, which filters results that match due to uninteresting regions, and *Mask*, which masks the query sequence according to the *Filter* parameter (Altschul 1991), were not necessary.

The sequences were screened and those with the best match were downloaded. Statistically, the best sequences were the ones with highest alignment score or lowest E-value (Altschul 1991). This criterion was disregarded only when the graphical representations of the sequences at the top of the search results page showed a shorter bar. As the graphical bar indicates coverage of the query by the aligned sequence (Altschul 1991), a shorter bar indicates less coverage. Less coverage could mean an incompletely sequenced protein and would skew the data. The sequences along with their accession numbers were copied into a Bioedit alignment file. To confirm the sequence selected was actually part of the Agr system, all of the sequences were also located within the organism's genome sequence. The presence of the Agr system components and arrangement or orientation were noted. For instance, if AgrD or AgrB were not flanked by each other, they did not meet this inclusion criterion. If AgrD and AgrB were flanked only by AgrA or AgrC, then the Agr A or AgrC was indicated as an orphan protein. Furthermore, at least one known conserved domain (Marchler-Bauer et al., 2017; Marchler-Bauer et al., 2015; Marchler-Bauer et al., 2011; Marchler-Bauer & Bryant, 2004) had to be present in one of the protein sequences of the operon for inclusion. Sequences that met these criteria were included in the alignment. Another method used for finding

sequences was researching for Clostridia that had the conserved domains of the Agr proteins. The Conserved Domain Architecture Retrieval Tool (Marchler-Bauer et al., 2015) provided the sequences containing the domains. The sequences were filtered through the search and retrieval system of NCBI, Entrez (Ostell, 2014), until the Agr protein sequence within the species of interest was found. Once the sequence was found, the same inclusion criteria were applied for inclusion in the alignment.

Sequences were grouped into alignments files (ALs) according to protein type and species. One set of alignment file contained the protein sequences of each Agr protein within each species (**AL1**: *C. difficile* AgrA2; **AL2**: *C. difficile* AgrB2; **AL3**: *C. difficile* AgrC2; and **AL4**: *C. difficile* AgrD2; **AL5**: *C. difficile* AgrB1; **AL6**: *C. difficile* AgrD1; **AL7**: *C. botulinum* AgrB1; **AL8**: *C. botulinum* AgrD1; etc.). The other set of ALs contained the consensus sequences of each Agr component within each species (**AL1**: *C. difficile* AgrA_consensus, *C. botulinum* AgrA_consensus, *C. perfringens* AgrA_consensus; **AL2**: *C. difficile* AgrB_consensus, *C. botulinum* AgrB_consensus, *C. perfringens* AgrB_consensus; etc.). Specifically, the ALs containing the consensus sequences were also split into two sets. One set contained the consensus sequences of Agr components with empirical quorum-sensing function and the other set contained all of the Agr components.

A consensus sequence contains the most prominent amino acids in each position given the sequences aligned. The consensus sequences were created in Bioedit (Hall, 1999). The *consensus sequence* function in Bioedit was set to ignore gaps, as the individual Agr protein sequences within species were highly similar and a full sequence was warranted for further analysis. Another option for the *consensus sequence* function allows setting a threshold frequency for inclusion of amino acids in consensus sequences. The threshold value assigned was found through testing values by

trial and error at 10% intervals down from 100% until the consensus sequence had all positions filled with an amino acid. All consensus sequences were produced based on alignments processed by the MUSCLE alignment program.

The MUSCLE program is one of the most widely used programs for rendering multiple sequence alignments. The program is highly rated and performs at higher speed and accuracy compared to other alignment programs (Baum & Smith, 2013). Creating alignments with MUSCLE is simple with the single code provided. The program aligns sequences in the FASTA format of sequence representation.

Once the alignments were ready, the identity between sequences was established. Identity analysis entailed rendering identity matrices demonstrating the percentage of amino acid similarity between the sequences. The lowest percentage of identity within an alignment was used as a measure of conservation. The lowest identity percentages were presented in reference to the 35 percent (Rost, 1999) homology cutoff for a given alignment.

Identity was also visually assessed in the alignments based on the decision of how similar the aligned residues were within a specific region of an Agr component. This decision was largely guided by Betts' and Russell's chapter on Amino Acid Properties and Consequences of Substitutions (Betts & Russell, 2003). Additionally, the BLOSUM62 amino acid similarity index (S. Henikoff & J. G. Henikoff, 1992) aided in finding similar and conserved residues within the alignments. The assessment entailed a comparison between the sequences of Agr components in Clostridium species in reference to that of *Staphylococcus aureus*. *S. aureus* was included in the alignments because its Agr system is well characterized. Some specific regions relevant to *S. aureus* were identified in the alignment to target these domains as potentially relevant. Thus, the ability to discern relevant differences and similarities was more focused. Some domains were

identified through NCBI (Marchler-Bauer et al., 2015) and some by simply aligning *S. aureus*' domains.

An even more focused assessment was used for the components with empirically proven quorum-sensing function by predicting secondary structures. The secondary structures of amino acid sequences can be conveniently and effectively predicted through PSIPRED, which offers a simple web user interface that does not depend on any parameters (Jones, 1999). Deeper assessments of similarity were also used to determine AgrD's similarity. The presence of an amphipathic helix in the AgrDs of *S. aureus* was used to identify similar properties in the AgrDs of Clostridia using HeliQuest (Gautier, Douguet, Antonny, & Drin, 2008). The HeliQuest program draws wheels as a top down overview of the residues in a helix, in this case an alpha helix. The residues that are close to each other are predicted to be on the same side of the helix potentially creating a face containing similar properties and a specific function. In addition, the quorum-sensing signaling peptide cleavage sites were predicted through SignalP (Nielsen, 2017), Predisi (Hiller et al., 2004), and Phobius (Käll et al., 2004; Käll et al., 2007). In general, the alignments allowed us to find similarities and differences within the sequences of Clostridia and infer probable functional regions of interest to target. Following identity analysis, MEGA X (Kumar et al., 2018) was used to render the maximum likelihood phylogenetic trees.

Maximum likelihood (ML) phylogenetic trees provide a phylogenetic or evolutionary history based on evolutionary distance. Evolutionary distance reflects the average number of differences in each position of a sequence. ML is the best method for calculating evolutionary distance between sequences due to its statistical power and foundation on proven mathematical models. Once evolutionary distances are established through ML, the tree with highest probability of reflecting these distances is rendered (Altschul et al., 1997).

The MEGA X (GUI) program used the comprehensive alignments of the quorum-sensing Agr components to conduct the evolutionary analysis. The evolutionary history was inferred by using the MLmethod and Jones-Taylor-Thornton (JTT) matrix-based model (Jones, Taylor, & Thornton, 1992). The bootstrap consensus tree inferred from 300 replicates was taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985). Branches corresponding to partitions reproduced in less than 50% of bootstrap replicates were collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (300 replicates) are shown next to the branches (Felsenstein, 1985). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model (Jones, Taylor, & Thornton, 1992), and then selecting the topology with superior log likelihood value.

RESULTS AND DISCUSSION

Apart from the Agr systems of the five Clostridial species that have proven quorum-sensing function, additional putative Agr components were found within the five Clostridial species. Furthermore, other Clostridia without previously reported Agr components also were found to have homologs to the Agr components of *S. aureus* and the five aforementioned Clostridia. The presence of the Agr system in several Clostridium species suggest the importance of this regulatory system in their biology and pathogenesis.

The results showed that the Agr components of Clostridial species are mostly similar between the strains of a particular species. The degree of similarity is directly proportional to the percent identity in the alignments of each component within each species and thus, the lowest percent identity indicates high dissimilarity and variability. **Figures 1-4** show the percent identities of the Agr components within Clostridial species that have more than one sequence for an Agr component. Overall, most of the alignments show identity proportions above the 35 percent cutoff for homology (shown on the figures as red lines). Some of the Agr components were found to vary significantly, including *C. botulinum's* AgrD3, which is a newly found autoinducing peptide, and *C. sordellii's* AgrB2, D2, D3 and A3. The sequences of different components have the same degree of variation within the same loci in *C. sordellii, C. difficile, C. botulinum, C. butyricum, C. pasteurianum, C. sphenoides, C. beijerinckii,* and *C. kluyveri*. The Agr components of *C. beijerinckii,* and *C. kluyveri,* however, have different degrees of variation within the same locus. Another noteworthy trend of conservation within the Agr loci is the tendency of the Agr operon to have greater conservation if the species only has a single Agr locus in its genome, as opposed to multiple Agr loci. Thus, there is no apparent sequence variation of the Agr proteins in species with

a single locus of AgrBDCA and AgrBD. The majority of the Agr components are homologous in their respective operons within their species, as they considerably surpass the homology cutoff devised by Rost (1999). The high degree of similarity supports the notion that the Agr system is important to Clostridia.

Although, the amino acid sequences of the Agr components of the same species are similar, the components might not necessarily be the same proteins. Using multiple sequence alignment, the sequences of the Agr components of the same operon in different strains of the five species were compared. The alignments allowed for assessment of identity and significant differences based on the comparability of the residues. However, some components were mostly identical within their alignments and were not included in the results. The components with significantly similar sequences and minimal differences were *C. acetobutylicum's* AgrD, B, C, and A; *C. difficile's* AgrD1 and B1; and *C. perfringens*' AgrD and B. On the other hand, the alignments of the other components contained significant differences and can be found in **Figures 5-16**.

Although the purpose of **Figures 5-16** is to show the Agr components' similarities within the Clostridial species, the *S. aureus* sequences were also included to demonstrate similar features between the Clostridial Agr components and confirm the presence of motifs.

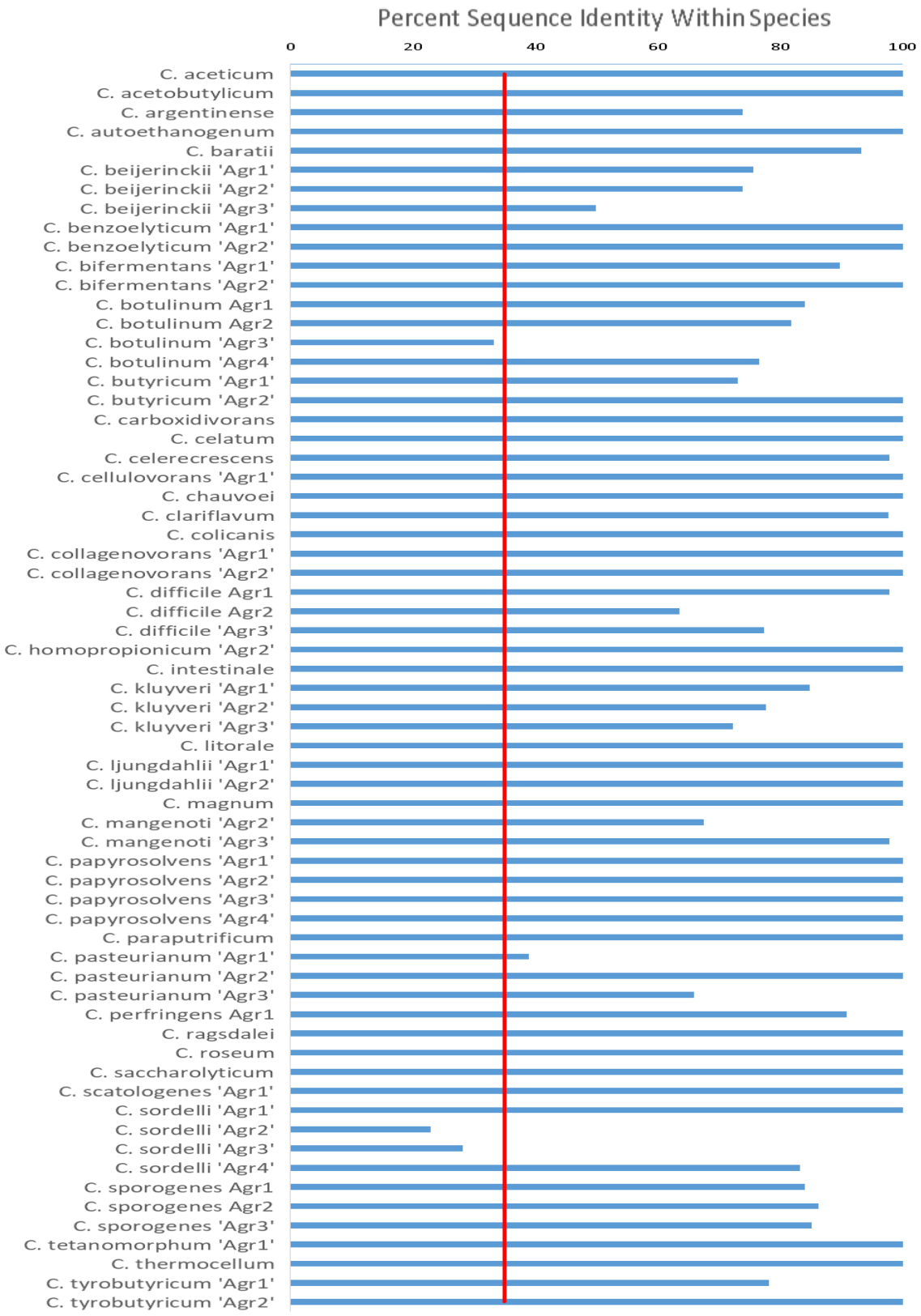

Figure 33: Percent sequence identity of all the homologs of AgrD proteins in Clostridial species.

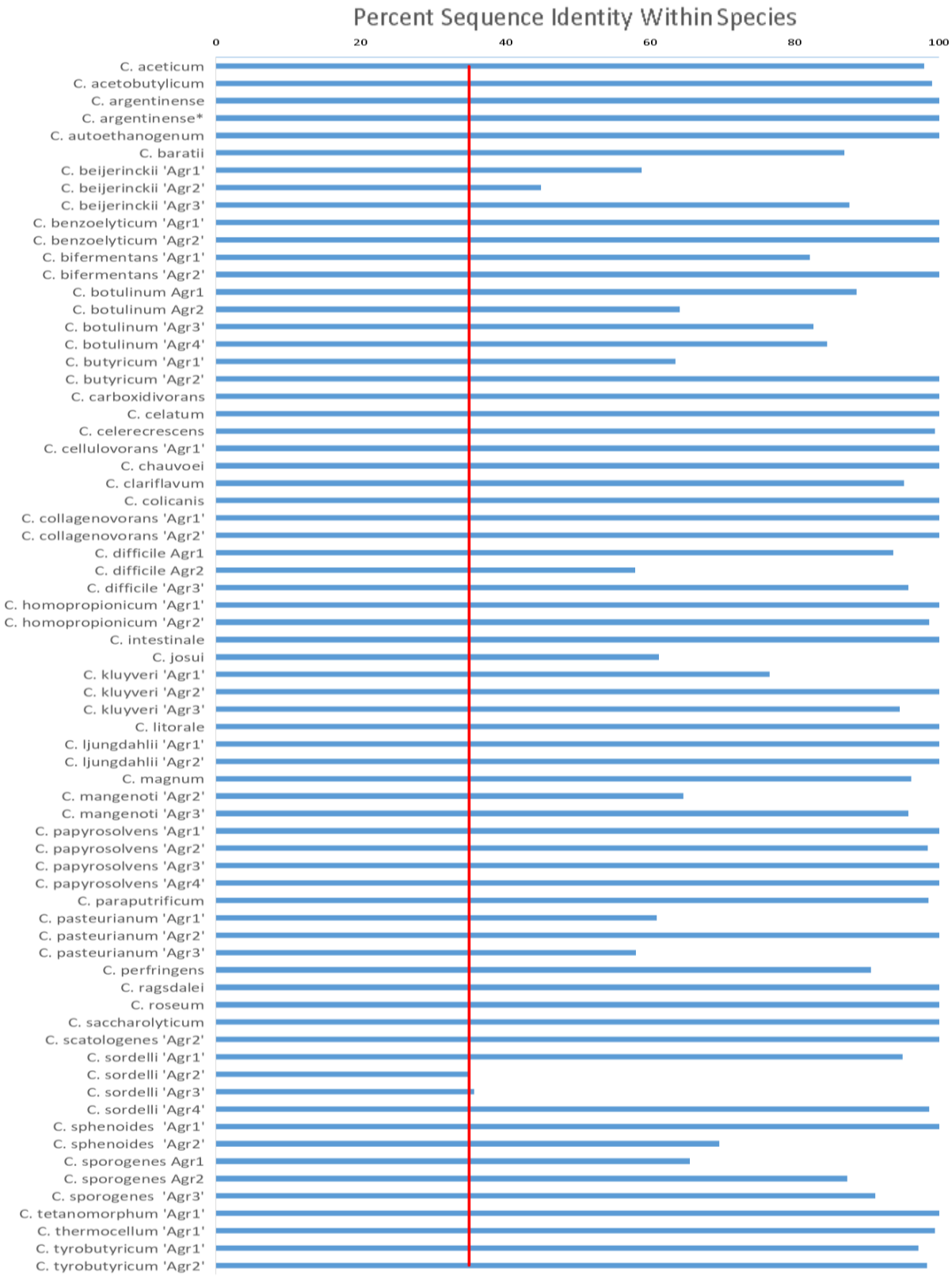

Figure 34: Percent sequence identity of all the homologs of AgrB proteins in Clostridial species.

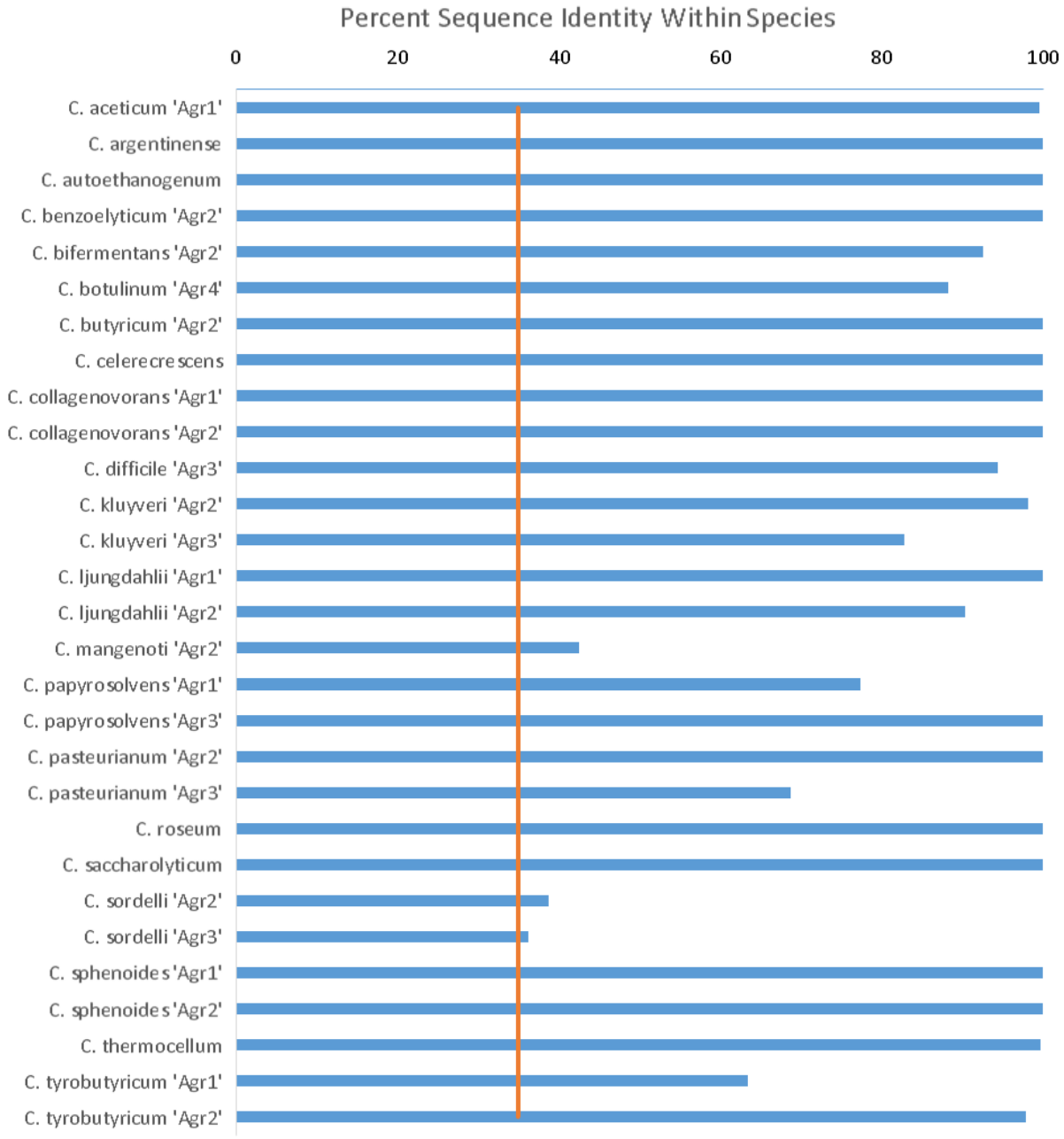

Figure 35: Percent sequence identity of all the homologs of AgrC proteins in Clostridial species.

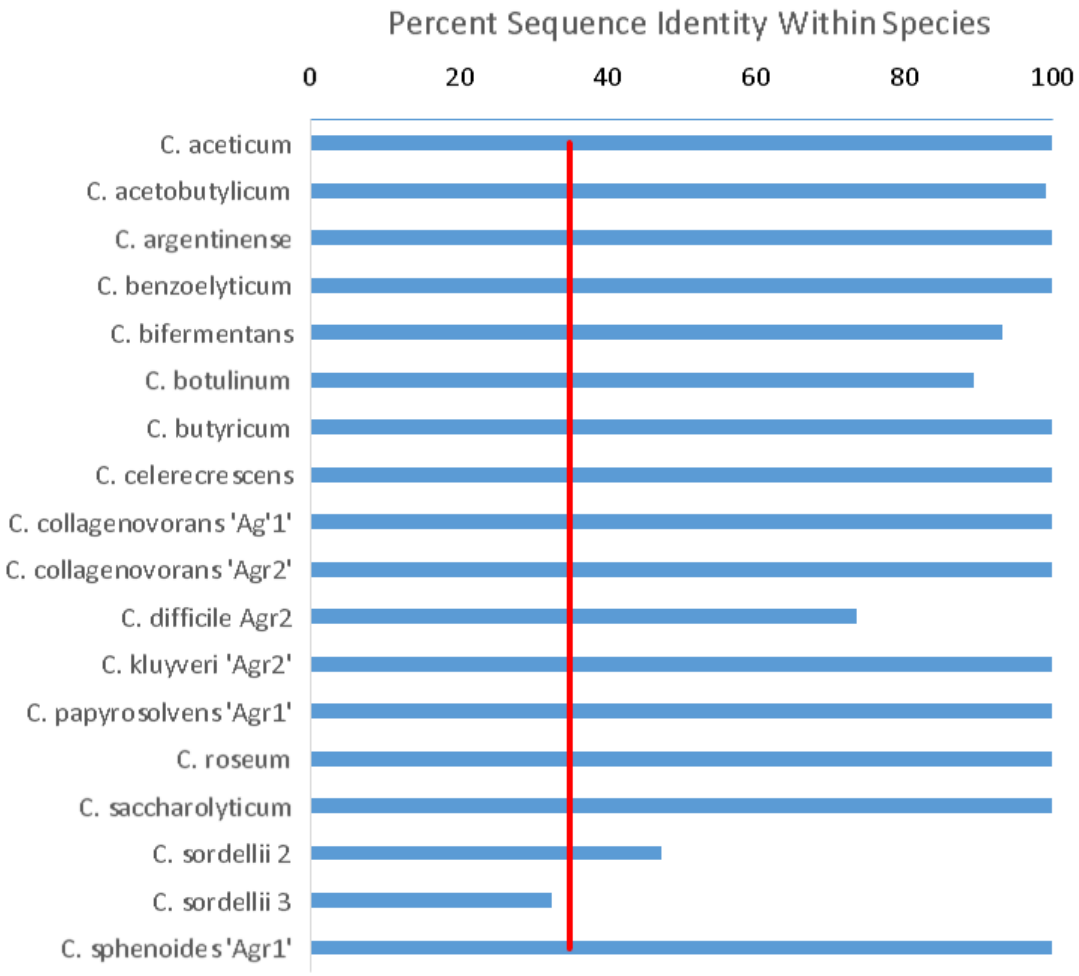

Figure 36: Percent sequence identity of all the homologs of AgrA proteins in Clostridial species.

The Sequence Identity of the Agr Components in *C. botulinum*

Figure 5A shows the alignment of *C. botulinum's* AgrD1, including the sequences of the *S. aureus* AgrDs that confirms the presence of domains and motifs commonly found in AgrD. The domains and motifs present in both species include the Cysteine at position 28, where cyclization happens, the AIP, and a hypothetical amphipathic helix. Additionally, both species have a C-terminus with a significant number of charged residues. However, the C-terminus of *C. botulinum's* AgrD1 has a Tyr33 instead of an Asp33, which is presumed to be the recognition site for AgrB. The different recognition site possibly indicates a different mechanism for *C. botulinum's* AgrB. Furthermore, Glu40 and Leu41 are not conserved in *C. botulinum's* AgrD1, even though they are necessary for AIP production in *S. aureus*, adding to the evidence of a different AgrB mechanism. Another point of contention for *C. botulinum's* AgrD1 is the possible absence of an amphipathic helix. Although there is a hydrophobic face that could tether the helix

Figure 37: (A) Comparative analysis of the *S. aureus* AgrDI-IV and AgrD1 sequences of *C. botulinum* strains. Relevant differences within *C. botulinum* are highlighted in yellow. (B) Wheel diagram mimicking the putative amphipathic helix of *C. botulinum* AgrD1. Color code for residues: yellow, hydrophobic; purple, serine and threonine; blue, basic; pink, asparagine; grey, alanine. The arrow in the helical wheel shows the direction of the hydrophobic moment.

to the membrane as shown in **Figure 5B**, AMPHIPASEEK does not recognize an amphipathic helix within the sequence. The domains within the *C. botulinum* AgrD1 are nearly identical between all strains. Despite the similarity throughout the signal peptide, strains AM533 and B2 450 did have significant differences within the AIP compared to the other strains. These differences might not seem crucial, however, *S. aureus'* AgrDI and IV are different AIPs that are distinguished by one amino acid difference. Considering the case of *S. aureus'* AgrDI and IV, categorizing *C. botulinum* AM533 and B2 450's AgrD1 sequences as a different protein is reasonable. Therefore, the sequences of *C. botulinum's* AgrD1 could be different between different strains.

Similarly, *C. botulinum's* AgrB1 sequences are mostly identical apart from a few significant differences (**Figure 6**). Although they are few, these significant differences are present in regions of *C. botulinum's* AgrB1 sequences that align with functionally-relevant regions of the *S. aureus* AgrB sequences. These functional regions are shown within boxes or in alignment with the coils represented by 'C' at the bottom of the alignments in **Figure 6A**. The coils represent the predicted secondary structure of the *C. botulinum* ATCC 3502 AgrB1. Since the secondary structure of *S. aureus'* AgrB1 is correlated with the location of some functional residues, the coils are a prediction of functional regions of *C. botulinum* AgrB1. There are several differences (shown in red) between *C. botulinum's* AgrB1 and *S. aureus'* AgrB sequences within the functional regions, but none of the functional residues required for AgrB catalytic activity are different. The differences between the species are expected, as they are merely homologs. On the other hand, significant differences within the *C. botulinum* sequences are not expected. Out of the six positions with significant differences, two of them are within the boxed regions and are within strains AM533 and B2 450. Two other positions outside of a functional region has significant differences

within the same strains, and two others within different strains. Similar to AgrD1, there is a possibility that the AgrD and AgrB of strains AM533 and B2 450 are different enough to interact exclusively and be considered different proteins from other sequences in their respective alignments.

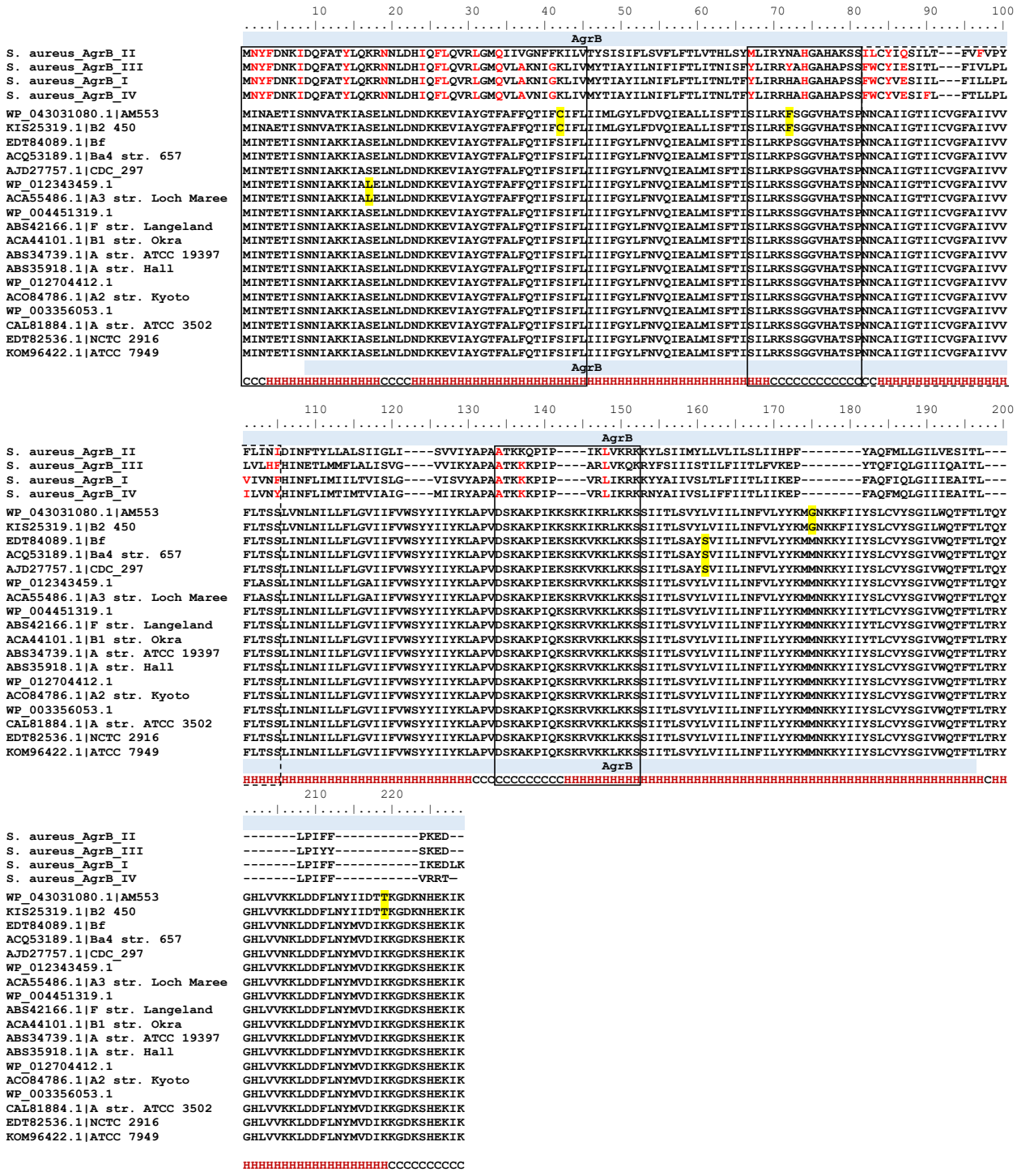

Figure 38: Comparative analysis of the sequences of *S. aureus* AgrBI-IV and AgrB1 sequences of strains of *C. botulinum*. Differences between *S. aureus* and *C. botulinum* are shown in red, whereas differences within *C. botulinum* are highlighted in yellow. Solid boxes highlight functional domains.

In *C. botulinum's* AgrD2 (**Figure 7**), the cyclization at the cysteine residue, AIP, and charged C-terminal are all present. There is also a hydrophobic patch on the helix (**Figure 7B**), but an amphipathic helix is not likely to occur according to the AMPHIPASEEK prediction. Furthermore, residues Asp34 and Glu41 at the C-terminus of the AgrD of *S. aureus* are also absent in *C. botulinum's* AgrD2, but Leu42 is present. Contrasting with *C. botulinum's* Agr1 sequences, yellow highlights and underscores in **Figure 7A** indicate various significant differences between *C. botulinum's* AgrD2 sequences, noticeable in every position except the conserved Cysteine.

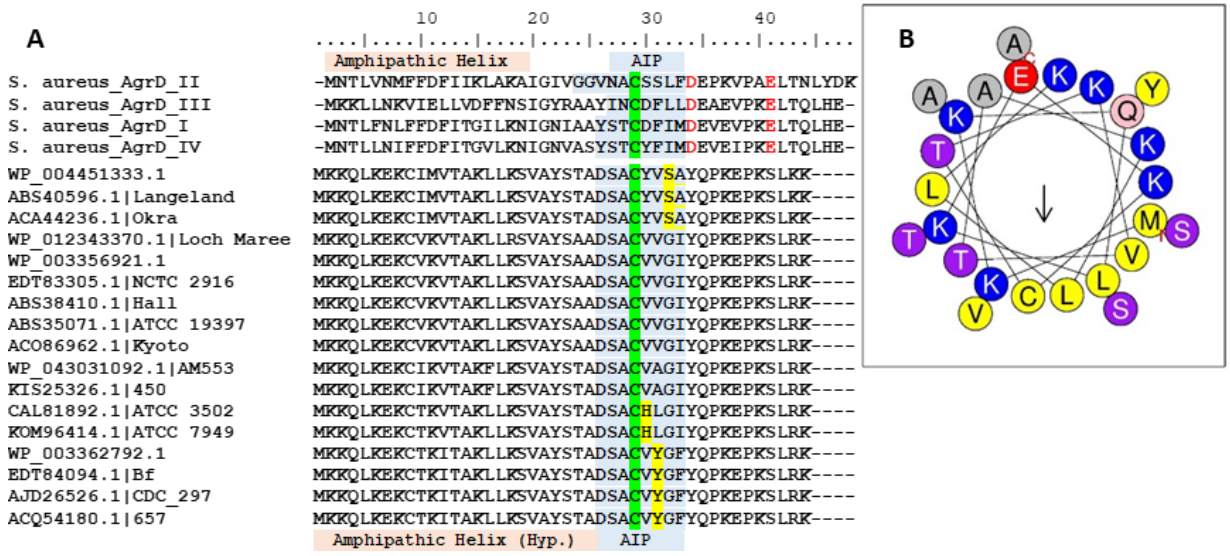

Figure 39: (A) Comparative analysis of the amino acid sequences of *S. aureus* AgrDI-IV and AgrD2 of *C. botulinum* strains. The conserved cysteine-28 is shown in green and the differences between *S. aureus* and *C. botulinum* are shown in red. (B) Wheel diagram mimicking the putative amphipathic helix of *C. botulinum* AgrD.

Therefore, the *C. botulinum* AgrD2 sequences are different between the different strains of the species, even more so than in their AgrD1.

C. botulinum AgrB2 proteins also have higher probability of being different. As shown in the alignment of **Fig. 8**, *C. botulinum's* AgrB2 sequences are different from *S. aureus'* AgrBs, but still have the same functional residues for peptidase activity. As with AgrD2, *C. botulinum's* AgrB2s have a larger number of significant differences around the hypothetical functional regions. However, most of these differences are not present in the strains that have variations in *C. botulinum's* AgrD2. Despite the lack of uniformity between the differences in the sequences of AgrD2 and AgrB2, there are strains that consistently have differences at the same positions. Examples of groups of strains that are different in the positions include AM533 and B2 450; Langeland, Okra, Bf, 657, and CDC_297; ATCC 3502 and ATCC 7949; and Kyoto, Hall, ATCC 19397, and NCTC 2916. Thus, there is a chance that the differences are not completely random and could lead to different categorization from the other sequences.

Due to the significant differences appearing consistently within the same strains in the sequences of both Agr components, it is plausible that the proteins are not the same within *C. botulinum*. *C. botulinum's* Agr1 sequences do have positions with significant differences within the same strains in both Agr components, creating a stronger argument for the different proteins. *C. botulinum's* Agr2 sequences also have locations with significant differences within the same strains. However, they do not vary within the same strains in both AgrD2 and B2 components. Due to the inconsistency in the strains' differences across Agr2 proteins, one might argue that Agr1 is more likely to have different proteins. However, *C. botulinum's* Agr2 components have more significant differences than Agr1.

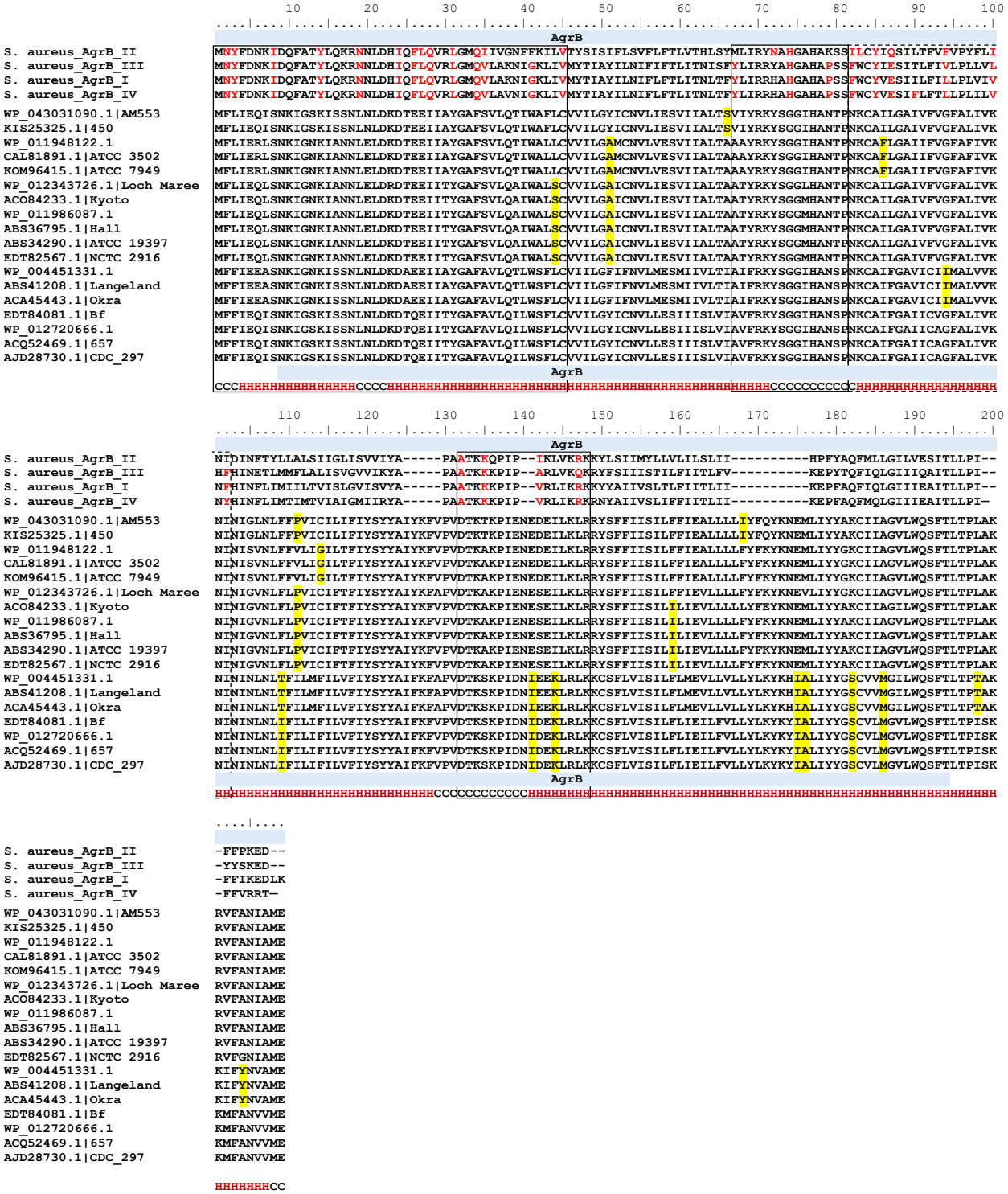

Figure 40: Comparative analysis of the amino acid sequences of *S. aureus* AgrBI-IV and AgrB2 of *C. botulinum* strains. Differences between *S. aureus* and *C. botulinum* are shown in red while differences within *C. botulinum* are highlighted in yellow. Solid boxes highlight functionally-relevant regions in *S. aureus*, including extracellular portions of AgrB-I (residues 1-45, and 132-148), and an intracellular loop (67-81). Dashed boxes show an extended region containing residues important for function in *S. aureus'* AgrB-I.

Sequence Identity in the Agr Components of *C. difficile*

The alignment of the AgrD2 of *C. difficile* is shown in **Figure 9A** in comparison to the AgrD alleles of *S. aureus*. The domains and motifs present in both species include the hypothetical amphipathic helix, confirmed by both AMPHIPASEEK (not shown) and the helix wheel in **Figure 9B**, and the charged C-terminal. However, the C-terminal of the *C. difficile* AgrD2 has a His33 instead of an Asp33, changing the presumed recognition site for AgrB2 and possibly indicating a different mechanism from that of *S. aureus* and *C. botulinum*. Furthermore, Glu40 is not conserved in *C. difficile's* AgrD2, while Leu41 is conserved. This difference indicates an alternative mechanism for AIP production in *C. difficile* AgrB2. Other factors that distinguishes *C. difficile* AgrD2 are the short tailless AIPs predicted by the bioinformatics tools and the lack of the Cysteine. The cysteine is replaced by a serine, which is found in other AIPs of different species (Thoendel & Horswill, 2009).

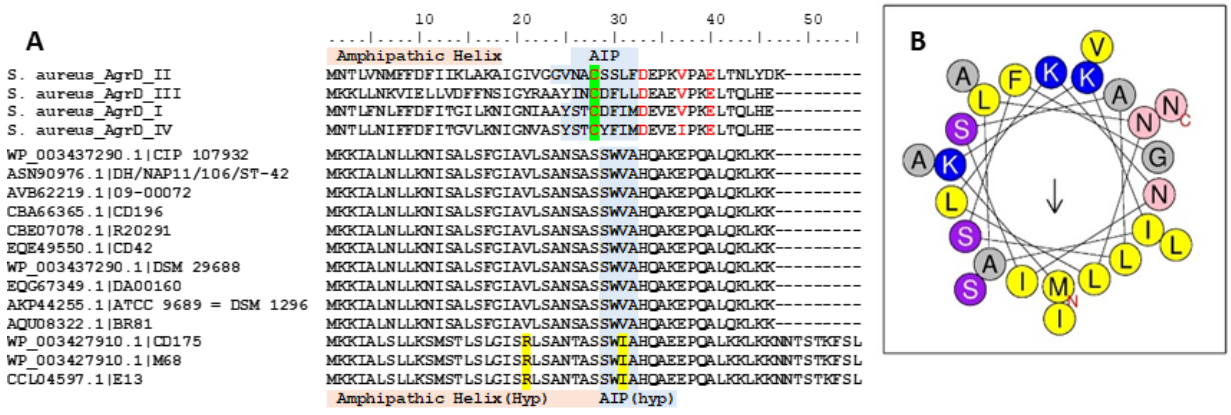

Figure 41: (A) Comparative analysis of the amino acid sequences of *S. aureus* AgrDI-IV and AgrD2 of *C. difficile* strains. Relevant differences between *S. aureus* and *C. difficile* are in red font. (B) Wheel diagram mimicking the putative amphipathic helix of *C. difficile* AgrD.

The *C. difficile* AgrD2 sequences are identical between all strains apart from strains CD175, M68, and E13. The significant differences within these strains are present in the same positions of the hypothetical amphipathic helix and the AIP. The Arg21 in the amphipathic helix might not affect the function of the signal peptide but might affect the interaction with the second peptidase that releases the AIP from the membrane, as seen in *S. aureus*. Additionally, the Val31 substitution for Ile31 might not have a great effect on the interaction between AgrD2 and AgrC2, as they are both hydrophobic and favored substitutes for each other. However, the recognition interaction with AgrC is sensitive, and the difference in bulkiness from valine to isoleucine might be enough to alter structure and function.

Due to the few significant differences in the *C. difficile* AgrD2, the AgrB2 is not expected to be different among the strains, yet, the level of differences between them was found to be high as shown in **Figure 10**. Even more interesting is that all but one of the different positions vary similarly among the same three strains (CD175, M68, and E13). The majority of these differences occur outside of the boxed areas, possibly reducing the functional significance of these differences. Nevertheless, the level of variation is significant and suggests the existent of a different Agr operon than the more prevalent version of *C. difficile* Agr2.

Further evidence for the hypothesis of another variant of the Agr2 operon is the differences within the *C. difficile* AgrC2 alleles. All the differences occur at the same position in strains M68 and E13, two of the strains that were consistently different in AgrD2 and agrB2. The boxes and coils represent hypothetical functional regions, specifically the extracellular loops of AgrC in the transmembrane sensor domain. In *S. aureus*, the first loop harbors residues involved in activation of AgrC, and the second and third loops have residues responsible for specificity to AgrD. All three loops show congruent differences only within the *C. difficile* M68 and E13 strains.

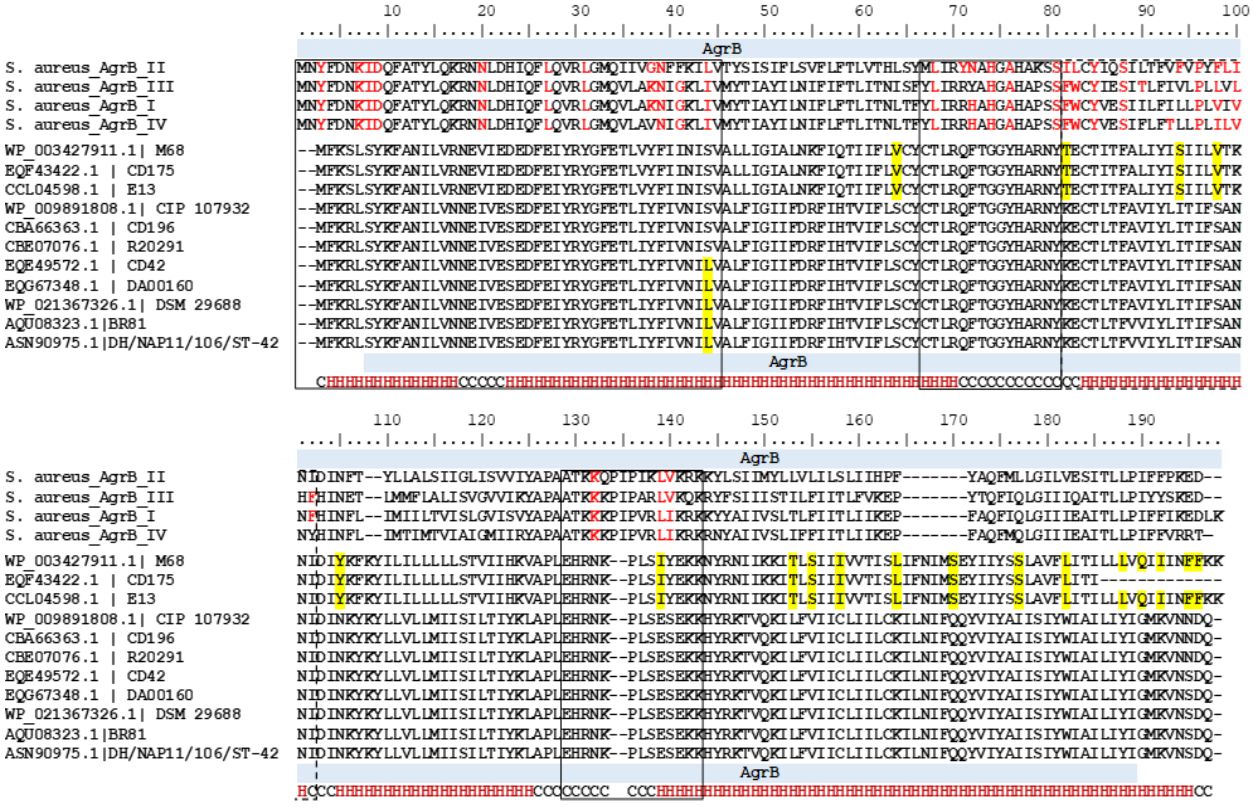

Figure 42: Comparative analysis of the amino acid sequences of *S. aureus* AgrBI-IV and AgrB2 of *C. difficile* strains. Differences between *S. aureus* and *C. difficile* are shown in red while differences within *C. difficile* are highlighted in yellow. Solid boxes highlight functionally-relevant regions in *S. aureus*, including extracellular portions of AgrB-I (residues 1-45, and 132-148), and an intracellular loop (67-81). Dashed boxes show an extended region containing residues important for function in *S. aureus'* AgrB-I.

The same pattern extends over to the catalytic regions of the dimerization and histidine phosphotransfer domain and the catalytic and ATP-binding domain. **Figure 11** shows an alignment of the hypothetical catalytic regions of the *C. difficile* AgrC2 with the H-box, N- box, and G-box of *S. aureus* AgrCI. These catalytic regions show significant differences between *C. difficile* sequences, suggesting different interactions with AgrA. Generally, however, the catalytic residues of these three catalytic boxes are conserved, indicating conserved function. The same catalytic residues are also conserved between *C. difficile* AgrC2 alleles and that of *S. aureus*. However, the sensor domain is not conserved, as the function of the domain is very specific to the AgrD variants it interacts with. Given that the AgrA proteins in each species interacts with their cognate histidine kinase and different nucleotides, the AgrA sequences are different between *C. difficile* and *S. aureus*.

Figure 12 shows the differences in *C. difficile* AgrA2, suggesting a similar pattern in the strains E13, CD175, and M68. There are only three positions with significant differences, and they are all within the recognition domain (REC) that interacts with AgrC. If there are no significant differences within the LytTR regulatory domain, then *C. difficile* AgrA can only interact with one promoter. Given that the AgrA of the three different strains may be binding to the same promoter as the rest of the strains, both would be upregulating the production of the same operon. Therefore, there are two immediately plausible situations assuming both operons function normally: either the promoters for both operons are the same, or the feedback loop would not be complete for the Agr system of E13, CD175, and M68. In the latter case, there would have to be a different step in the mechanism where another Agr component cross-interacts between both possible systems. In reference to the minimal differences between the AIPs of *C. difficile*, both versions of the AIP could interact with one or both of the AgrBs to provide the feedback loop for the Agr system of

E13, CD175, and M68. Although there is evidence for the existence of variants within the *C. difficile* Agr2 system, the lack of differences in the regulatory domain of AgrA might confirm that they are all the same protein.

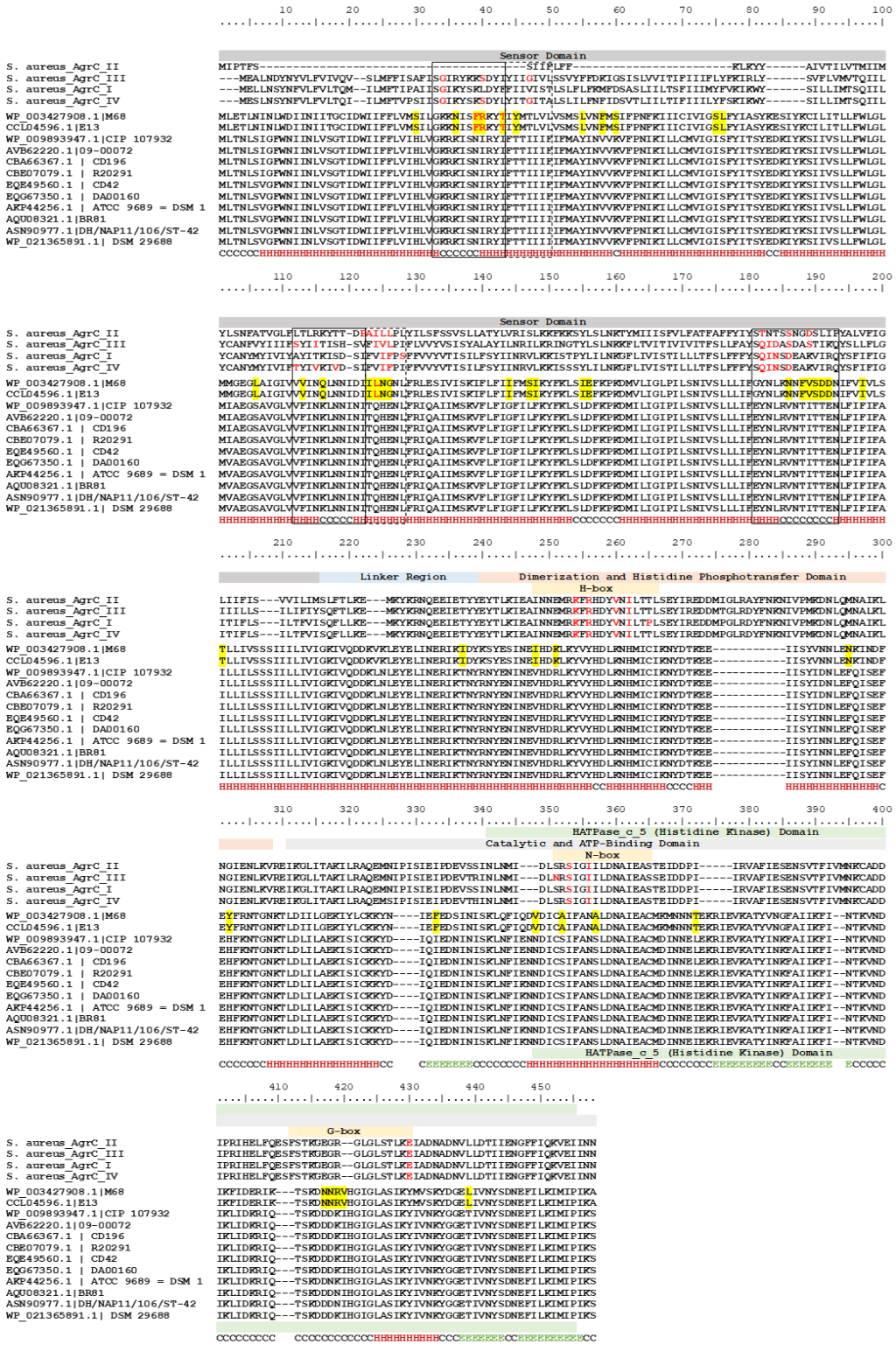

Figure 43: Comparative analysis of the amino acid sequences of *S. aureus* AgrCI-IV and AgrC2 of *C. difficile* strains. Differences between *S. aureus* and *C. difficile* are shown in red while differences within *C. difficile* are highlighted in yellow. Solid boxes highlight functionally relevant regions in *S. aureus*, including extracellular portions of AgrC-I (residues 29-39, 104-113 and 178-190). Dashed boxes show an extended region that contains functionally relevant residues of *S. aureus'* AgrC-I.

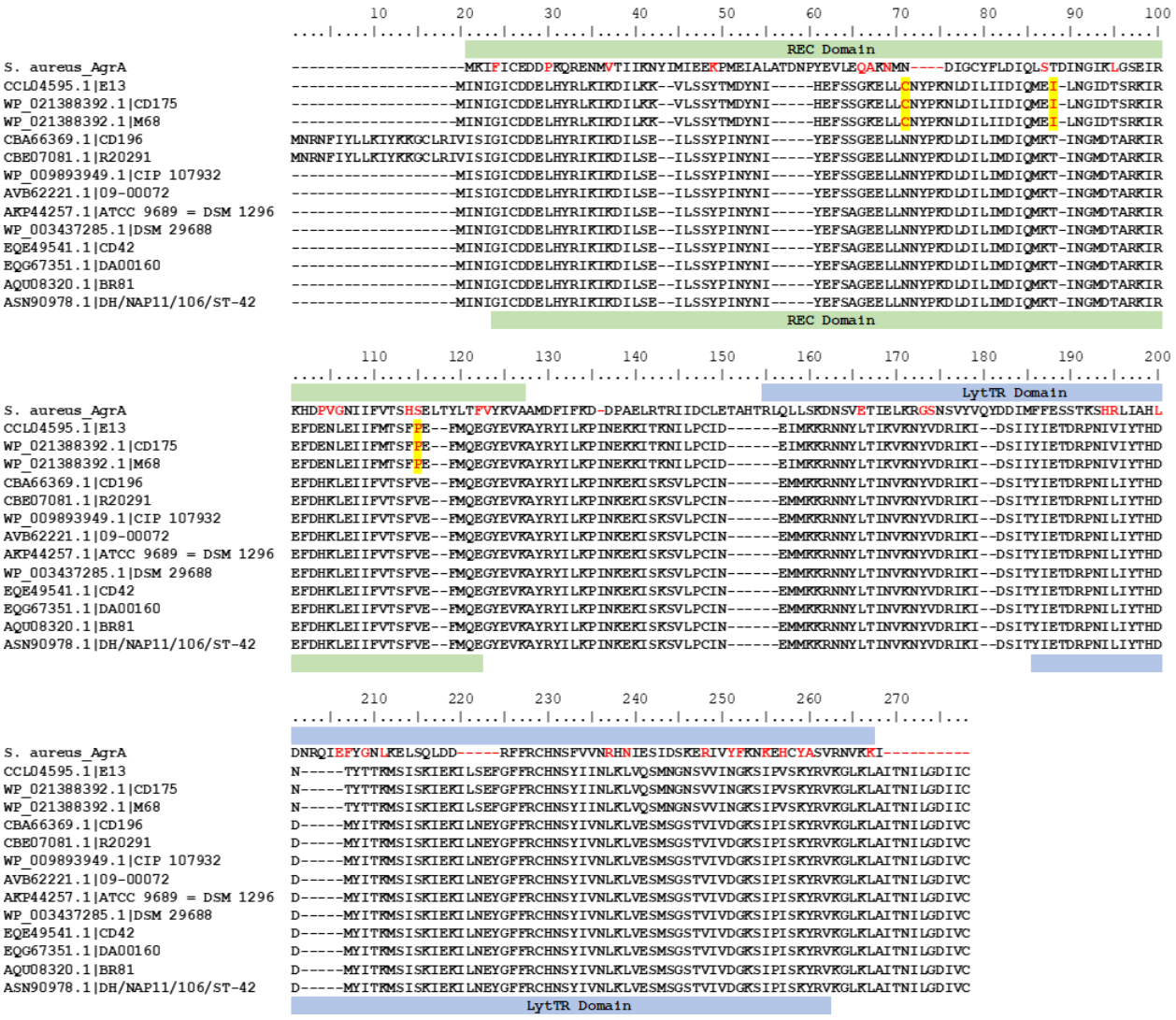

Figure 44: Comparative analysis of the amino acid sequences of *S. aureus* AgrA and AgrA2 of *C. difficile* strains. Differences between *S. aureus* and *C. difficile* are shown in red while differences within *C. difficile* are highlighted in yellow.

Sequence Identity in the Agr Components of *C. sporogenes*

The *C. sporogenes* Agr components are very similar to that of *C. botulinum*. The *C. sporogenes* AgrD1, like the *S. aureus* AgrDs, has a charged C-terminus, cyclization cysteine, and an AIP (**Figure 13A**). While the C-terminus has the charged amino acids in *C. sporogenes* AgrD1, the Asp34 is not conserved. Conversely, Glu41 and Leu42 are present. Instead of Asp34, the *C. Sporogenes* AgrD sequences have Tyr34, similar to the *C. botulinum* AgrD. Another motif that differs from the *S. aureus* AgrD is the amphipathic helix. While AMPHIPASEEK predicts a low likelihood of formation of an amphipathic helix, it likely forms a hydrophobic face (**Figure 13B**). However, only an experimental approach will be able to determine the presence of an amphipathic helix. An experimental approach will also be necessary to determine potential functional differences within the AIPs of *C. sporogenes*. The AIPs contain significant differences in every position of the macrocycle apart from the conserved cysteine (shown in yellow in **Figure 13A**). The residues at position 31 are different among the strains.

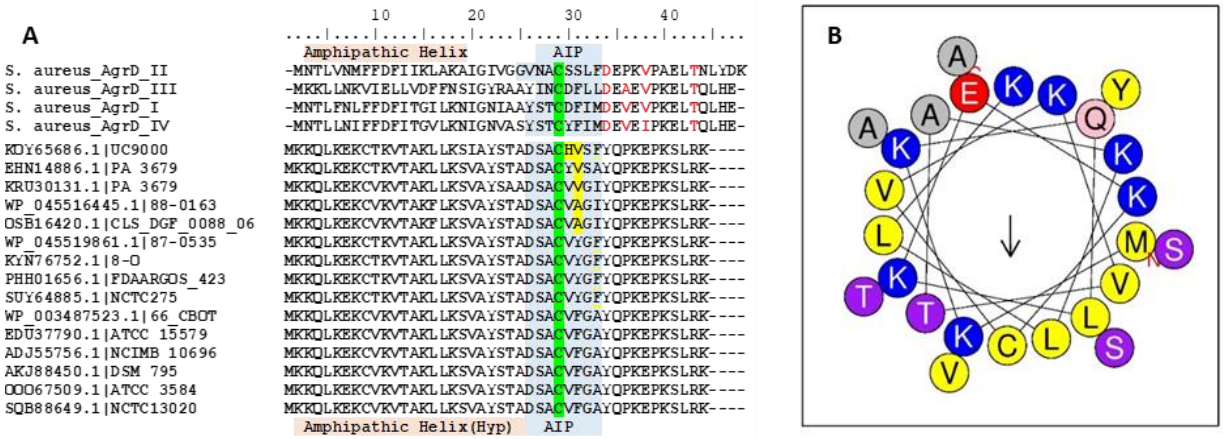

Figure 45: (A) Comparative analysis of the *S. aureus* AgrDI-IV and AgrD1 sequences of *C. sporogenes* strains. Relevant differences within *C. sporogenes* are highlighted in yellow. (B) Wheel diagram mimicking the putative amphipathic helix of *C. sporogenes* AgrD.

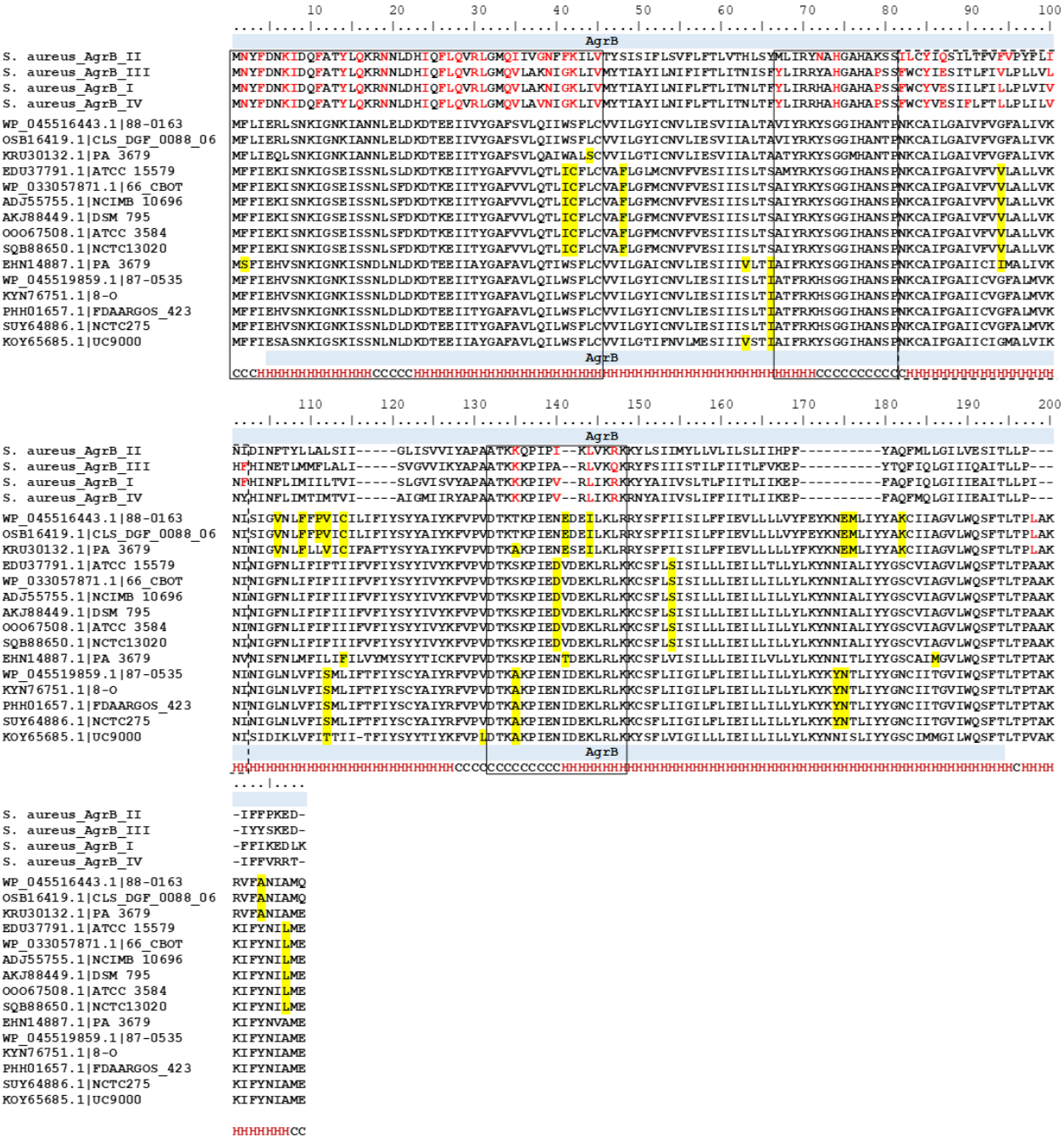

Figure 46: Comparative analysis of the sequences of *S. aureus* AgrBI-IV and AgrB1 sequences of strains of *C. sporogenes*. Differences between *S. aureus* and *C. sporogenes* are shown in red, whereas differences within *C. sporogenes* are highlighted in yellow. Solid boxes highlight functional regions in *S. aureus*, including extracellular portions of AgrB-I (residues 1-45, and 134-152), and an intracellular loop (67-81). Dashed boxes show an extended region that contains functional residues of *S. aureus'* AgrB-I.

The variety of the AIPs expected from the *C. sporogenes* AgrD1 sequences is reflected in the AgrB1 as well. The *C. sporogenes* AgrB1 is significantly different from *S. aureus* AgrBs. The boxed hypothetic functional regions show significant differences, as well as the regions outside of the box (**Figure 14**). Despite the differences, the catalytic residues are still present, suggesting a conserved function. In contrast to catalysis, the specific interactions between the *C. sporogenes* AgrD1 and agrB1 might not be the same throughout the strains due to the differences that are consistent within same positions. The *C. sporogenes* strains that have differences at the same position within the AgrD1 and AgrB1 include PA 3679, 88-0163, and CLS_DGF_0088_06; 87-0535, 8-O, FDAARGOS_423, and NCTC275; and ATCC 15579, 66_*C. botulinum*OT, NCIMB 10696, DSM 795, ATCC 3584, and NCTC13020. The significant differences in the sequences of AgrD1 and AgrB1 suggest the Agr1 operon of *C. sporogenes* is different.

The sequences of AgrD2 of *C. sporogenes* show fewer differences than in the AgrD1 (**Figure 15A**). The AIP found in AgrD2 has only one difference and the general motif of the cyclization cysteine is present. The charged C-terminal is also present in the AgrD2, even though AgrD2 is missing the Asp33, Glu40, and Leu41 that are present in *S. auerus*. These residue-specific differences indicate a possible change in AgrB mechanism from *S. aureus* to Clostridial species. The presence of an amphipathic helix could also be a contrast between *S. aureus* and Clostridia, as AMPHIPASEEK predicted that a helix does not exist in the *C. sporogenes* AgrD2, even when the helix wheel in **Figure 15B** shows otherwise.

The sequences of *C. sporogenes* AgrB2 do not show many differences. The differences present are random and only two out of the six varying positions are within hypothetic functional regions. Thus, it appears the sequences of the Agr2 components are similar.

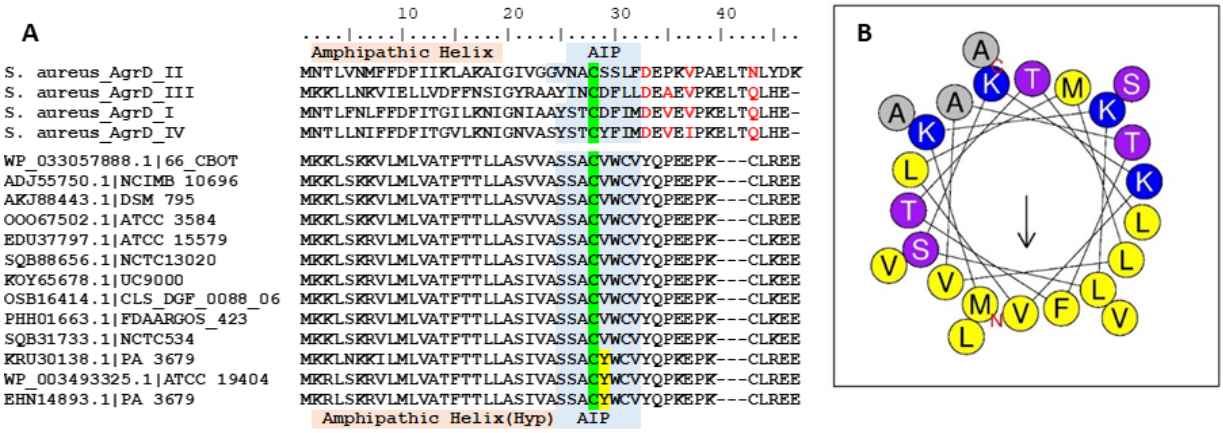

Figure 47: (A) Comparative analysis of the Staphylococcus aureus AgrDI-IV and AgrD2 sequences of *C. sporogenes* strains. Relevant differences within *C. sporogenes* are highlighted in yellow. (B) Wheel diagram mimicking the putative amphipathic helix of *C. sporogenes* AgrB1.

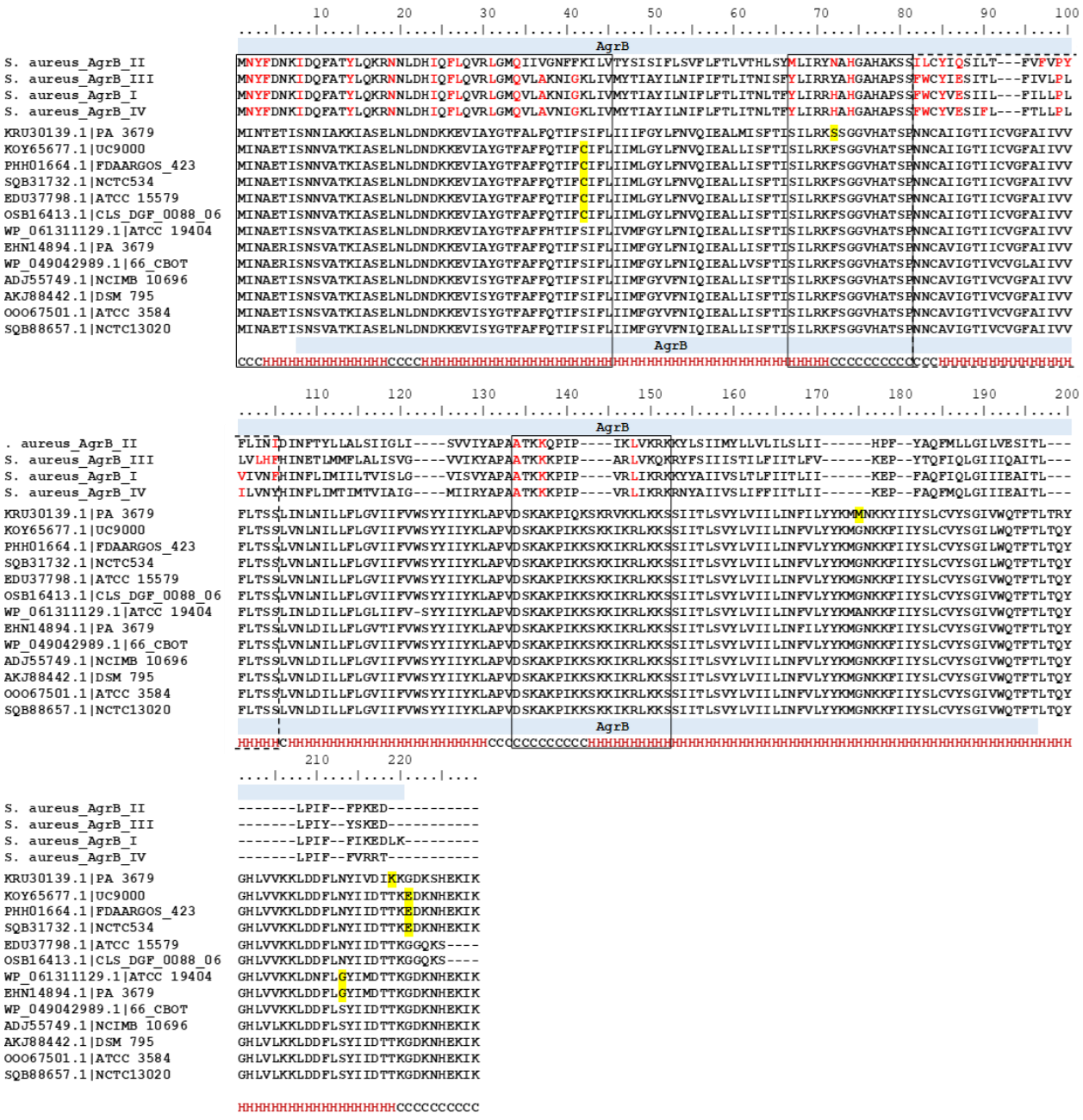

Figure 48: Comparative analysis of the sequences of S. aureus AgrBI-IV and AgrB2 sequences of strains of *C. sporogenes*. Differences between *S. aureus* and *C. sporogenes* are shown in red, whereas differences within *C. sporogenes* are highlighted in yellow. Solid boxes highlight functional regions in *S. aureus*, including extracellular portions of AgrB-I (residues 1-45, and 134-152), and an intracellular loop (67-81). Dashed boxes show an extended region that contains functional residues of *S. aureus'* AgrB-I.

Sequence Identity in the Agr Components Between Clostridial Species

Comparison of the AgrD Sequences between the five Clostridial Species

Although they all have the same domain, the AgrDs of Clostridial species are different from each other. **Figure 17** is an alignment of the AgrD showing the differences. Some of the species do not have a tail on their AIPs and their macrocycles are completely different. In addition, the cysteine residue is the most conserved residue of the macrocycles, but AgrD2 of *C. difficile* has a serine in that position that most likely cyclizes into a lactone. The specific residue differences

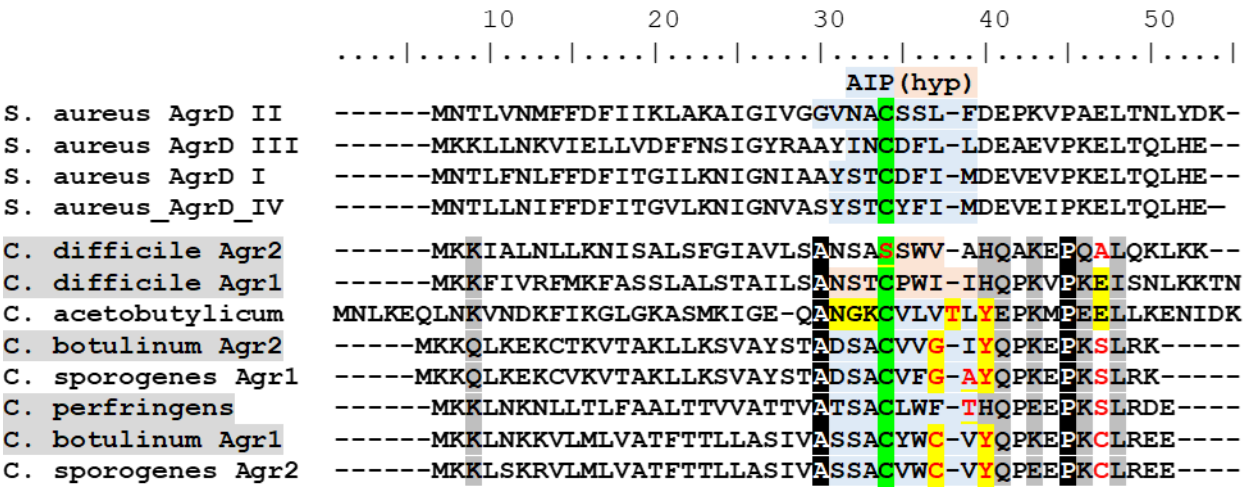

Figure 49: Comparative analysis of the *S. aureus* AgrDI-IV and AgrD consensus sequences of Clostridium species with quorum-sensing Agr components. Relevant differences between *S. aureus* and Clostridium species are shown in red, whereas differences between Clostridium species are highlighted in yellow. Black and grey highlighting of amino acids indicates full conservation and similar residues, respectively. The grey shading of the species name indicates -pathogenic or toxigenic. The light blue shading shows empirically proven autoinducer peptides (AIPs) and the orange shading shows predicted AIPs based on bioinformatics analyses through SignalP, Predisi, and Phobius.

in the macrocycle include variations from the hydrophobic and bulky residues that normally populate the last two positions of the AIPs of *S. aureus*. Although the other residues in the Clostridial AIPs are hydrophobic and bulky, the small, polar, and potentially catalytic residues in Clostridia are still different between the species.

Other motifs also show differences in the AgrDs of Clostridial species, for example, the cleavage site recognized by SpsB, which includes the three residues preceding the AIP and a conserved glycine or proline at position -5 or -6 from the AIP. Interestingly, the glycine and proline, which are thought to present the cleavage site to the peptidase, are absent. Nevertheless, there are conserved alanine residues at positions 26 and 30 that could fit the description. Downstream, the cleavage site is usually a variation of an A–X–A motif, but with significant wobble to the residues as shown in yellow highlight. The C-terminal recognition site for AgrB, where the *S. aureus* Asp40 aligns right after the AIP (**Figure 17**), is not exactly conserved amongst Clostridial species. The Glu47 that is essential for AIP production in *S. aureus* is also not conserved in Clostridia. The lack of conservation within these residues' hints at differences in mechanism between AgrBs of Clostridia. The differences between the AIPs of the species are expected as they are specific molecules that have sensitivity and specific binding action.

On the other hand, the similarities between the AgrD of Clostridia could make it easier to target therapeutically. This is because a single drug could be used to target the Agr system in multiple Clostridia. As shown in **Fig. 17,** the proline residues at positions 42 and 45 have specific function and structure that could provide a target for regulation within Clostridia (grey shading in **Fig. 17**). Similarly, if one of the alanine residues conserved at the N-terminal positions 26 and 30 were found to function as the cleavage site for SpsB, these alanine residues could be another target for exclusive modulation of Clostridial regulatory pathways.

Comparison of the AgrB Sequences between the five Clostridial Species

The consensus AgrB sequences of Clostridial species (**Figure 18**) show conservation of catalytic residues and significant differences in the boxed regions. The significant differences occur between Clostridia and between Clostridial species and *S. aureus*. The first 34 residues that are conserved and necessary in the *S. aureus* AgrBI show differences even between the same species of Clostridia. A specific residue, Gln38, when mutated to Pro38 in *S. aureus* led to the destabilization of the protein; the AgrBs of *C. botulinum* Agr2, *C. sporogenes* Agr1, *C. perfringens*, *C. botulinum* Agr2, and *C. sporogenes* Agr2 show an aromatic residue at that position. Another specific residue in the vicinity, Asn43, when mutated to Ile43 or Tyr43 in *S. aureus* led to loss of peptidase activity (Thoendel & Horswill, 2013). Ile43 is present in *C. sporogenes* Agr1 and Phe43 in both *C. difficile* AgrBs. Although these mutations probably do not hinder the AgrBs of the Clostridial species, they do indicate that the proteins are likely different and that the positions might not be as crucial in the Clostridium genus.

The other boxed regions (solid and dashed in **Figure 18**) include the catalytic residues of His81 and Cys88 and show differences amongst *C. difficile* AgrBs. Additionally, the two *C. difficile* AgrBs are the only ones with a significant difference at an experimentally tested position, Thr142, which if mutated to Ile142 in the *S. aureus* AgrBI would abolish peptidase activity (Thoendel & Horswill, 2013). Therefore, *C. difficile* might have a significantly different AgrB

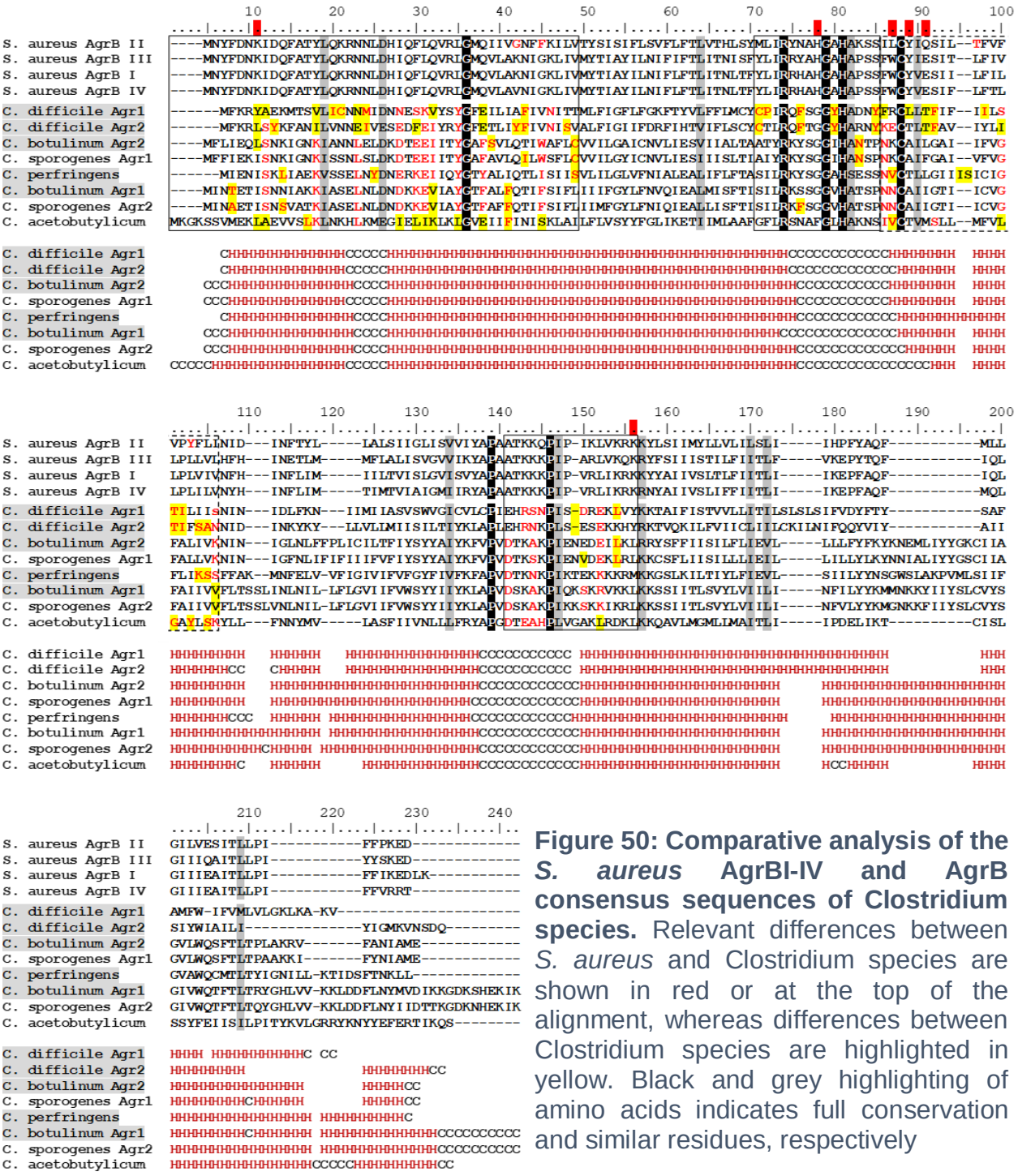

Figure 50: Comparative analysis of the *S. aureus* AgrBI-IV and AgrB consensus sequences of Clostridium species. Relevant differences between *S. aureus* and Clostridium species are shown in red or at the top of the alignment, whereas differences between Clostridium species are highlighted in yellow. Black and grey highlighting of amino acids indicates full conservation and similar residues, respectively

respectively. The grey shading of the species' name indicates pathogenicity or toxigenicity. Solid boxes highlight functional regions in *S. aureus*, including extracellular portions of AgrB-I (residues 1-49, and 141-156), and an intracellular loop (71-85). Dashed boxes show an extended region that contains functional residues of the *S. aureus* AgrB-I. Below the amino acid alignment is the alignment of the secondary structure of the AgrB of each species presented in the previous figures.

compared to other Clostridial AgrB components. Continuing downstream, a lysine patch in positions 143-5 of the *S. aureus* AgrBs was found to be crucial for secretion of the cleaved AIP. The following mutations, Lys145Glu, Lys143Gln, Lys144Gln, or Lys145Gln, abolished secretion of the cleaved AIP (Thoendel & Horswill, 2013). Various inconsistent mutations are present across all of these positions, suggesting different AgrB processing mechanisms across Clostridial species.

Despite the significant differences between Clostridial AgrBs and between Clostridial and *S. aureus* AgrBs, there are a few conserved residues at positions 36, 74, 79, 139, and 146. Out of these residues, Gly36 stabilizes the *S. aureus* AgrB, Arg74 and Gly79 are known as necessary for AIP production in *S. aureus*, and Pro139 is necessary for AgrB cleavage activity (Zhang et al., 2002; Thoendel & Horswill, 2013). Interestingly, Pro146 is not known to have a specific function in the *S. aureus* AgrBs but could be involved in producing a specific shape for the interacting coiled-coil region alongside Pro139. Furthermore, PSIPRED program predicted that the secondary structures of all AgrBs are similar (**Figure 18)**. The conserved residues and secondary structure could establish homology between the proteins, but the residue-based analysis above shows lack of significant similarity between the proteins. Given the significant differences in their amino acids, the proteins are different between and within Clostridial species.

Comparison of the AgrC Sequences between the five Clostridial Species

The AgrC sequences of Clostridial species are homologs, as they contain the catalytic residues of the histidine kinase, but within the hypothetically functional regions there are definite differences between the *C. acetobutylicum* and *C. difficile* AgrC sequences. These regions include the AgrD sensing and binding specificity regions, the AgrC activation region, binding sites for ATP and AgrA, and residues responsible for protein structure stability.

The sensor domain's three functional extracellular loops are shown in **Figure 19** as solid boxes, while dashed boxes surround buried regions with functional residues. All boxes have significant differences and most of the differences are dissimilar between *C. acetobutylicum* and *C. difficile*. The transmembrane domains between the boxes have few similar residues (shaded gray) and one fully conserved lysine between the AgrCs of *S. aureus*, *C. acetobutylicum*, and *C. difficile*. Although these similar residues appear within the membrane, they could still make a significant functional difference within the protein as other residues have been shown to affect the protein from within membranes (Thoendel & Horswill, 2013). Considering the activation and specificity properties of the sensor domain region, the differences observed are expected, and similarities could be further investigated for specific functions in the histidine kinase.

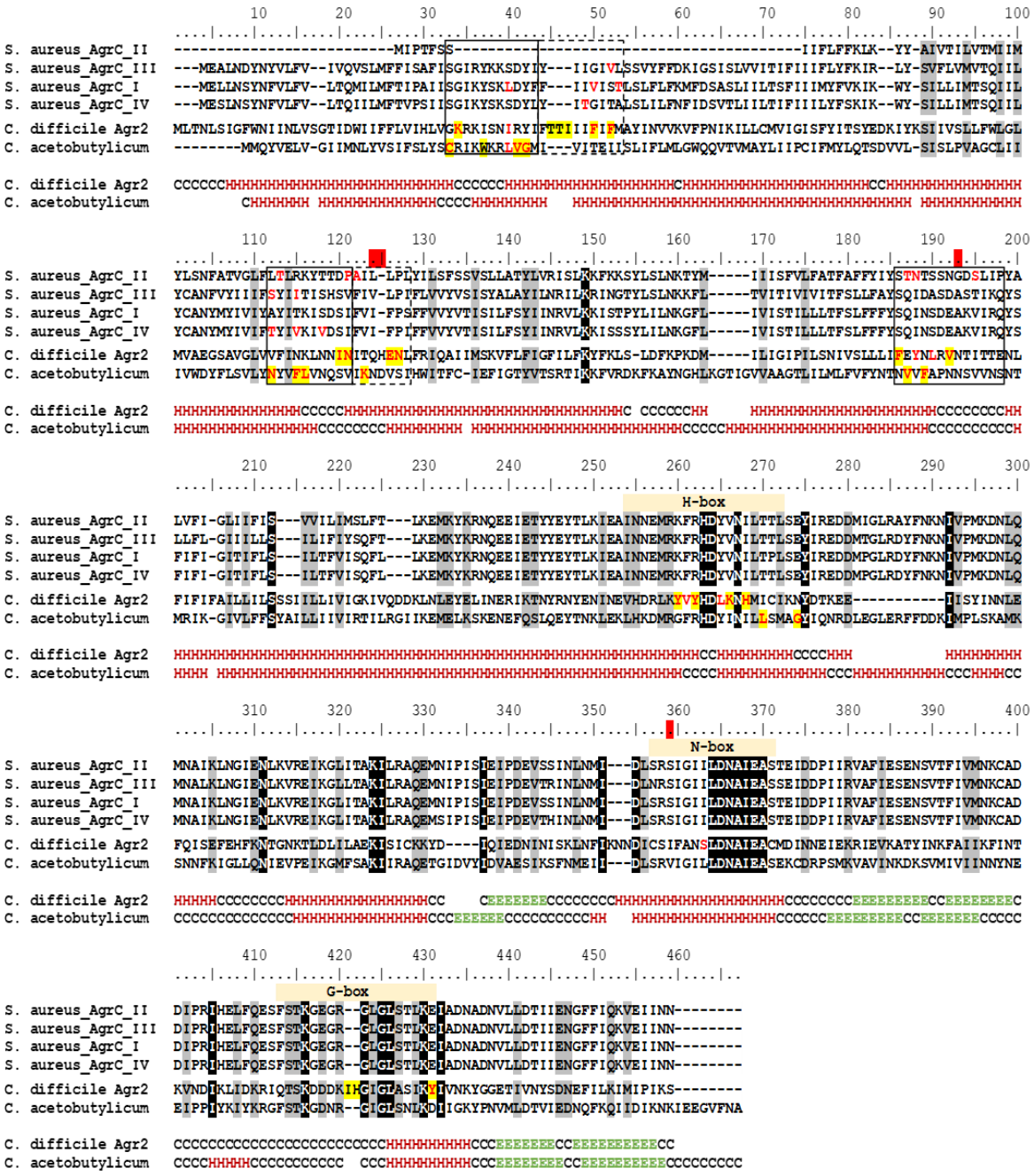

Figure 51: Comparative analysis of the S. aureus AgrCI-IV and AgrC consensus sequences of Clostridium species with quorum-sensing Agr components. Relevant differences between *S. aureus* and Clostridium species are shown in red font or at the top of the alignment, whereas differences between Clostridium species are highlighted in yellow. Black and grey highlighting of amino acids indicates full conservation and similar residues, respectively. The grey shading of the species' name indicates a pathogenicity or toxigenicity. Solid boxes highlight functional regions in SA, including extracellular portions of AgrC-I (residues 33-43, 112-121 and 186-198). Dashed boxes show an extended region that contains functionally relevant residues of SA's AgrC-I. Below the amino acid alignment is the alignment of the secondary structure of the AgrB of each species presented in the previous figures.

Continuing downstream, there are also a significant number of conserved residues that appear within the end of the *S. aureus* AgrCI's last transmembrane segment and the linker to the dimerization and histidine phosphotransfer domain (positions 207-245). These partially conserved residues might be necessary to maintain the shape and orientation of the helix to allow for proper sequestration and exposure of the ATP-binding domain. Supporting this point is the destabilization of *S. aureus* AgrC's interaction with AgrA when Tyr247 is substituted for Cys247 (Norrby-Teglund et al., 2016). While *C. acetobutylicum* has Tyr247, *C. difficile* has Asn247 that possibly implies a different mechanism for the *C. difficile* AgrC. In addition to the dimerization and histidine phosphotransfer domain, the second part of the protein also holds the catalytic and ATP-binding domains. These two domains have three functional boxes, including the H-box, where phosphotransfer occurs, and the two boxes that shape the ATP-binding cleft, N-box and G-box. The sections of the sequences of the *C. acetobutylicum* and *C. difficile* AgrCs that align with the *S. aureus* catalytic boxes all show significant differences. The H-box, containing the phosphoryl acceptor motif F[RK]HDYXN, shows significant variation from the *C. difficile* AgrC2 to *C. acetobutylicum* AgrC, which is almost identical to the *S. aureus* motif. The same box also has residues that interact with AgrA between positions 266 and 275. The other two functional boxes are similar between the Clostridial species. Another similarity between the sequences lies in the predicted secondary structure of Clostridial AgrCs. Loops in the transmembrane sensor domain are reasonably aligned, as are the beta-sheets and helices that form the dimerization and histidine phosphotransfer domain and the catalytic and ATP-binding domain.

As the AgrC is a histidine kinase commonly found in two-component regulatory systems, more similarities are expected between the AgrCs of the Clostridial species and even between the Clostridial species and *S. aureus*. Therefore, the nature of the histidine kinase combined with the

homology of the proteins explains the similar secondary structure and similarities among the residues in the functional regions. Despite the similarities between the Clostridial AgrCs, their activation, sensory, and phosphotransfer regions may have differences that distinguish them.

Comparison of the AgrA Sequences between the five Clostridial Species

AgrA is also part of the two-component regulatory system where it promotes the expression of the Agr system and the RNA that will further regulate cellular functions. The AgrA of *C. acetobutylicum* and *C. difficile* contain all the catalytic residues necessary for function and are the components with the most similarities among all the Agr components. The majority of the conserved residues (**Figure 20)** are present within the recognition (REC) domain that spans positions 1-103 and interacts with AgrC. The differences between the *C. acetobutylicum* and *C. difficile* AgrA components appear mostly in the LytTR domain that binds DNA.

One of the differing sites between the *C. acetobutylicum* and *C. difficile* AgrAs includes the intermolecular recognition motif located at positions 111 and 112. While the AgrAs of *C. acetobutylicum* and *S. aureus* have the same residues for intermolecular recognition, *C. difficile*

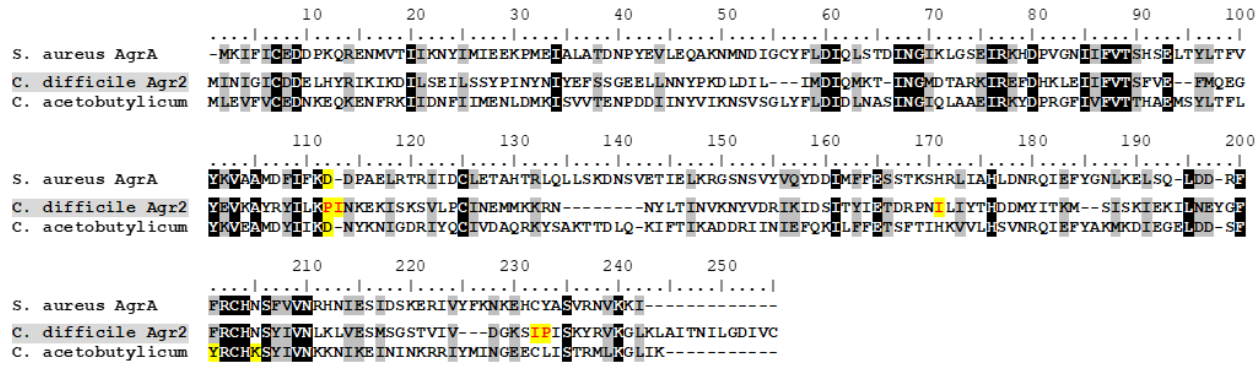

Figure 52: Comparative analysis of the *S. aureus* AgrA and AgrA consensus sequences of Clostridium species. Relevant differences between *S. aureus* and Clostridium species are shown in red, whereas differences between Clostridium species are highlighted in yellow. Black and grey highlighting of amino acids indicates full conservation and similar residues, respectively. The grey shading of the species' name indicates pathogenicity or toxigenicity.

has a different motif composed of Lys-Pro-Ile (positions 111-113). The residues in AgrA that make contact with specific bases in DNA vary throughout the LytTR family of proteins. This is also the case for the *C. difficile* and *C. acetobutylicum* LytTR domains. *C. difficile* has Ile171 instead of His171. *C. acetobutylicum* also has different residues in one of the most conserved DNA-binding motifs, where it has Tyr201 and Lys205 instead of the conserved F201 and N205. Interestingly, the *S. aureus* AgrA has the ability to respond to oxidative stress by creating a disulfide bond between Cys203 and Cys232 (Sun et al., 2012). Both cysteines are conserved in *C. acetobutylicum*, but not in *C. difficile*. The Tyr233, following the second cysteine involved in disulfide bonding, bears a significant role in transcription activation by AgrA, as substitution by alanine led to a significant decrease in transcription (Wang & Muir, 2016). The Tyr233 is substituted in *C. acetobutylicum* for Leu233, which is similar enough to Tyr233. However, *C. difficile* has a

substitution for Pro233, which most likely indicates a significant difference between the AgrA of Clostridial species.

Despite the relatively extensive similarities and the small number of significant differences between the AgrA sequences of *C. acetobutylicum* and *C. difficile*, the differences are enough to suggest that the proteins are different. The recognition domain of AgrA should show similarities between species, as it is an essential part for relaying the signal of the two-component regulatory system. Similarly, the LytTR domain should be different between species as the DNA-binding bases have to be specific to the different promoters of each Agr operon. The differences in the intermolecular recognition motifs adds to the evidence suggesting that the AgrA proteins are different amongst species.

Comparison of all the AgrD Sequences between Clostridial Species

AgrD carries a lot of information within its residues and to achieve specificity, the signal peptide needs a reasonable degree of variation. The alignment of the different AgrD sequences of Clostridia against the AgrD alleles of *S. aureus* is shown in **Figure 21**. There is extensive variation in the N-terminus and AIP portions of the protein. The C-terminus contains significant differences, but it is generally more similar between the species. The alignment of the AgrD sequences shows that the Agr component is different between the species.

The N-terminus of the *S. aureus* AgrD is not conserved, although the amphipathic helix is conserved in all of *S. aureus* AgrDs. The same amphipathic helix was found only in some of the Clostridia species, indicating differences in Clostridia. The amphipathic helix is followed by a helix breaking motif composed of Ile42-Gly43 that allows a turn in the helix, but is not necessary

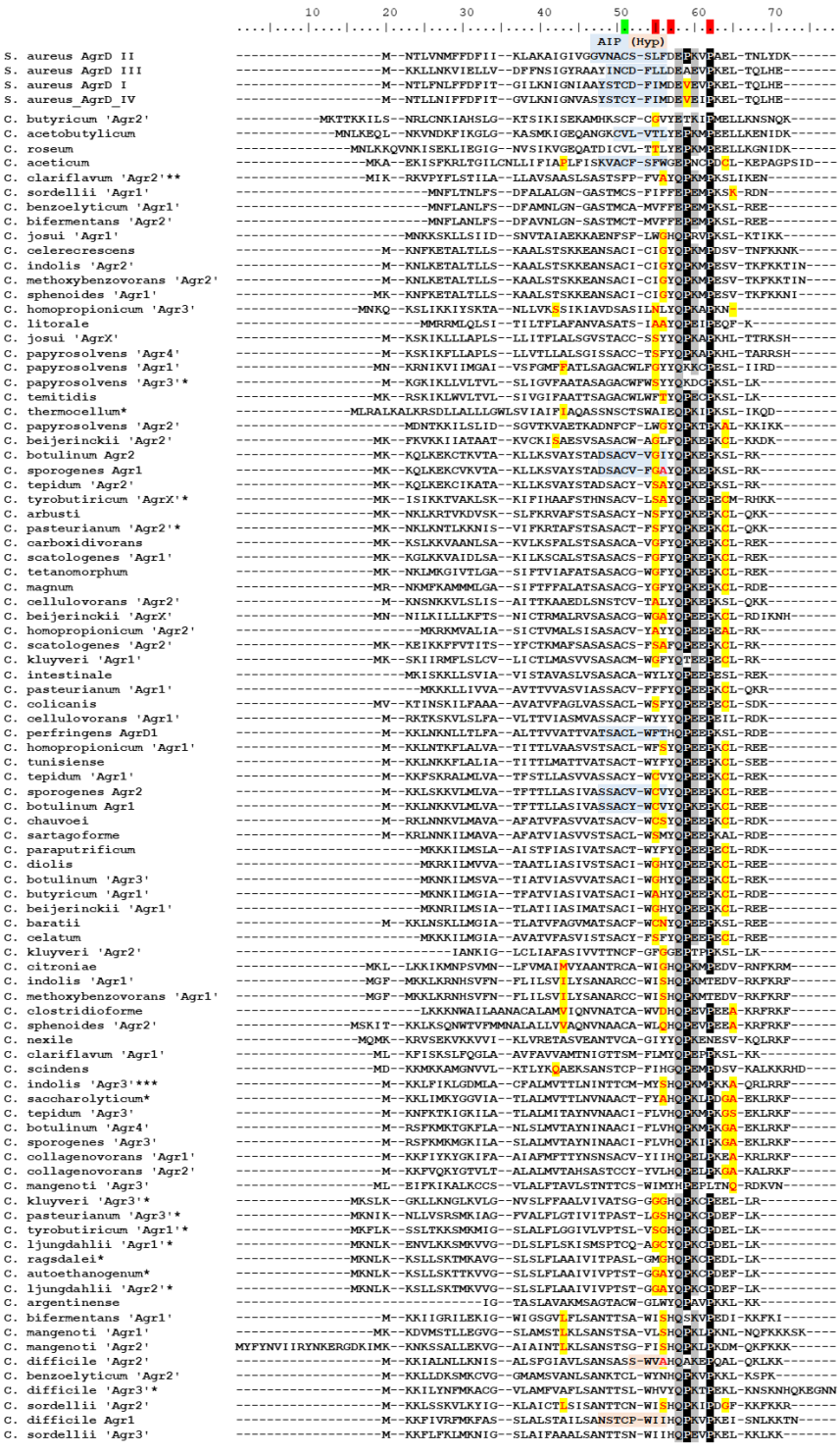

Figure 53: Comparative analysis of the *S. aureus* AgrDI-IV and AgrD consensus sequences of all Clostridium species. Relevant differences between *S. aureus* and Clostridium species are shown in red whereas differences between Clostridium species are highlighted in yellow. Black and grey indicates full conservation and similar residues, respectively. The grey shading of the species' name indicates pathogenicity or toxigenicity. The light blue shading shows empirically proven autoinducer peptides (AIPs) and the orange shading shows predicted AIPs based on bioinformatics analyses through SignalP, Predisi, and Phobius.

for AIP production (Cisar et al., 2009). Although there are many small residues in position 46, they do not possess the helix-breaking property of Gly43, therefore the position is not conserved in Clostridia. The lack of a conserved glycine following the amphipathic helix indicates that Clostridia might have a different method of presenting the AgrD prepeptide to AgrB and a SpsB-like peptidase. The N-terminal cleavage site is also necessary for AIP processing and is composed of certain residues recognizable by SpsB in the *S. aureus* AgrD I, II, and IV (Kavanaugh et al., 2007). These recognition residues include a proline/glycine at the position -5 or -6, a small or branched chain residue at -3, and a glycine/serine/alanine at the -1-position relative to the cleavage site or beginning of the AIP (Kavanaugh et al., 2007). The putative recognition sites are shown in **Figure 21** as the three residues preceding the shaded AIP sequences and show significant variation. The proline/glycine at the position -5 or -6 are missing, but the small or branched chain residue at -3, and a glycine/serine/alanine at the -1 position are present in most Clostridia. The AIP sequences that are empirically unknown were predicted by SingalP, Predisi, and Phobius and shown as orange in the shaded area. Thus, the N-terminus of Clostridia contains different motifs necessary for processing of the AgrD compared to *S. aureus* and within the species, suggesting modified peptidase-interactions.

Similar to the N-terminus, the AIPs within the AgrDs do not demonstrate conservation. Position 51, highlighted in green at the top of the alignment, is the only semi-conserved position, as it mostly has cysteine and serine residues. These small residues form the important thioester (cysteine) and ester (serine) bonds in the macrocycles of the AIPs (Thoendel & Horswill, 2009). Another important motif within the AIP macrocycles of *S. aureus* is the hydrophobic motif composed of two or three bulky hydrophobic residues at the end of the AIP. The corresponding positions in the alignment (54-56) do not show conserved hydrophobicity in at least the last two

positions. Evidently, many AIPs have polar or small residues at positions 55 or 56 instead of bulky hydrophobic residues. The different cyclization residues and variation in the macrocycle residues indicate that the AgrDs of Clostridia are different and specific to each species. Although there is little conservation within features of the AIP, some sequences of Clostridial AgrD are exactly the same across species.

Despite its conserved positions, the C-terminus also has significant differences throughout Clostridia. In *S. aureus*, the entire terminus is considered charged due to the presence of 5 or 6 charged residues (Thoendel & Horswill, 2009). In Clostridia, however, the number of charged residues in the sequences varies between two and seven. The lack of conservation is demonstrated by the large presence of small, uncharged residues at position 64, which is a conserved position essential in *S. aureus* given that mutation from Glu64 to alanine abolished AIP production (Thoendel & Horswill, 2009). Given that the charge of the *S. aureus* C-terminus is necessary for proper interaction and cleavage of AgrD, the different degrees of charge present in the C-termini of Clostridia suggest other mechanism of interaction or fewer necessary charged residues for cleavage.

In contrast to the variation in charge, position 65, one of the most conserved residues in the *S. aureus* AgrDs and essential for endopeptidase activity and AIP production, has hydrophobicity conserved in Clostridia. The C-terminus also has a small patch between positions 58 and 62 that shows strong conservation. The first position, Glu58 shaded in grey, has one of the two residues presumed to allow recognition by AgrB in *S. aureus*. The other AgrD recognition residue is Asp57 (George & Muir, 2007), occupied with mostly aromatic residues in the Clostridia sequences. Positions 58 and 57 are likely to hold residues with similar recognition functions in Clostridia as well given their higher degree of conservation. Notice that the other three residues in the shaded

positions do not have specific functions like Glu58. Following position 58, is the first conserved position, Pro59. The six Clostridial sequences without Pro59 contain acceptable substitutions. The next shaded position is Lys/Glu60, conserved with these two residues, even though five Clostridial sequences have acceptable substitutions instead. The last conserved position is the Pro62 with only four Clostridial sequences diverging from the conservation, although only one has an unacceptable Leu65 substitution. Although the C-terminus is necessary for AIP production in *S. aureus*, the conservation of these C-terminal residues in Clostridia suggests they should be further investigated in both *S. aureus* and Clostridia. Pro59 in addition to Pro62 could provide a binding cleft for regulation of AgrD activity and possibly the interaction with AgrB.

The AIP sequences of Clostridia are different, reflecting their role and specific interactions with AgrC. Their N-termini have different amphipathic helices and cleavage recognition sites, their AIPs do not follow a specific pattern besides the cyclization cysteine and serine, and their C-termini are not significantly charged. However, the semi-conserved recognition site for cleavage by a SpsB-like peptidase and the conserved residues within their C-termini are a promising therapeutic targets of the Clostridia's Agr system.

Comparison of all the AgrB Sequences between Clostridial Species

The AgrB component of the Agr system is a unique protein without homologs apart from other AgrBs (Novick et al., 1995), indicating how specific its role is within the Agr system. The alignment of the Clostridial AgrBs (**Figure 22)** shows conserved residues aligning with the catalytic residues in active sites of the *S. aureus* AgrBs. Proportionally, however, the AgrBs of Clostridia have fewer conserved positions compared to the other Agr components, matching the diversity of AgrD proteins.

The conserved residues of Clostridial AgrBs are present in the protein's binding site, including the His84 and Cys91 necessary for AIP production. Only two Clostridial AgrB components do not have these catalytic residues, *C. josui 'Agr1'*, probably due to a sequencing error, and *C. argentinense* AgrB^. Although not catalytic, Arg77 is a transmembrane residue required for AIP production in *S. aureus* and is conserved in Clostridial AgrBs through both Arg77 and Lys77. Another conserved and required residue present in the vicinity is Gly82 (Thoendel & Horswill, 2013), which follows an additional conserved G81. The Gly81 is exclusive to Clostridia and can indicate a less strict interaction with AgrD or a different type of interaction altogether due to the glycine's hydrogen side-chain and freedom in movement.

The AgrBs of Clostridial species also lack functionally-relevant residues of *S. aureus* AgrB. The Staphylococcal AgrB is dependent on A85, which is not conserved in all Clostridia. The ability of AgrB to cleave AgrD in *S. aureus* is dependent on the fully conserved P139 and AgrD secretion is dependent on a lysine patch that precedes the proline (Thoendel & Horswill, 2013). Although Pro139 is conserved, the patch (positions 143-5) only shows a semi-conserved Lys145 with occasional Arg145 and His145. As the lysine patch allows for secretion of the processed AgrD, the significant differences may indicate a potentially different secretion mechanism for the AgrD of Clostridia.

Apart from residues responsible for direct interactions with other proteins, the Clostridial AgrB components also have residues vital for stability of the protein. In *S. aureus*, residues Gly39 and Gln41 resulted in a destabilized AgrB when mutated to Val39 and Pro41 (Thoendel & Horswill, 2013). Apart from the first few sequences, Gly39 is conserved throughout Clostridia. Residue Gln41, however, is not conserved in Clostridia. There is also a position with hydrophobic and mostly aromatic residues conserved at position 38 that is not mentioned in literature and could

be relevant to Clostridia. Given that the residues are hydrophobic, they are likely transmembrane residues involved in protein stability. Furthermore, position 46 also harbors a necessary asparagine in *S. aureus*, as isoleucine or tyrosine mutation lead to inhibition of cleavage of AgrD (Thoendel & Horswill, 2013), but the position does not have asparagine nor polarity conserved in Clostridia. A transmembrane mutation at Ser167 lead to similar destabilization most likely due to the introduction of a charged residue in the membrane (Thoendel & Horswill, 2013). Although there is no charged residue at position 167, a majority of bulky and hydrophobic residues occupy position 167. Residues that hold the protein together are bound to vary and possibly lose their function across homologs. Therefore, even if there were significant similarities between the AgrBs of Clostridial species and *S. aureus*, they would probably be less significant than the differences in their catalytic regions.

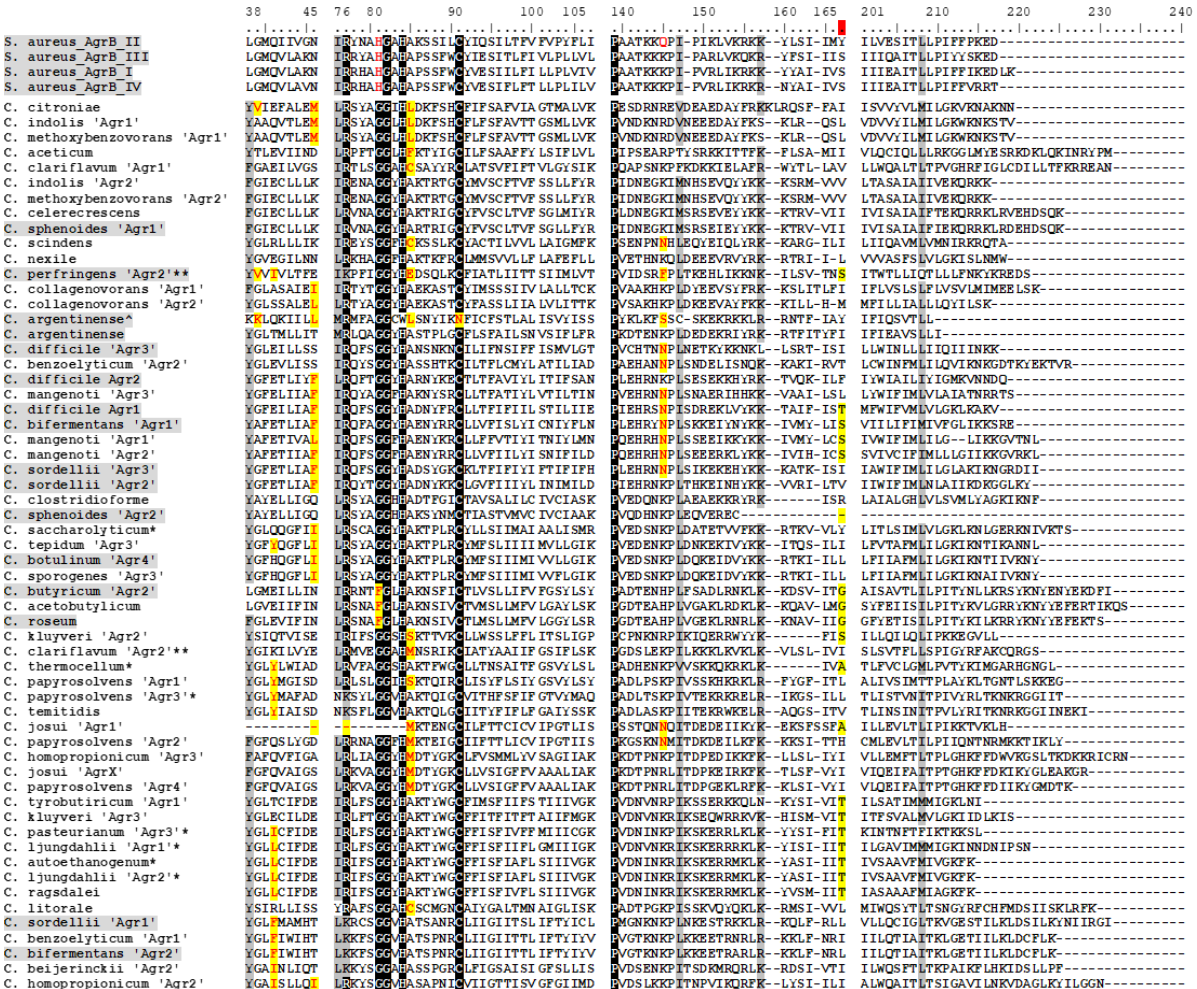

Figure 54: Comparative analysis of regions with relevant similarities and differences between the sequences of *S. aureus* AgrBI-IV and AgrB consensus sequences of all Clostridium species. Relevant differences between *S. aureus* and Clostridium species are shown in red or highlighting at the top of the alignment, whereas differences between Clostridium species are highlighted in yellow. Black and grey highlighting of amino acids indicates full conservation and similar residues, respectively. The grey shading of the species' name indicates pathogenicity or toxigenicity.

Altogether, the interacting residues of the AgrBs in Clostridia support the proteins function, but there are sufficient differences to distinguish between the proteins. The similarities lie in the catalytic intracellular membrane loop and in its AgrD cleaving motif. The differences, on the other hand, are also present in positions aligning with residues in the catalytic region and other relevant positions of *S. aureus*. Interestingly, the AgrBs of Clostridia also have additional conserved residues that are not relevant in *S. aureus* but are still located in relevant regions of the protein.

Comparison of all the AgrC Sequences between Clostridial Species

Most Clostridial species have the main functional residues of the *S. aureus* AgrCs conserved. The regions of most conservation surround the active sites of the catalytic boxes in the dimerization and histidine phosphotransfer domain and the catalytic ATP binding domain. The conservation is enough to maintain the function of the protein, but there are still significant differences within these domains. The least conserved domain is the sensor domain, as it varies from species to species. The specific motif and residues are shown in **Figure 23** and outlined below.

The three fully conserved residues within the AgrC of Clostridia are the ones that define the ATP binding cleft and the ATP binding motif and composed of the H-box, N-box, and the G-box. The essential residues corresponding to the boxes include His399, Asn524, and G593. The only sequences that do not contain all of these residues are the AgrC of *C. indolis* Agr3, both of the operons of *C. mangenoti*, and *C. ragsdalei*. The absence of these crucial residues is probably a result of mutations or sequencing errors, as it appears their sequences are incomplete. Both the H-box (positions 379-410) and the N-box (positions 515-528) have three semi-conserved positions

in addition to the fully conserved His399 and Asn524. However, the functionality of the residues of the N-box has not been as thoroughly explored.

Generally, the H-box shows low conservation and significant differences regarding *S. aureus* in a significant number of sequences. Researchers have found specific residues within the *S. aureus* H-box that interact with AgrA, these are Val402, Ile404, Leu405, and Leu408. Out of these positions, 402 and 404 (shaded in red at the top of the alignment, **Fig. 23**) showed significant differences from *S. aureus*. Conversely, the other two leucine residues are conserved through hydrophobic residues. Specific mutations within the H-box of *S. aureus,* for example, the mutation of Met383 to Leu383 lead to constitutive activation of the protein, which is significantly present in the alignment. Arg387 mutations to histidine/cysteine/glycine387, none of which was found in the alignment, also lead to constitutive activation. However, a significant number of the species have Leu387 at that position, a residue that is not favorable in place of arginine. Lastly, Tyr401 mutation to Cys401 that also turns on the constitutive phenotype, is absent from the other Clostridial species and has hydrophobicity conserved. Some of the functional residues of the *S. aureus* H-boxes form a motif (F[RK]HDYXN) around His399 that is conserved in other histidine kinases and is part of the HPK10 category of Histidine Kinase (HK) domains. This pattern is not conserved in the putative H-boxes of Clostridial species and indicates significant variation in DNA-binding residues within the genus. Despite the variation in the H-box, the G-box (positions 583-599) contains the least number of similar residues between the AgrCs of Clostridium species whilst being the largest box motif. The similarities are limited to the conserved G-X-G (591-X-593) motif, with X being hydrophobic and bulky in most sequences.

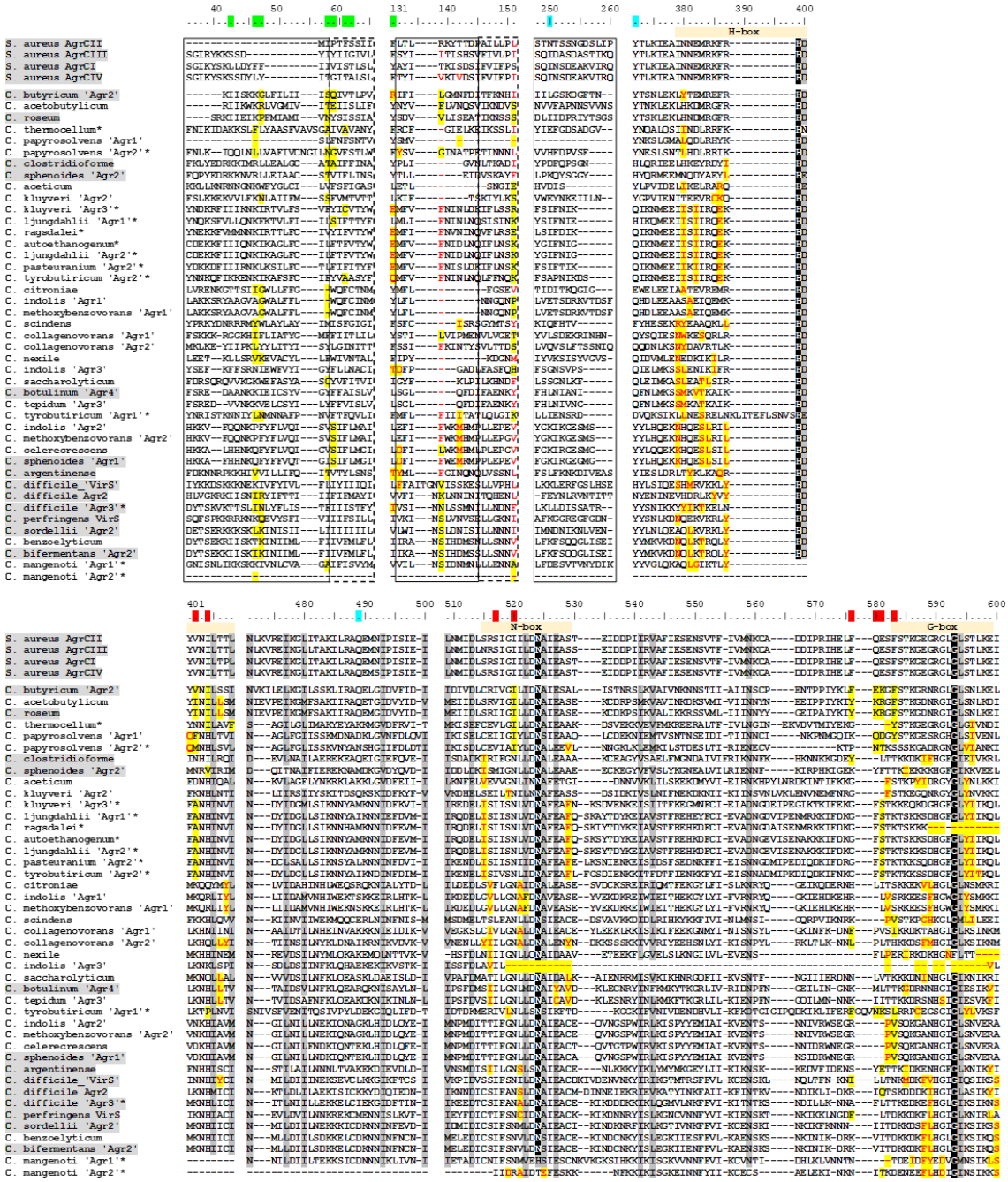

Figure 55: Comparative analysis of regions with relevant similarities and differences between the sequences of *S. aureus* AgrCI-IV and AgrC consensus sequences of all Clostridium species. Relevant differences between *S. aureus* and Clostridium species are shown in red or highlighting at the top of the alignment, whereas differences between Clostridium species are highlighted in yellow. Blue highlights at the top of the alignment indicate a specific mutation of potential importance and green highlights indicate positions with conserved amino acids. Black and grey highlighting of amino acids indicates full conservation and similar residues, respectively. Solid boxes show functionally relevant regions in *S. aureus*, including extracellular portions of AgrC-I. Dashed boxes show an extended region that contains functionally-relevant residues of the *S. aureus* AgrC-I. Specific domains span the colored bars above the alignment. The grey shading of the species name indicates pathogenicity or toxigenicity.

An additional G-box is present in other HK domains whereas in most Clostridia, a conserved Asn559 takes the place of the G-box's aspartate residue. This G-box is also absent in other HPK10 categories, but most HPK10 HK domains have the original aspartate. Only a few sequences contain the original aspartate, including *C. kluyveri* 'Agr2', *kluyveri* 'Agr3', *ljungdhali* 'Agr1'*, *ragsdalei**, *autoethanogenum**, *ljungdhali* 'Agr2'*, almost the same cluster from the H- and N-boxes. This difference in addition to the significant differences that the (F[RK]HDYXN) motif carries raises the question of whether the Agr HK domain of Clostridia can be categorized differently. Although there are still significant differences in relevant positions of the histidine kinase domain, it is the most conserved domain across Clostridia in comparison to the sensor domain.

The sensor domain significantly varies in the four *S. aureus* Agr groups and also varies between Clostridia. In *S. aureus*, the first extracellular loop is responsible for activating interactions as alanine mutations of residues Leu43, Phe46, Phe47, Ile58, Val59, Ser61, and Thr62 abolished activation in AgrC–AIP interactions in group I and diminished activation in group IV (Cisar & Elizabeth, 2009). In the second extracellular loop, the *S. aureus* AgrC I has residues necessary for its activation by and responsible for specificity with AIP I, including Tyr131, Ala132, Thr139, Ser142, and Ser151 (Cisar & Elizabeth, 2009). These positions are within highly variable recognition domains so the lack of conservation of any residue is justified. Some of the positions also have many gaps, rendering them obsolete. However, the following positions (highlighted in green, **Figure 24**) have traits conserved; most residues at position Tyr131 have a bulky hydrophobic character; position Phe46 has mostly Lys46 and Arg46; Phe47, Ile58, Val59, and Ser61 all have hydrophobicity conserved and a reasonable number of aromatic residues. The characters of these residues suggest the possibility of interaction, given the capability of

hydrophobic interactions between aromatic rings and aliphatic chains or hydrogen bond through the charge on position 46. However, the functions of these residues can only be confirmed through *in vitro* testing.

A few specific sensor domain mutations at positions Arg259, Ser262, Thr286, and Leu294 lead to constitutive activity of AgrC (Geisinger, Muir, & Novick, 2009), but the Arg259 aligned to that position with mostly gaps. Ser262 and Thr268 had mostly hydrophobic residues and some were aromatic, matching the mutation leading to AgrC's constitutive activity. The Ser262 position is conserved through polarity of the residues. These significant differences and variations support the fact that the sensor domain, and consequently the AgrC components are different from each other. They also confirm that the mechanisms of recognition of the AIP are different between *S. aureus* and Clostridia, as the mutations leading to constitutive activation in *S. aureus* are unlikely to lead to constitutive activation in Clostridia.

There is data on very specific mutations in the *S. aureus* AgrC that could be relevant, and their positions are highlighted in blue (**Fig. 24**). A mutation at Ile250 to lysine led to lack of sensitivity to AIPs of other groups (Geisinger et al., 2009). The Ile250Lys mutation is present in a significant number of Clostridial sequences, and other sequences have a charged residue at position 250. At the least, the presence of this mutation in Clostridia indicates a difference in AgrC-AgrD interactions between *S. aureus* and Clotridia. The Tyr372 of the *S. aureus* AgrC, located in the sensor domain, has been implicated in AgrC-AgrA interaction, as a cysteine mutation in the position led to different genetic regulation resulting in a colonizing phenotype rather than a cytotoxic effect (Norrby-Teglund et al., 2016). Many Clostridia have a tyrosine residue at position 372, however, other polar residues are also present, such as glutamine and histidine. Another

mutation that lead to constitutive activation, glutamine489 to histidine/arginine/glutamate489 (Geisinger et al., 2009), is present in some of the species.

The functions of the Clostridial residues that align with the mutated functional *S. aureus* residues are unknown within Clostridia. The mutations that showed some effect on the *S. aureus* AgrCs are probably obsolete within the Clostridium genus, however, the variation in these positions provide evidence that the AgrC components are different from *S. aureus* and different between the Clostridial species. Additionally, the variations within both the Senor and HK domains suggest that the AgrCs of Clostridia are different, even within the same species. Despite these differences, the AgrCs of Clostridia and *S. aureus* are probably homologous and carry out the same function within their Agr systems.

Comparison of all the AgrA Sequences between Clostridial Species

The AgrA components of Clostridia have the majority of the functional residues of the LytTR response regulator conserved throughout the sequences of all species. **Figure 24** shows the conservation of the catalytic residues, dimerization domain, and intermolecular recognition domain. On the other hand, **Figure 24** also demonstrates that the DNA-binding domain and

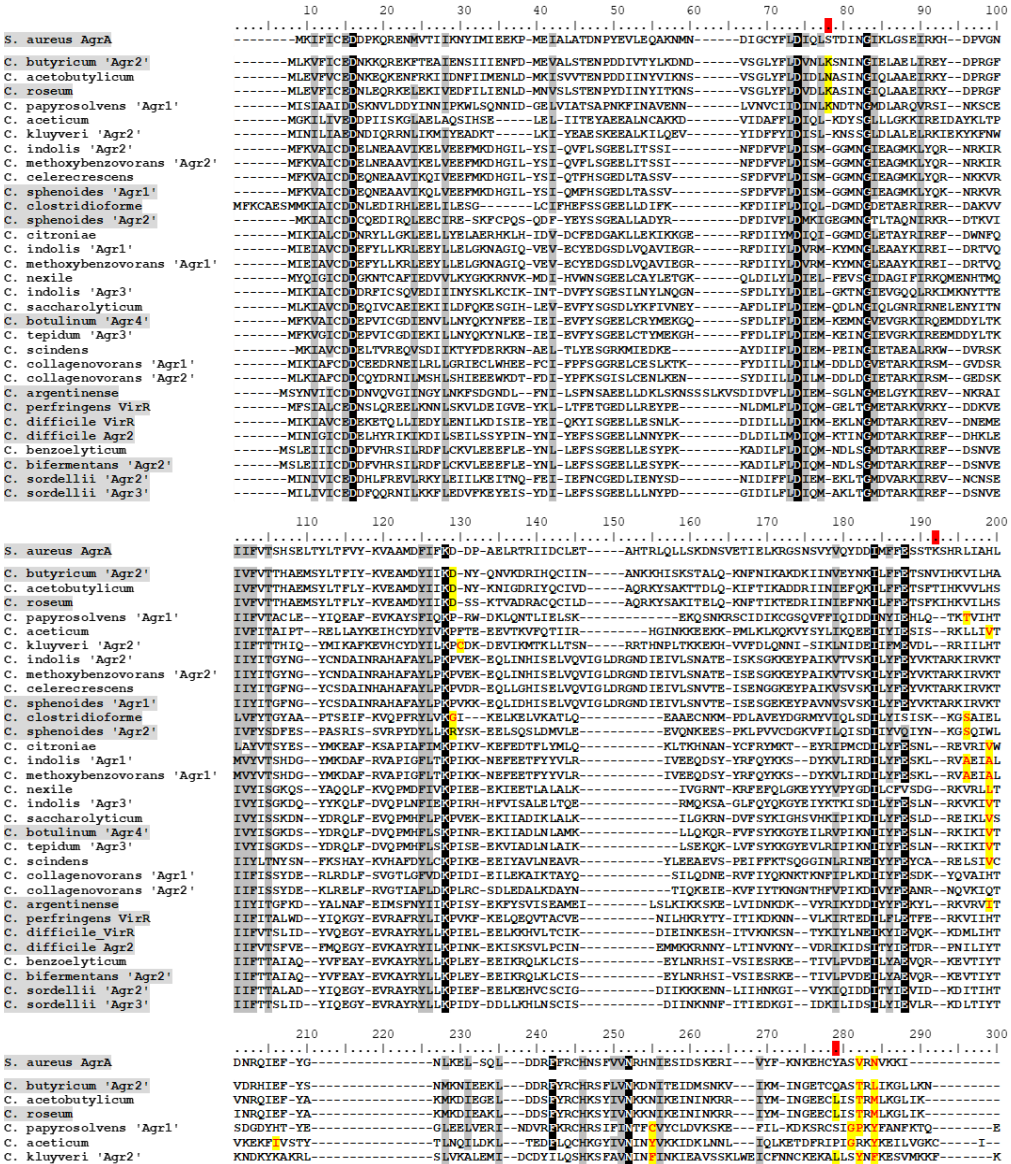

Figure 56: **Comparative analysis of the *S. aureus* AgrA and AgrA consensus sequences of all Clostridium species.** Relevant differences between *S. aureus* and Clostridium species are shown in red or highlighted at the top of the alignment, whereas differences between Clostridium species are highlighted in yellow. Black and grey highlighting of amino acids indicates full conservation and similar residues, respectively. The grey shading of the species name indicates pathogenicity or toxigenicity.

response regulator recognition domain is not conserved in the Clostridia AgrA. Therefore, the data establishes the Clostridial AgrA as distinguishable proteins with relevant similarities.

All Clostridia contain the three aspartates (or two aspartates and one glutamate) in the shaded positions 15, 16 and 74. The sequences have a partially conserved Lys-Pro-Ile (KPI) dimerization domain, as only the Lys128 residue is fully conserved, but positions 129 and 130 have significant conservation of proline and a bulky hydrophobic residue, respectively. The conserved Lys128 and semi-conserved Pro129 are also functional in the binding of ATP, as the lysine forms a salt-bridge with the phosphorylation site and the proline directs the site lysine towards the active site (Gao & Stock, 2009; Marchler-Bauer et al., 2017). Given the prevalence and possible dual function of Lys128 and Pro129, these residues could be valuable targets for deactivation of the Agr system in pathogens toxins are regulated by the operon.

The DNA-binding residues of *S. aureus*, His194, and Asn247 are not conserved in the Clostridial species, but all sequences have acceptable polar substitutions in their place. The third DNA-binding residue of the *S. aureus* AgrA, Arg283, does not have polarity conserved in the position. These positions, however, do not necessarily represent DNA-binding sites, as they vary considerably within the LytTR domain (Sidote et al., 2009). The other DNA-binding motif present in the LytTR domain is composed of FFRCHNS (McGowan et al., 2002). In the AgrA alignment, the only truly conserved residues within the motif is phenylalanine(242), serine(248), and histidine/tyrosine(246). The other positions of the motif, however, are not conserved, although the majority of the residues at positions arginine(244) and arginine(247) are polar. Other locations that affect DNA-binding by AgrA in *S. aureus* include position 196, which has a conserved hydrophobicity, and position 206, which has polar residues (Nicod et al., 2014). The *S. aureus* AgrA also has a residue that is necessary for the beginning of transcription even after binding to

promoter P3 (Wang & Muir, 2016). Transcription at the P3 promoter is halted in an Ala mutant at position Tyr279, where Clostridia mostly have a disfavored Pro as a substitute. The residues that affect AgrA interaction with DNA are not conserved in Clostridia, suggesting different and specific mechanisms of regulation from *S. aureus*, including within the Clostridium genus. However, the residues between positions 242-8 are the best candidates for targeted modulation of the Agr system by impeding DNA-binding.

A peculiar trait of *S. aureus* AgrA is the ability to form a disulfide bond between C245 and C278 in oxidative conditions (Sun et al., 2012), interrupting its activity. Similarly, a few Clostridial species have the cysteine residues conserved at the same position, including *C. acetobutylicum, roseum, papyrosolvens,* and *aceticum*. All species but *C. roseum* are non-pathogenic or non-toxigenic.

Apart from DNA, AgrA also interacts with AgrC through the response regulator recognition domain, which has an intermolecular recognition domain (IMRD) (Marchler-Bauer et al., 2017). In *S. aureus*, the IMRD is composed of residues leucine77, serine78, isoleucine81, asparaginen82, and glycine83, out of which position 77 has conserved hydrophobicity and glycine83 is conserved in Clostridia. The polarity of asparagine82 is also conserved and the hydrophobicity of isoleucine81 is somewhat conserved through isoleucine, valine, tyrosine, methionine, and leucine. serine78 is the only position that has the first few sequences with conserved polarity but has a gap in most of the species. The higher conservation primes IMRD as a target for halting the Agr system by severing the interaction between AgrA and AgrC, removing the intracellular response.

Some positions do not have direct implications on the interactions of AgrA but do keep the integrity of the protein. Asn252 and Ile256 of the *S. aureus* AgrA are examples of such residues

that are conserved. The former residue is fully conserved, and the latter has hydrophobicity conserved. Locations nearby (250 and 251) also have hydrophobicity conserved throughout. While these residues demonstrate similarity between the AgrA of Clostridia, other residues that are necessary for detection of AgrA expression in *S. aureus* distinguish between Clostridial the AgrA. Among these residues are Lys192, His199, and Asn255 (Nicod et al., 2014). Lys192 aligns with a position that includes a gap in most of the species. His199 is not conserved as there are a series of hydrophobic residues at that position. Lastly, Asn255 is not conserved, as there are many significantly different residues at that position.

Met228 of the AgrA in *C. perfringens* is usually conserved through a leucine in similar response regulators, as it is responsible for stabilizing the response regulator-DNA complex. In the Clostridial AgrA, the position has hydrophobicity fully conserved, presenting another interesting residue for intervention and supporting a degree of similarity in AgrA. *C. perfringens*'s VirR has a serine-lysine-histidine-arginine motif at positons 281-284 (McGowan et al., 2002; McGowan, O'Connor, Cheung, & Rood, 2003) with side chains essential for DNA-binding activity. However, there is significant variability within these positions, as shown previously through AgrA's Arg283.

The AgrA components of Clostridia have relevant residues that are conserved, but also have residues in relevant regions that are not conserved. While the AgrA of Clostridia do not show full conservation of any motif, the dimerization domain, a DNA-binding motif, stabilizing residues, and the IMRD are the most similar across the genus. Therefore, they provide the most uniform targets for modulation of Agr function. Despite the similarities, the AgrA of Clostridia are still different as other residues in the DNA binding motifs, and the response regulator recognition domain are not conserved.

Evolutionary inference of the Agr components

The Agr system could be split into two operons with AgrC and AgrA on one hand and the AgrB and AgrD on the other. This categorization originates from their function and is reflected in their maximum likelihood phylogenetic trees as shown in **Figures 25** and **26**. The trees of AgrD and AgrB show a more dissimilar topology compared to the trees of AgrC and AgrA. The trees show pathogenic Clostridia in bold, which are dispersed throughout the leaves of all four trees. The dispersion of pathogens throughout the tree is evidence of the lack of relationship between the structure of the Agr components and the pathogenicity of the species. However, the pathogens *C. difficile*, *C. perfringens*, *C. sordellii*, and *C. bifermentans* do form a polytomous clade in both AgrA and AgrC trees, meaning the sequences do not provide enough information to discern branching, or the nodes were not statistically significant. The polytomous clade of AgrA is much more statistically robust than the corresponding clade in AgrC. AgrB also shows clustering of sequences of these four pathogens, but the clade is statistically significant and not polytomous. Another clade containing sequences from pathogens *C. roseum* and *C. butyricum*, in addition to *C. acetobutylicum*, is the most related to *S. aureus* compared to the rest of the clades. Interestingly, this clade is present in all four trees. Many of the AgrC without the AgrA at their flanks (AgrC*) cluster into a clade with statistical robustness, although some other sequences with these orphan AgrCs are not in the same clade. Most of the orphan AgrC*s have a corresponding AgrD2* and AgrB2*. The AgrD2* and AgrB2* sequences are found in mostly polytomous clades that are topologically equivalent to the AgrC2* sequences. Interestingly, all of the AgrA orphan sequences are present in non-pathogenic genomes, apart from the orphan histidine kinase of *C. difficile*. Furthermore, *C. difficile* is the only species that has all its sequences most related to each other in every tree.

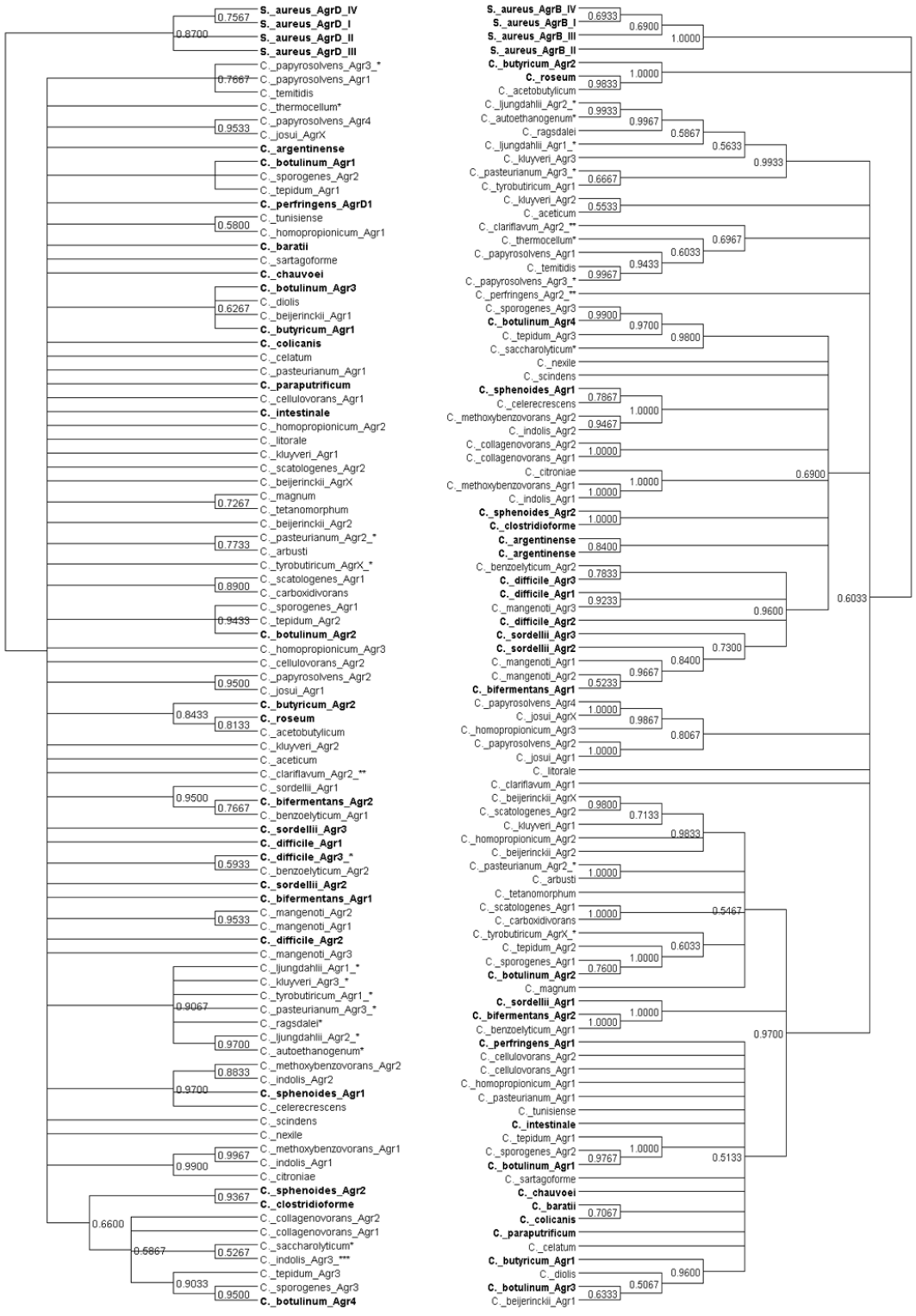

Figure 57: The bootstrap phylogenetic trees of the consensus AgrD (right) and AgrB (left) sequences of Clostridium species. The trees inferred by using the maximum likelihood method and JTT matrix-based model. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (300 replicates) are shown next to the branches. The species in bold are pathogenic or toxigenic. *Evolutionary analyses were conducted in MEGA X.*

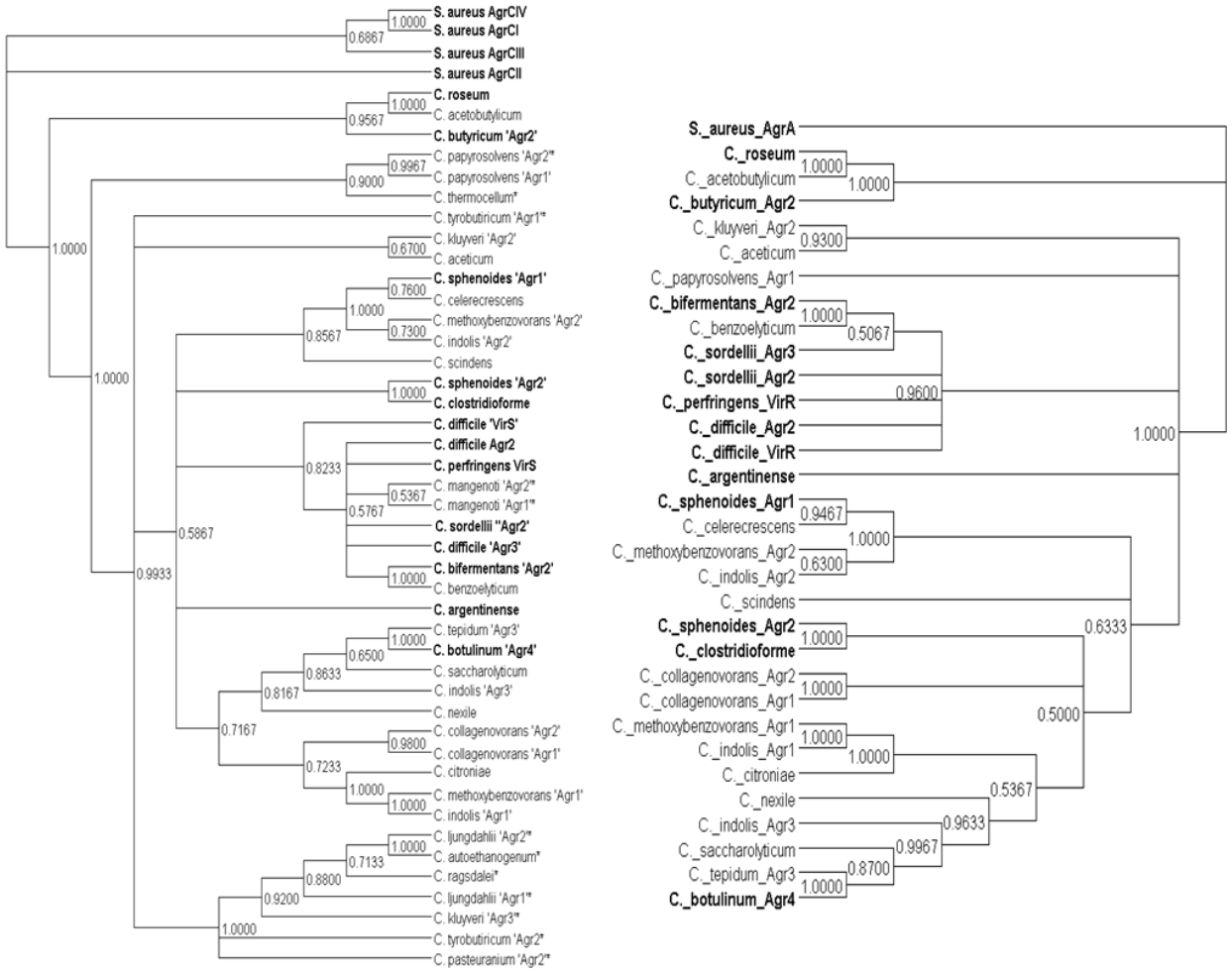

Figure 58: The bootstrap phylogenetic trees of the consensus AgrC (left) and AgrA (right) sequences of Clostridium species. The trees were constructed using the maximum likelihood method and JTT matrix-based model. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (300 replicates) are shown next to the branches. The species in bold are pathogenic or toxigenic. Evolutionary analyses were conducted in MEGA X.

CONCLUSION

Antibiotic resistance has been a major threat to our best defense against bacterial infections. Antibiotic-based treatment of bacterial infections has saved many lives, since the first antibiotic was discovered in 1928. Now, the effectiveness of many antibiotics is under threat due to antimicrobial resistance. Thus, antimicrobial resistance poses a major to public health. . The genus Clostridium has its own multidrug-resistant bug, *C. difficile* (Davies J & Davies D, 2010), a bacterium that itself has become a major threat and an enormous burden to public health authorities (Gupta & Khanna, 2014). As a result, various non-antibiotic therapies that pose minimum risk of resistance are being explored. *C. perfringens* is also a concern, since it is a common perpetrator of foodborne illnesses with strains that are resistant to antibiotics (Labbe & Juneja 2017). The Agr system controls toxin production and virulence in both of these pathogens as well as other Clostridial species. The virulence-associated processes controlled by the Agr operon include toxin production, colonization (Darkoh & Asiedu, 2014; Darkoh et al., 2015; Darkoh et al., 2016; Martin et al., 2013) and motility (Martin et al., 2013) in *C. difficile*, and sporulation in *C. perfringens* (McClane et al., 2012). In this study, the components of the Agr system in different pathogenic Clostridia was compared and the results identified similarities and differences that could serve as targets for the development of non-antibiotic, anti-virulence therapies against these pathogens.

Given the virulence and other important functions of the Agr system, similarities in Agr proteins across Clostridium species can be targeted for single therapies that could inhibit different Clostridial pathogens. Apart from the catalytic residues present in *S. aureus* and Clostridia, novel similarities were found between Clostridium species. The AgrD sequences of Clostridia demonstrate a similar C-termini composed of a rigid proline-based motif with charged residues. Given the charges and rigid motif, a therapeutic drug could be developed to bind to the C-termini

of AgrD with high affinity to abort its interaction with AgrB, the peptidase mediating AIP cyclization. The absence of cyclization inhibits the production of the AIP, which in turn would halt toxin production and other cellular processes. The AgrBs of Clostridia are also similar in their catalytic loop region that is novel and could be explored as a potential target. The catalytic loop appears to be very flexible given the presence of a second conserved glycine residue and because it is the active site of the protein, these residues are likely important in the processing of AgrD. Targeting this site may interrupt the ability of the AgrB protein to cyclize AIP and toxin production would be abolished. In Addition, the dimerization domain of AgrA is another similar motif that could be targeted for anti-virulence treatment against Clostridial pathogens. Since AgrA dimerization is necessary for function, sequestering the binding site between AgrA would inhibit the regulation and promotion of toxin producing genes.

New Agr operons were also found in species with reported functional Agr systems, such as *C. botulinum* and *C. sporogenes*. Most interestingly, the *C. botulinum* Agr1, *C. difficile* Agr2, and *C. sporogenes* Agr1 have components where the same groups of strains have similar sequence variations in both components, suggesting divergence into different components that could interact with each other. These findings, combined with the data on sequence identity, indicated that some of the Agr components could be different and must be further investigated for different functions or interactions compared to the other Agr systems. This may potentially lead to categorizing the Agr components into different groups. The implications of a different Agr operon within the same species, or even within the same operon, could indicate the ability to cross-regulate their Agr operons. Although *S. aureus* does not seem to have two Agr operons in the same strain's genome, its different Agr components can regulate each other, either by activation or inhibition of the Agr systems (Geisinger et al., 2009). Research on *S. aureus* demonstrates that this cross-regulation of

the operons has physiological consequences in mouse models (Wright, Jin, & Novick, 2005) and suggests that the same can happen within Clostridia, especially given the different operons are within the same bacterium. The ability to cross-regulate could provide *C. difficile*, for example, the ability to increase the efficiency of the system, depending on the AIP activate the other AgrC. This could result in increased toxin production and possibly, contribute to hypervirulence.

Some Clostridia have components with significant differences in key functional motifs. Despite the differences, the components are likely functional within their own species. Therefore, these differences are interesting but do not necessarily have phenotypic and systemic effect. However, some differences might have an effect, for example, AgrD has differences in the size and presence of the amphipathic helix. If the helix is non-existent, then the Agr system is probably less efficient than others as the tethering of AgrD to the membrane by the amphipathic helix allows for quick processing of the AgrD. The cyclization residue might have a more significant effect on the system, as an AIP cyclized at a cysteine residue is more ephemeral than a serine-based AIP (Gorske & Blackwell, 2006). Therefore, the thiolactone AIP could lead to a more stable AIP and an increased effect of the system, such as the transcription of the downstream genes it regulates.

There was no relationship between the structures of the Agr proteins and pathogenicity or toxigenicity. However, a significant number of the sequences of the pathogenic species cluster together in all trees, indicating that they have a closer common ancestor and are somewhat similar. Therefore, the phylogenetic trees support the idea of having a unique therapy to treat a subset of pathogenic Clostridia. There is also clustering of many sequences of species with operons that are missing an AgrA.

Despite the thorough comparisons made between Agr components within and between species, the study demonstrates limitations. The sequences of the components varied in size

between species, reducing the overall effectiveness of the comparisons. Some species, such as *C. difficile, C. sordellii*, *C. botulinum*, and *C. perfringens*, have more strains published in NCBI, therefore, some species have more data with increased validity compared to other species. Another source of uncertainty includes the taxonomy of the species. Species' names change due to misclassification and there is a possibility that some Clostridia included in this analysis are not truly part of the Clostridium genus. Thus, some Clostridia might have to be removed from the analysis if their taxonomy is changed. Another detail to notice is that the sequences were stopped being collected in September of 2018, meaning there could be more sequences of Agr components that have not been included in the analysis. On the other hand, the breadth of the analysis would not have been possible without the approach used to retrieve the sequences from NCBI. The BLASTP and the Entrez search methods enabled the retrieval of Agr sequences from most, if not all, Clostridium species containing the operon, providing largest collection of sequences of Clostridial Agr proteins in the literature. The residue-centered comparisons provided specific blueprints for experiments that will explore the function of the Agr components in both pathogenic and industrially relevant Clostridia. The deeper analysis of the functional Agr components focused on the sequences of proteins with empirical function, supporting a stronger argument and clearer understanding of the potential applications and implications of the differences and similarities. Furthermore, it provides a library of sequenced Agr proteins that could be used by synthetic biologists to develop custom regulator systems. Looking forward, it would be interesting to investigate predictive methods, such as a modeling the docking of the interacting components to predict what areas or residues of the Agr components would to better understand the function of the proteins.

Although the interactions and mechanisms of the Agr components may likely be different between species, the results from this study showed similarities in Clostridia species that could be explored for drug development. It is envisioned that small molecule drugs designed to target the motifs in the Agr system identified to be similar in the pathogenic Clostridia may be harnessed to develop non-antibiotic therapies against these public health important pathogens. These potential non-antibiotic therapies are less likely to stimulate resistance, since the Agr system is not directly associated with growth (Darkoh & DuPont, 2017).

APPENDIX I

List of species included in analysis

C. aceticum
C. acetobutylicum
C. arbusti
C. argentinense
C. autoethanogenum
C. baratii
C. beijerinckii
C. benzoelyticum
C. bifermentans
C. botulinum
C. butyricum
C. carboxidivorans
C. celatum
C. celerecrescens
C. cellulovorans
C. chauvoei
C. citroniae
C. clariflavum
C. clostridioforme
C. colicanis
C. collagenovorans
C. difficile
C. diolis
C. homopropionicum
C. indolis
C. intestinale
C. josui
C. kluyveri
C. litorale
C. ljungdahlii
C. magnum
C. mangenoti
C. methoxybenzovorans
C. nexile
C. papyrosolvens
C. paraputrificum
C. pasteuranium
C. perfringens
C. ragsdalei
C. roseum
C. saccharolyticum
C. sartagoforme
C. scatologenes
C. scindens
C. sordellii
C. sphenoides
C. sporogenes
C. temitidis
C. tepidum
C. tetanomorphum
C. thermocellum
C. tunisiense
C. tyrobutyricum

References

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution, 35*(6), 1547-1549.

Kumar, S., Mashooq, M., Gandham, R. K., Alavandi, S., & Nagaleekar, V. K. (2018). Characterization of quorum sensing system in clostridium chauvoei. *Anaerobe*,

Darkoh, C., & DuPont, H. L. (2017). The accessory gene regulator-1 as a therapeutic target for C. difficile infections. *Expert Opinion on Therapeutic Targets, 21*(5), 451-453.

Labbe, R. G., & Juneja, V. K. (2017). Clostridium perfringens. Foodborne diseases (pp. 235-242) Elsevier.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., . . . Bryant, S. H. (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research, 45*(D1), D200-D203.

Nielsen, H. (2017). Predicting secretory proteins with SignalP. *Protein Function Prediction: Methods and Protocols,* , 59-73.

Rodrigues, R., Barber, G. E., & Ananthakrishnan, A. N. (2017). A comprehensive study of costs associated with recurrent clostridium difficile infection. *Infection Control & Hospital Epidemiology, 38*(2), 196-202.

Udaondo, Z., Duque, E., & Ramos, J. (2017). The pangenome of the genus Clostridium. *Environmental Microbiology, 19*(7), 2588-2603.

Verbeke, T. J., Giannone, R. J., Klingeman, D. M., Engle, N. L., Rydzak, T., Guss, A. M., . . . Elkins, J. G. (2017). Corrigendum: Pentose sugars inhibit metabolism and increase expression of an AgrD-type cyclic pentapeptide in clostridium thermocellum. *Scientific Reports, 7*, 46875.

Wang, B., Zhao, A., Xie, Q., Olinares, P. D., Chait, B. T., Novick, R. P., & Muir, T. W. (2017). Functional plasticity of the AgrC receptor histidine kinase required for staphylococcal virulence. *Cell Chemical Biology, 24*(1), 76-86.

Yu, Q., Lepp, D., Mehdizadeh Gohari, I., Wu, T., Zhou, H., Yin, X., . . . Gong, J. (2017). The agr-like quorum sensing system is required for pathogenesis of necrotic enteritis caused by clostridium perfringens in poultry. *Infection and Immunity, 85*(6), 10.1128/IAI.00975-16. Print 2017 Jun.

Canovas, J., Baldry, M., Bojer, M. S., Andersen, P. S., Gless, B. H., Grzeskowiak, P. K., . . . Ingmer, H. (2016). Cross-talk between staphylococcus aureus and other staphylococcal species via the agr quorum sensing system. *Frontiers in Microbiology, 7*, 1733.

Darkoh, C., Odo, C., & DuPont, H. L. (2016). Accessory gene regulator-1 locus is essential for virulence and pathogenesis of clostridium difficile. *Mbio, 7*(4), 10.1128/mBio.01237-16.

Dawn Thompson, J. (2016). 1 - introduction. In J. D. Thompson (Ed.), *Statistics for bioinformatics* (pp. 3-25) Elsevier.

Mairpady Shambat, S., Siemens, N., Monk, I. R., Mohan, D. B., Mukundan, S., Krishnan, K. C., . . . Norrby-Teglund, A. (2016). A point mutation in AgrC determines cytotoxic or colonizing properties associated with phenotypic variants of ST22 MRSA strains. *Scientific Reports, 6*, 31360.

Ohtani, K. (2016). Gene regulation by the VirS/VirR system in clostridium perfringens. *Anaerobe, 41*, 5-9.

Wang, B., & Muir, T. W. (2016). Regulation of virulence in staphylococcus aureus: Molecular mechanisms and remaining puzzles. *Cell Chemical Biology, 23*(2), 214-224.

Darkoh, C., DuPont, H. L., Norris, S. J., & Kaplan, H. B. (2015). Toxin synthesis by clostridium difficile is regulated through quorum signaling. *Mbio, 6*(2), e02569-14.

Kök, M. S. (2015). An integrated approach: Advances in the use of clostridium for biofuel. *Biotechnology and Genetic Engineering Reviews, 31*(1-2), 69-81.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., . . . Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research, 43*(Database issue), D222-6.

Church, G. M., Elowitz, M. B., Smolke, C. D., Voigt, C. A., & Weiss, R. (2014). Realizing the potential of synthetic biology. *Nature Reviews Molecular Cell Biology, 15*(4), 289.

Darkoh, C., & Asiedu, G. A. (2014). **Quorum sensing systems in clostridia**. In Vipin Chandra Kalia (Ed.), ***Quorum sensing vs quorum quenching: A battle with no end in sight*** (Kalia, Vipin Chandra ed., pp. 133--154). New Delhi, India: Springer India.

Gupta, A., & Khanna, S. (2014). Community-acquired clostridium difficile infection: An increasing public health threat. Infection and Drug Resistance, 7, 63-72.

Hargreaves, K. R., Kropinski, A. M., & Clokie, M. R. (2014). What does the talking?: Quorum sensing signalling genes discovered in a bacteriophage genome. *PloS One, 9*(1), e85131.

Kadariya, J., Smith, T. C., & Thapaliya, D. (2014). Staphylococcus aureus and staphylococcal food-borne disease: An ongoing challenge in public health. BioMed Research International, 2014, 827965.

Nicod, S. S., Weinzierl, R. O., Burchell, L., Escalera-Maurer, A., James, E. H., & Wigneshweraraj, S. (2014). Systematic mutational analysis of the LytTR DNA binding domain of staphylococcus aureus virulence gene transcription factor AgrA. *Nucleic Acids Research, 42*(20), 12523-12536.

Ostell, J. (2014). The entrez search and retrieval system.

Pearson, W. R. (2014). BLAST and FASTA similarity searching for multiple sequence alignment. In D. J. Russell (Ed.), *Multiple sequence alignment methods* (pp. 75-101) Humana Press.

Peter, D. (2014). *Systems biology of clostridium* World Scientific.

Sully, E. K., Malachowa, N., Elmore, B. O., Alexander, S. M., Femling, J. K., Gray, B. M., . . . Edwards, B. S. (2014). Selective chemical inhibition of agr quorum sensing in staphylococcus aureus promotes host defense with minimal impact on resistance. PLoS Pathogens, 10(6), e1004174.

Wang, B., Zhao, A., Novick, R. P., & Muir, T. W. (2014). Activation and inhibition of the receptor histidine kinase AgrC occurs through opposite helical transduction motions. *Molecular Cell, 53*(6), 929-940.

Baum, D. A., & Smith, S. D. (2013). *Tree thinking: An introduction to phylogenetic biology* Roberts.

Grass, J. E., Gould, L. H., & Mahon, B. E. (2013). Epidemiology of foodborne disease outbreaks caused by clostridium perfringens, united states, 1998–2010.*Foodborne Pathogens and Disease, 10*(2), 131-136.

Gray, B., Hall, P., & Gresham, H. (2013). Targeting agr- and agr-like quorum sensing systems for development of common therapeutics to treat multiple gram-positive bacterial infections. *Sensors (Basel, Switzerland), 13*(4), 5130-5166.

Jabbari, S., Steiner, E., Heap, J. T., Winzer, K., Minton, N. P., & King, J. R. (2013). The putative influence of the agr operon upon survival mechanisms used by clostridium acetobutylicum. *Mathematical Biosciences, 243*(2), 223-239.

Martin, M. J., Clare, S., Goulding, D., Faulds-Pain, A., Barquist, L., Browne, H. P., . . . Wren, B. W. (2013). The agr locus regulates virulence and colonization genes in clostridium difficile 027. *Journal of Bacteriology, 195*(16), 3672-3681.

Tal-Gan, Y., Ivancic, M., Cornilescu, G., Cornilescu, C. C., & Blackwell, H. E. (2013). Structural characterization of native autoinducing peptides and abiotic analogues reveals key

features essential for activation and inhibition of an AgrC quorum sensing receptor in staphylococcus aureus. *Journal of the American Chemical Society, 135*(49), 18436-18444.

Thoendel, M., & Horswill, A. R. (2013). Random mutagenesis and topology analysis of the autoinducing peptide biosynthesis proteins in S taphylococcus aureus.*Molecular Microbiology, 87*(2), 318-337.

Chen, J., & McClane, B. A. (2012). Role of the agr-like quorum-sensing system in regulating toxin production by clostridium perfringens type B strains CN1793 and CN1795. *Infection and Immunity, 80*(9), 3008-3017.

Steiner, E., Scott, J., Minton, N. P., & Winzer, K. (2012). An agr quorum sensing system that regulates granulose formation and sporulation in clostridium acetobutylicum. *Applied and Environmental Microbiology, 78*(4), 1113-1122.

Sun, F., Liang, H., Kong, X., Xie, S., Cho, H., Deng, X., . . . He, C. (2012). Quorum-sensing agr mediates bacterial oxidation response via an intramolecular disulfide redox switch in the response regulator AgrA. *Proceedings of the National Academy of Sciences of the United States of America, 109*(23), 9095-9100.

Vidal, J. E., Ma, M., Saputo, J., Garcia, J., Uzal, F. A., & McClane, B. A. (2012). Evidence that the Agr-like quorum sensing system regulates the toxin production, cytotoxicity and pathogenicity of clostridium perfringens type C isolate CN3685. *Molecular Microbiology, 83*(1), 179-194.

Li, J., Chen, J., Vidal, J. E., & McClane, B. A. (2011). The agr-like quorum-sensing system regulates sporulation and production of enterotoxin and beta2 toxin by clostridium perfringens type A non-food-borne human gastrointestinal disease strain F5603. *Infection and Immunity, 79*(6), 2451-2459.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., . . . Bryant, S. H. (2011). CDD: A conserved domain database for the functional annotation of proteins. *Nucleic Acids Research, 39*(Database issue), D225-9.

Saujet, L., Monot, M., Dupuy, B., Soutourina, O., & Martin-Verstraete, I. (2011). The key sigma factor of transition phase, SigH, controls sporulation, metabolism, and virulence factor expression in clostridium difficile. *Journal of Bacteriology, 193*(13), 3186-3196.

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., . . . Griffin, P. M. (2011). Foodborne illness acquired in the united states--major pathogens. *Emerging Infectious Diseases, 17*(1), 7-15.

Cooksley, C. M., Davis, I. J., Winzer, K., Chan, W. C., Peck, M. W., & Minton, N. P. (2010). Regulation of neurotoxin production and sporulation by a putative agrBD signaling system in proteolytic clostridium botulinum. *Applied and Environmental Microbiology, 76*(13), 4448-4460.

Davies, J., & Davies, D. (2010). Origins and evolution of antibiotic resistance. Microbiology and Molecular Biology Reviews : MMBR, 74(3), 417-433.

Cisar, G., & Elizabeth, A. (2009). Mechanism of signal transduction by the staphylococcus aureus quorum sensing receptor AGRC.

Gao, R., & Stock, A. M. (2009). Biological insights from structures of two-component proteins. *Annual Review of Microbiology, 63*, 133-154.

Geisinger, E., Muir, T. W., & Novick, R. P. (2009). Agr receptor mutants reveal distinct modes of inhibition by staphylococcal autoinducing peptides. *Proceedings of the National Academy of Sciences of the United States of America, 106*(4), 1216-1221.

George Cisar, E. A., Geisinger, E., Muir, T. W., & Novick, R. P. (2009). Symmetric signalling within asymmetric dimers of the staphylococcus aureus receptor histidine kinase AgrC. *Molecular Microbiology, 74*(1), 44-57.

Ohtani, K., Yuan, Y., Hassan, S., Wang, R., Wang, Y., & Shimizu, T. (2009). Virulence gene regulation by the agr system in clostridium perfringens. *Journal of Bacteriology, 191*(12), 3919-3927.

Stabler, R. A., He, M., Dawson, L., Martin, M., Valiente, E., Corton, C., . . . Rose, G. (2009). Comparative genome and phenotypic analysis of clostridium difficile 027 strains provides insight into the evolution of a hypervirulent bacterium. *Genome Biology, 10*(9), R102.

Thoendel, M., & Horswill, A. R. (2009). Identification of staphylococcus aureus AgrD residues required for autoinducing peptide biosynthesis. *The Journal of Biological Chemistry, 284*(33), 21828-21838.

Underwood, S., Guan, S., Vijayasubhash, V., Baines, S. D., Graham, L., Lewis, R. J., . . . Stephenson, K. (2009). Characterization of the sporulation initiation pathway of clostridium difficile and its role in toxin production. *Journal of Bacteriology, 191*(23), 7296-7305.

Gautier, R., Douguet, D., Antonny, B., & Drin, G. (2008). HELIQUEST: A web server to screen sequences with specific α-helical properties. *Bioinformatics, 24*(18), 2101-2102.

Sidote, D. J., Barbieri, C. M., Wu, T., & Stock, A. M. (2008). Structure of the staphylococcus aureus AgrA LytTR domain bound to DNA reveals a beta fold with an unusual mode of binding. *Structure, 16*(5), 727-735.

Wuster, A., & Babu, M. M. (2008). Conservation and evolutionary dynamics of the agr cell-to-cell communication system across firmicutes. *Journal of Bacteriology, 190*(2), 743-746.

George, E. A., & Muir, T. W. (2007). Molecular mechanisms of agr quorum sensing in virulent staphylococci. *Chembiochem, 8*(8), 847-855.

Käll, L., Krogh, A., & Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction—the phobius web server. *Nucleic Acids Research, 35*(suppl_2), W429-W432.

Kavanaugh, J. S., Thoendel, M., & Horswill, A. R. (2007). A role for type I signal peptidase in staphylococcus aureus quorum sensing. *Molecular Microbiology, 65*(3), 780-798.

Gorske, B. C., & Blackwell, H. E. (2006). Interception of quorum sensing in staphylococcus aureus: A new niche for peptidomimetics. Organic & Biomolecular Chemistry, 4(8), 1441-1445.

Hall, T. (2005). *Bioedit* (7.0.5 ed.). North Carolina: North Carolina State University, Department of Microbiology.

Qiu, R., Pei, W., Zhang, L., Lin, J., & Ji, G. (2005). Identification of the putative staphylococcal AgrB catalytic residues involving the proteolytic cleavage of AgrD to generate autoinducing peptide. *The Journal of Biological Chemistry, 280*(17), 16695-16704.

Wright, J. S.,3rd, Jin, R., & Novick, R. P. (2005). Transient interference with staphylococcal quorum sensing blocks abscess formation. Proceedings of the National Academy of Sciences of the United States of America, 102(5), 1691-1696.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792-1797.

Hiller, K., Grote, A., Scheer, M., Münch, R., & Jahn, D. (2004). PrediSi: Prediction of signal peptides and their cleavage positions. *Nucleic Acids Research, 32*(suppl_2), W375-W379.

Käll, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology, 338*(5), 1027-1036.

Marchler-Bauer, A., & Bryant, S. H. (2004). CD-search: Protein domain annotations on the fly. *Nucleic Acids Research, 32*(Web Server issue), W327-31.

Zhiqiang, Q., Yang, Z., Jian, Z., Youyu, H., Yang, W., Juan, J., . . . Di, Q. (2004). Bioinformatics analysis of two-component regulatory systems in staphylococcus epidermidis. *Chinese Science Bulletin, 49*(12), 1267-1271.

Betts, M. J., & Russell, R. B. (2003). Amino acid properties and consequences of substitutions. *Bioinformatics for Geneticists, ,* 289-316.

McGowan, S., O'Connor, J. R., Cheung, J. K., & Rood, J. I. (2003). The SKHR motif is required for biological function of the VirR response regulator from clostridium perfringens. *Journal of Bacteriology, 185*(20), 6205-6208.

Geer, L. Y., Domrachev, M., Lipman, D. J., & Bryant, S. H. (2002). CDART: Protein homology by domain architecture. *Genome Research, 12*(10), 1619-1623.

McGowan, S., Lucet, I. S., Cheung, J. K., Awad, M. M., Whisstock, J. C., & Rood, J. I. (2002). The FxRxHrS motif: A conserved region essential for DNA binding of the VirR response regulator from clostridium perfringens. *Journal of Molecular Biology, 322*(5), 997-1011.

Nikolskaya, A. N., & Galperin, M. Y. (2002). A novel type of conserved DNA-binding domain in the transcriptional regulators of the AlgR/AgrA/LytR family. *Nucleic Acids Research, 30*(11), 2453-2459.

Zhang, L., Gray, L., Novick, R. P., & Ji, G. (2002). Transmembrane topology of AgrB, the protein involved in the post-translational modification of AgrD in staphylococcus aureus. *The Journal of Biological Chemistry, 277*(38), 34736-34742.

Arnon, S. S., Schechter, R., Inglesby, T. V., Henderson, D. A., Bartlett, J. G., Ascher, M. S., . . . Layton, M. (2001). Botulinum toxin as a biological weapon: Medical and public health management. *Jama, 285*(8), 1059-1070.

Stock, A. M., Robinson, V. L., & Goudreau, P. N. (2000). Two-component signal transduction. *Annual Review of Biochemistry, 69*(1), 183-215.

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symposium Series, , 41*(41) 95-98.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology, 292*(2), 195-202.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, 12*(2), 85-94.

Stackebrandt, E., Kramer, I., Swiderski, J., & Hippe, H. (1999). Phylogenetic basis for a taxonomic dissection of the genus clostridium. *FEMS Immunology & Medical Microbiology, 24*(3), 253-258.

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

Novick, R. P., Projan, S. J., Kornblum, J., Ross, H. F., Ji, G., Kreiswirth, B., . . . Moghazeh, S. (1995). The agr P2 operon: An autocatalytic sensory transduction system in staphylococcus aureus. *Molecular & General Genetics : MGG, 248*(4), 446-458.

Altschul, S. F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *Journal of Molecular Evolution, 36*(3), 290-300.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915-10919.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*(3), 275-282.

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology*, *219*(3), 555-565.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, *39*(4), 783-791.