MANCHESTER
1824

The University of Manchester

A COMPUTATIONAL APPROACH TO STUDYING THE EVOLUTION OF STREPTOCOCCAL QUORUM SENSING SYSTEMS

A thesis submitted to The University of Manchester for the degree of

DOCTOR OF PHILOSOPHY (PhD)

in the Faculty of Life Sciences

2014

RAJA FARHANA RAJA KHAIRUDDIN

TABLES OF CONTENTS

Word count: 45,713

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

For many years, researchers have studied the social lives of bacteria to understand intra- and inter-species interactions. Cell-cell communication, also known as quorum sensing (QS), is used by bacteria to coordinate their behaviour in response to environmental conditions. The QS system in *Streptococcus* species is well known to regulate competence. Studies show that *Streptococcus pneumoniae* has two homologous QS systems: 1) the competence (Com) system that regulates competence; and 2) a bacteriocin-like peptide (Blp) system that regulates the production of bacteriocins. Both functions are widespread in the genus. In *S. pneumoniae*, the Blp QS system shares a common ancestor and has similar features to the Com QS system. However, the evolutionary relationship between these QS systems remains obscure.

SUCRE methodology was developed to identify the QS homologous genes in the streptococcal species. SUCRE uses four complementary approaches: homology search, putative gene finding, regulon construction, and evolutionary analysis. The performance of SUCRE was assessed in comparison with other orthology detection methods. SUCRE is precise in identifying the QS homologous genes and has similar performance to OrthoMCL. The QS system structures are found to be conserved across the streptococcal species.

A streptococcal species phylogeny was constructed from the ribosomal and tRNA synthetase gene families. Using the QS genes identified from SUCRE and the streptococcal species phylogeny, the study infers the evolution of the QS systems in *Streptococcus* species. The study shows that the QS systems evolved as a regulon unit. The paralogous relationship between each of the QS systems suggests that duplication has a huge influence on functional divergence of the QS systems in the genus. Although, horizontal gene transfer (HGT) is commonly found in bacteria, little evidence is found to support that the effect of HGT on the functional divergence of the QS systems in this genus. However, the QS regulon genes of the same QS system are found to be non-vertically transferred across species that signifies that the HGT event promotes the sequence variation between these genes.

DECLARATION

The University of Manchester
PhD by published work Candidate Declaration

Candidate Name: Raja Farhana RAJA KHAIRUDDIN

Faculty: Life Sciences

Thesis Title: A computational approach to studying the evolution of the streptococcal quorum sensing systems.

Declaration to be completed by the candidate:

I declare that no portion of this work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Signed: Date: 2 November, 2014

COPYRIGHT

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude to my supervisor Dr. Simon Whelan for being such an amazing mentor, who with knowledge, patience and enthusiasm guided me from the beginning of the journey until this paper is structured. And a very special thanks to my co-supervisor Dr. Daniel Rozen for the advice and the constructive suggestions throughout the research. My experience learning from both of you has been priceless.

It is my pleasure to thank Prof. Simon Lovell, Dr. Christopher Knight, and my fellow colleagues who had been there to support me and made this journey remarkable. I would also like to thank Prof. David Robertson and Prof. Peter Young for making my defence an enjoyment moment and for the brilliant comments and positive insights.

My deepest thanks and love to my beloved family, who continually encouraged me, whom without, none of this would be possible. The thesis is especially dedicated to them. Thank you.

CHAPTER 1

Introduction

1.1 Quorum sensing systems in bacteria

For the longest time, it was thought that a single bacterium lives alone. However, studies have shown that this is not the case. Bacteria actually live in communities where they talk to each other in order to find nutrients and reproduce. There are trillions of cells in the human body, and there are approximately ten times as many bacteria that live in and on our body. These bacteria can be either beneficial or harmful. A single bacterium would not be sufficient to affect an organ or tissue in our body. A bacterium takes in nutrients and multiplies to increase its number within a population. Communication between bacteria cells, which later become known as quorum sensing (QS) (Fuqua et al., 1994), is important in co-ordinating bacterial behaviour within a population. Once there are sufficient bacterial numbers to engage in a particular action, bacteria communicate with each other to synchronise their actions.

This control mechanism in both Gram-positive and Gram-negative bacteria regulates gene expression, which governs physiological functions, in response to the density of the signalling molecules produced by the bacteria (Bassler, 1999, Dunny and Winans, 1999). Numerous studies have examined the mechanism of QS, both within and between different bacterial species (Bassler, 1999, Dunny and Winans, 1999, Claverys and Havarstein, 2002).

The QS mechanism involves producing, detecting and responding to small signalling molecules. This mechanism allows bacteria: to coordinate their behaviour to environmental changes, such as nutrient availability (Finkel and Kolter, 2001); to engage in defence mechanisms against other bacteria (Høyland-Kroghsbo et al., 2013); to control virulence factor expression (Rutherford and Bassler, 2012) and biofilm production (Vuong et al., 2003, Rathsam et al., 2005, Joelsson et al., 2006); and adapt to stimuli at the population level (Baysse et al., 2005). These responses are crucial for increasing the chance of survival of a bacterial species in population.

Each bacterium has specific signalling molecules that regulate different bacteria functions. A study on QS systems between bacteria showed that most bacteria have the ability to produce and transmit at least one of the signalling molecules (Miller and Bassler, 2001). Although each QS system regulates a different function, in general, how the QS

systems response to the molecular signals is similar (Bassler, 1999, Dubnau, 1999, Miller and Bassler, 2001). First, the molecular signals are produced in the cell. Then, these molecules are secreted into the extracellular environment. The signalling molecules increase as the number of bacteria cells in a population increases. When the accumulation of the molecular signals reaches a certain concentration threshold, the signalling molecules bind to a sensor protein, which becomes phosphorylated. The transfer of a phosphate group to a receptor protein triggers a signal transduction cascade that leads to changes in the expression of genes with particular functions. Because QS systems allow the cells to work together as a group, changes in gene expression happen collectively among the bacterial species in a population.

In streptococcal species, the QS mechanism is known to regulate bacterial competence. In fact, the regulation of bacterial competence in *S. pneumoniae* was the first process that was shown to involve a QS mechanism (Dawson and Sia, 1931). In 1964, an experiment by Tomasz and Hotchkiss (1964) inferred that the competence in *S. pneumoniae* species is controlled by the accumulation of a signalling molecule, called the Competence Stimulating Peptide (CSP) (Havarstein et al., 1995).

Out of a million microbial species, there are approximately a few hundred that cause diseases. *S. pneumoniae* is one of the most infectious microbial species that can grow, metabolize and have independent interactions with other organisms, including humans. *S. pneumoniae* is known to cause about 5 million fatal pneumonia infections per year in children in the United States, and around 7% of the 1 million cases per year in the elderly are fatal (University of Warwick, 2008). Over the last two decades, the incidence of pneumococcal resistance to penicillin and other antibiotics has alarmed the world and led to a growing concern regarding the control of pneumococcal disease. The natural competence of this species has been well documented since then, as it is the major determinant that contributes to the reservoir of penicillin-resistant bacteria (Appelbaum, 1987).

The competence (Com) QS in *S. pneumoniae* is used as a model of a QS cascade to study the evolution of QS systems in *Streptococcus* species. A homologous QS system, bacteriocin-like peptide (Blp), regulates the production of bacteriocins in *S. pneumoniae* (de Saizieu et al., 2000, Martin et al., 2006). Both functions are widespread

in the *Streptococcus* genus. Detailed descriptions of the Com QS and the homologous QS systems, Blp, in *S. pneumoniae* are presented in the following sections.

1.1.1 Regulation of *S. pneumoniae* Com QS systems

S. pneumoniae has been one of the best-studied bacteria for more than 80 years, as it was first used to demonstrate transformation by Fred Griffith in 1928 (Griffith, 1928). The presence of the Com regulatory cascade has led to the categorisation of competence genes into two classes based on two main processes: i) the induction of competence in the bacterial cell; and ii) the mechanism of DNA uptake into the cell. Seven early competence genes (*comABCDE*, *comX* and *comW*), which are dependent on the phosphorylation of ComE (ComE~P) for their expression, are employed to induce competence (Pestova et al., 1996). Subsequently, 14 late competence genes that are dependent on the *comX* gene, which encodes an alternative sigma factor, are used to process the DNA that was taken up (Claverys and Havarstein, 2002).

An overview of the Com QS mechanism in *S. pneumoniae* is illustrated in Figure 1.1. In *S. pneumoniae*, the QS mechanism has been found to initiate the regulation of the competence system, which comprises an unmodified peptide pheromone, a secretion apparatus and a two-regulatory system (TCS) (Morrison and Lee, 2000, Claverys and Havarstein, 2002). Although *S. pneumoniae* has been reported to be competent only for a short period of time during its exponential growth (Palmen et al., 1994), this cell-cell signalling triggers a global change in gene expression, which leads to the transcription of the late genes. The late genes then encode proteins involved in the binding, processing, uptake and recombination of DNA.

Pakula and Walczak (1963), together with Tomasz and Hotchkiss (1964), discovered that streptococci use QS to measure the external concentration of their secreted signalling molecules, CSPs, to induce competence (Claverys and Havarstein, 2002). The *comC* gene was shown to encode CSP, which is a precursor peptide pheromone (Havarstein et al., 1995). CSP is an unmodified, 17-residue peptide containing a double-glycine (GG) type leader at its N-terminal end (Havarstein et al., 1995). CSP is exported and matured

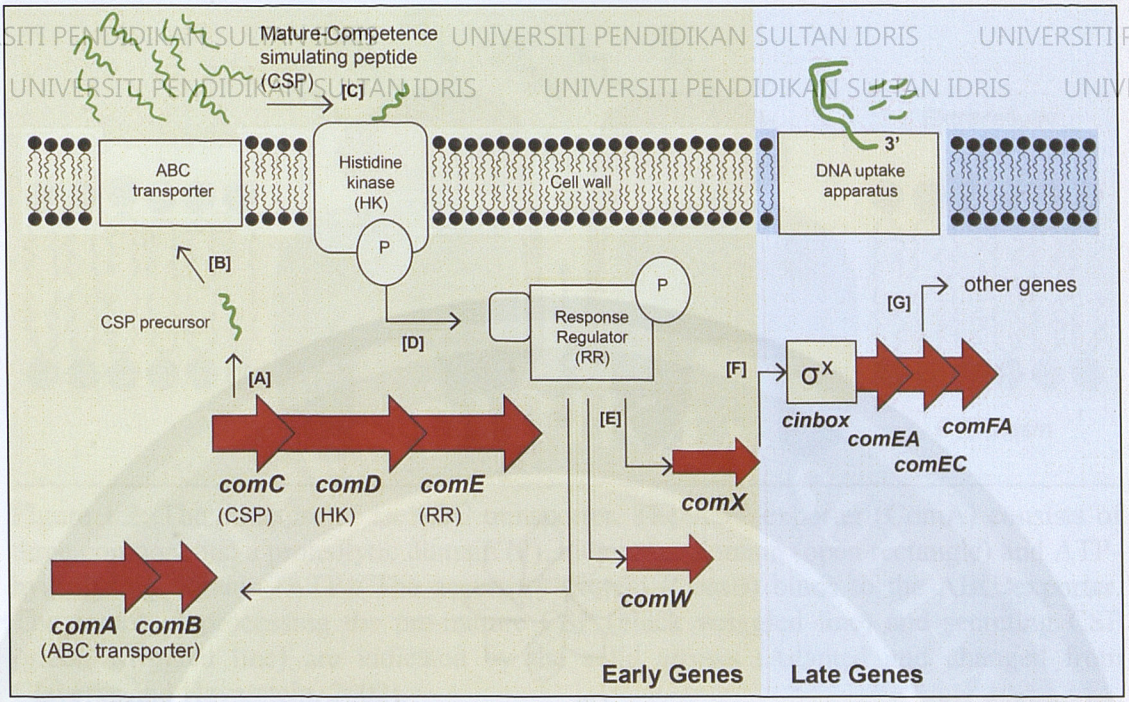

Figure 1.1: The regulation of the S. pneumoniae Com QS system. There are two main processes: i) competence induction (yellow box) involves early genes, which encode competence induction regulatory control components (comAB: an ABC transporter; comC: competence simulating peptide (CSP); comD: histidine kinase (HK); comE: response regulator (RR), ComW, ComX), respectively; ii) the DNA uptake mechanism (blue box) involves comEA, comEC, comEC, comFA and other 11 genes. The routes of the competence mechanisms are indicated by the black arrows in alphabetical order (Adapted and changed from Peterson et. al., 2006).

via the actions of ComAB, a dedicated ABC-transporter that consists of an ABC exporter (ComA) and an accessory protein (ComB) (Fath and Kolter, 1993).

ComA comprises three domains: a proteolytic domain, a membrane domain and an ATP-hydrolysing domain (Figure 1.2). An accessory protein, ComB, is required to assist the ABC exporter. Inactivation of either ComA or ComB affects the production of the peptide pheromone (Hui and Morrison, 1991, Hui et al., 1995). *comB* is located immediately downstream from *comA*. Although it is not clear how ComB assists ComA, ComB shares a similar feature with the accessory protein HlyD (Wang et al., 1991) at its N-terminus. The HlyD protein is co-expressed with HlyB (Hui and Morrison, 1991) to

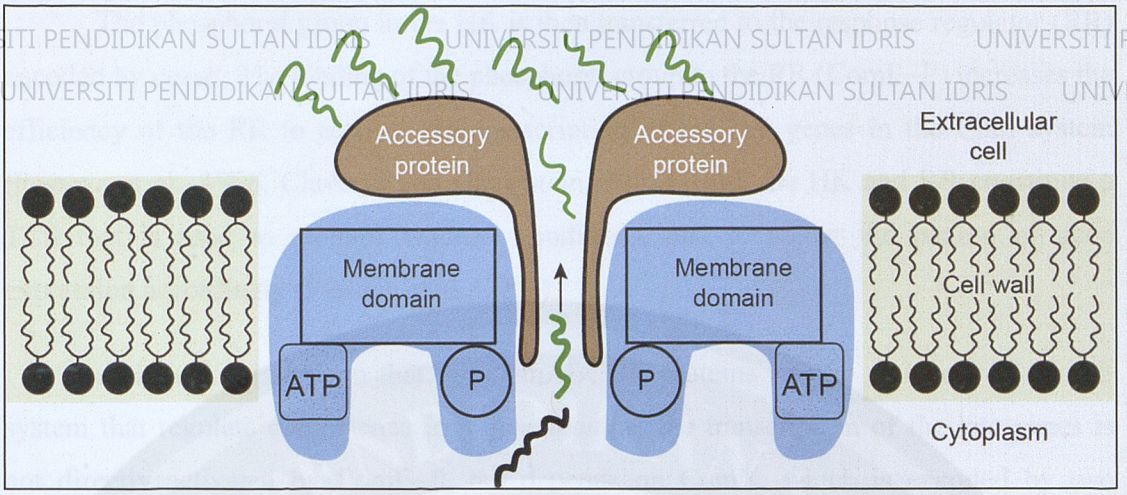

Figure 1.2: The structure of an ABC transporter. The ABC exporter (ComA) consists of three components: a proteolytic domain (P), membrane domain (open rectangle) and ATP-hydrolysing domain (ATP). The accessory protein (ComB) binds to the ABC exporter. The routes of processing the pre-mature CSP (black wriggled line) and secreting CSP (green wriggled line) are indicated by the solid arrows (Adapted and changed from Claverys and Havarstein, 2002).

provide a membrane structure for the export of α-haemolysin in the *Escherichia coli* haemolysis system that is involved in toxin production. ComA and ComB are most similar to LcnC and LcnD, which are the components of a bacteriocin ABC transporter in *Lactococcus lactis* (Hui et al., 1995).

Amino acid residues in the GG-leader of ComC are cleaved by the proteolytic domain at the N-terminus of the ABC transporter. The mature peptide pheromones are transported across the membrane, mediated by the membrane domain, with the energy provided by the ATP-hydrolysing domain. While the ABC transporter processes and exports CSP outside the cell, a dedicated TCS, the histidine kinase (HK) encoded by the *comD* gene (Pestova et al., 1996) detects the extracellular CSP. It was found that the HK (ComD) receptor belongs to the class of orthodox kinases that are composed of a highly variable membrane-spanning N-terminal sensor region and a more conserved C-terminal kinase that enables histidine to accept a phosphoryl group (Havarstein et al., 1996). The binding of CSP changes the conformation of the HK, which results in its dimerization and, subsequently, its autophosphorylation (Stock et al., 2000).

The phosphoryl group in the HK is then transferred to the response regulator (RR) encoded by *comE*. The binding of the phosphoryl group to the RR (ComE~P) increases the efficiency of the RR to activate the transcription of the late genes in the Com system (Pestova et al., 1996, Claverys and Havarstein, 2002). Both the HK and RR constitute a TCS that is used to monitor external conditions, and to adjust the pattern of gene expression accordingly (Pestova et al., 1996).

Although it is known that the ComABCDE proteins are the elements of the QS system that regulate competence in *S. pneumoniae*, the transcription of the late genes is not directly activated by ComE~P, but depends on ComX, which is encoded by two identical genes, *comX*1 and *comX*2 (Lee and Morrison, 1999). They have been shown to have an imperfect direct repeat motif similar to the ComE binding site. This similar character suggests that ComE activates the expression of ComX1 and ComX2, which encode a specific alternative sigma factor required for the synthesis of competence proteins that are responsible for the DNA uptake process.

The late genes are expressed through a common regulatory element called the *cin*-box (Campbell et al., 1998, Morrison and Lee, 2000). The *cin*-box contains a conserved 8-mer motif (TACGAATA) that represents a typical binding site for the sigma factor. This conserved motif is usually centred approximately 10 base pairs upstream from the initiation site of transcription. Later, *comW* was shown to encode a protein that mediates the post-transcriptional control of *comX* (Luo et al., 2004, Sung and Morrison, 2005). ComW is required for the stabilization and activation of the sigma factor ComX. The activation of the sigma factor ComX results in the recognition of the motif conserved in the promoter regions of the late genes.

It has been shown by Lee and Morrison (1999) that the expression of the late genes is dependent on *comX*. Several independent studies (Martin et al., 1995, Campbell et al., 1998, Dubnau, 1999, Petersen and Scheie, 2000) have identified four major classes of ComX, based on the function of the proteins expressed by the late genes.

The first class of proteins encoded by the late genes has the ability to bind all types of double-stranded DNA present in the environment, and to process and transport one strand into the cytoplasm (Dubnau, 1999). The proteins in this class are known as DNA

uptake proteins. The second class of proteins is involved in chromosomal integration, e.g., RecA, which was shown to be essential for DNA recombination (Martin et al., 1995), and SsbB, which was found to have a major effect on the transformation rate of *S. pneumoniae* (Campbell et al., 1998). These proteins are classified as recombination proteins. The third class of proteins is dispensable for transformations, and includes LytA (Mortier-Barriere et al., 1998), and the last class consists of the proteins with unknown functions (Peterson et al., 2000).

1.1.2 DNA uptake mechanism and transformation

When a competent cell is in the 'transient' state of being competent, the *comX*-encoded sigma factor activates the transcription of the late genes that are essential for DNA uptake (Lee and Morrison, 1999). A double-stranded DNA (dsDNA) that is bound at the membrane of the competent cell is transported into the cytosol by this mechanism (Dubnau, 1999, Chen and Dubnau, 2004). The DNA uptake mechanism was characterized long before the uptake genes were identified. The characterisation of the DNA uptake mechanisms was mostly done by Dubnau and his colleagues, firstly in *Bacillus subtilis*, and subsequently on *S. pneumoniae* (Dubnau, 1999, Chen and Dubnau, 2004, Chen et al., 2005).

As illustrated in Figure 1.3, the uptake mechanism is divided into three steps. The first step involves the initial binding of the exogenous dsDNA to the cell surface. In this step, a transformation pseudopilus and a surface endonuclease are required to transport and process the dsDNA. The pseudopilus comprises a major pilin-like protein (ComGC), and minor pilin-like proteins (ComGD, ComGE and ComGG). The pseudopilus is evolutionarily related to Type IV pili (Chen and Dubnau, 2004, Chen et al., 2005). The binding of the dsDNA to the cell surface has been suggested to occur via the interaction between the negatively charged DNA backbone and positively charged residues in the pseudopilus proteins (Craig and Li, 2008).

In the next step, the retraction of the ComG proteins of the pseudopilus allows the dsDNA to pass through the peptidoglycan and contact the receptor protein, ComEA (Dubnau, 1999). The receptor then delivers the dsDNA to EndA. The EndA endonuclease

degrades one strand of the dsDNA, as only a single-stranded DNA (ssDNA) can traverse through the trans-membrane channel protein ComEC. The transportation of the ssDNA through ComEC is assisted by the ComFA, which is an ATP-binding protein that mediates the ssDNA internalisation (Dubnau, 1999).

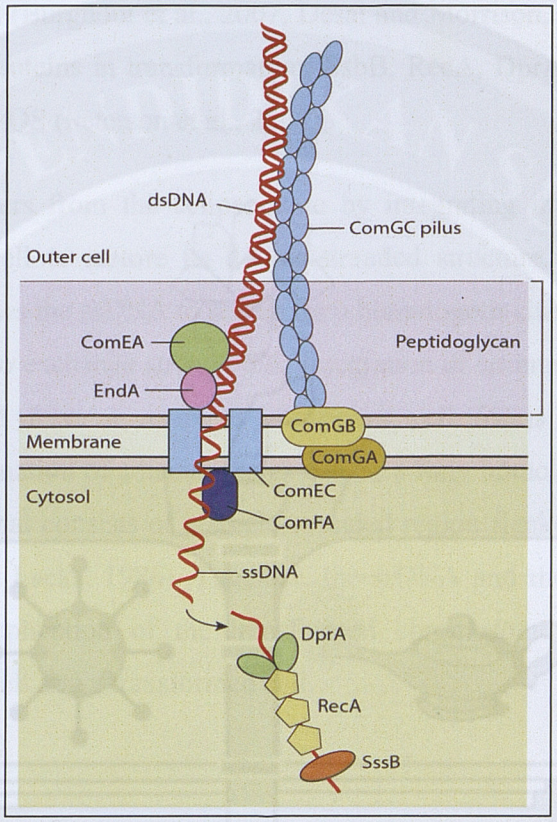

Figure 1.3: The DNA uptake and transformation mechanism. The recombination process by which the transported ssDNA is spliced into the chromosome of the recipient competent cell is shown.

Upon entry of the ssDNA into the cytosol of the competent cell, the ssDNA enters a transient state of eclipse. During this state, the ssDNA loses its ability to transform (Fox, 1957). Later, the ssDNA is bound by a set of proteins to form an eclipse complex (EC) (Morrison, 1977, Morrison, 1978). One of the major protein components of an EC is the SsbB protein, and the other components include SsbA, an SsbB paralog (Thanassi et al., 2002), DprA and RecA, which are all important for the DNA transformation process in

S. pneumoniae (Peterson et al., 2004). SsbB (probably together with SsbA) protects the ssDNA from degradation by endonucleases and assists with the recombination process (Morrison et al., 2007). DprA, a recombinase loader, binds to the ssDNA and promotes the interactions of the ssDNA with RecA (Mortier-Barriere et al., 2007) which is an essential DNA-strand exchange protein (Martin et al., 1992). Other protein components found in the EC are CoiA and RadA, which are crucial contributors to the optimisation of the recombination process (Burghout et al., 2007, Desai and Morrison, 2007). The expression of the five essential proteins in transformation: SsbB, RecA, DprA, RadA, and CoiA, is dependent on the Com QS (Peterson et al., 2004).

The EC recovers from the eclipse state by integrating into the genome of the recipient competent cell to restore its double-stranded structure. The incorporation of RecA into the EC allows the ssDNA to search for a homologous site in the chromosome of the recipient cell and to exchange strands. The integration of an internalised ssDNA into a complementary strand from a recipient competent cell forms a structure called a heteroduplex. The formation of a heteroduplex can be a fully homologous or heterologous recombination event that consists of a single-stranded region flanked by the homologous regions (Claverys and Lacks, 1986). Together, the ssDNA and the flanking regions are covalently joined. Replication of the transformed chromosome during cell division produces a wild-type cell and a transformed cell.

1.1.3 Regulation of competence via QS systems in other *Streptococcus* species

The first discovery of natural transformation occurred in 1928, when Griffith observed that non-virulent strains of *S. pneumoniae* were able to escape the immune system and cause pneumonia (Griffith, 1928). Later, the finding contributed to the discovery of DNA as the genetic material (Avery et al., 1944). Not all bacteria are naturally competent. However, the ability of a competent cell to naturally transform is unpredictably sporadic in the bacterial population, especially in the *Streptococcus* genus.

Most of the species, which are previously found to be naturally competent only in Mitis group, have also been observed in other streptococcal group: Anginosus group, Mutans group, Pyogenic group, and recently in *S. suis* (Leblanc et al., 1978, Gaustad,

1979, Perry and Kuramitsu, 1981, Havarstein et al., 1997) To date, approximately 67 prokaryotes, including 17 streptococci (Appendix C), are known to be naturally transformable (Lorenz and Wackernagel, 1994, Johnsborg et al., 2007).

1.1.4 The bacteriocin QS system: a closely related competence QS system in *S. pneumoniae*

S. pneumoniae contains a second QS regulation system (ComABCDE), known as BlpABCHR (de Saizieu et al., 2000). Both systems consist of: a peptide pheromone, encoded by *comC* and *blpC*; secretion apparatuses, *comAB* and *blpAB*; and also TCSs, *comDE* and *blpHR* (see Figure 1.4). Studies revealed that the function of BlpABCHR is to regulate the production of bacteriocin-like peptides (hence the name Blp) (de Saizieu et al., 2000, Reichmann and Hakenbeck, 2000). It has also been shown that BlpA and BlpB are highly related to ComA and ComB (ABC exporter), which have been shown to have an important role in exporting and processing the ComC peptide (de Saizieu et al., 2000). Both ComC and BlpC have the same responsibility to synthesise precursor peptide pheromones with double-glycine leaders at their N- termini.

Both QS systems appear to be regulated similarly; however, a study by de Seizieu et al. (de Saizieu et al., 2000) reported that these two QS systems appear to operate at different cell densities. The autoinduction of the Com QS system occurs in the early logarithmic phase, whereas the autoinduction of the Blp QS system occurs in the late logarithmic phase. The difference between these QS systems suggests that they operate independently of each other, probably without cross-communication at the pheromone and HK receptor levels (de Saizieu et al., 2000).

1.1.5 The importance of the Com QS system

Competent bacteria cells are able to uptake extracellular DNA fragments. The fates of the incoming DNAs are either for: i) nutritional purposes (Finkel and Kolter, 2001); ii) to facilitate the adaptation of the population through DNA repair (Michod et al., 1988, Wojciechowski et al., 1989, Hoelzer and Michod, 1991); or even iii) to increase the

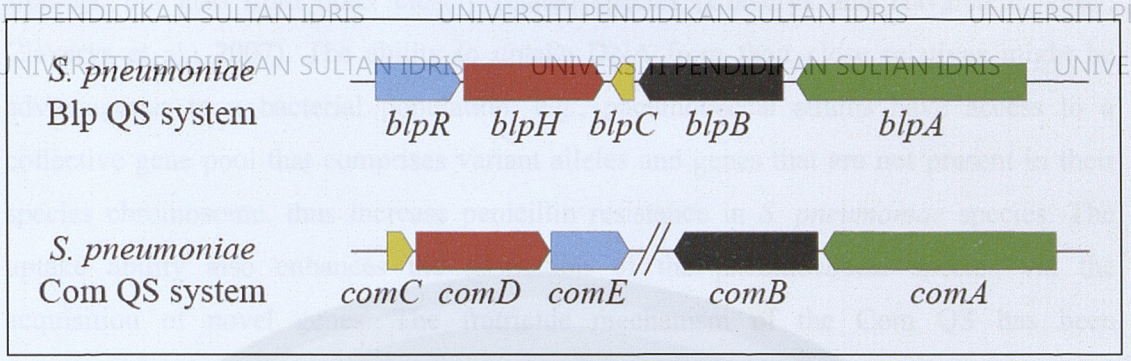

Figure 1.4: The QS regulon structure of the bacteriocin-like peptide (Blp) and competence (Com) system of *S. pneumoniae*. The homologous genes in both QS systems are coloured accordingly: ABC exporter (blpA/comA), accessory protein (blpB/comB), peptide pheromones (blpC/comC), histidine kinase (blpH/comD) and response regulator (blpR/ comE).

genetic diversity in the bacterial population (Tortosa and Dubnau, 1999, Morrison and Lee, 2000). However, the benefit of the competent bacteria cells remains unclear.

The first argument is that the incoming DNA is a source of food. Considering the complex and tightly regulated DNA uptake mechanism, as previously discussed in section 1.1.2, it is a bit crude to assume that the DNA is used as food for the bacteria. The fact that only a single strand of the extracellular dsDNA is translocated into the cytosol, while the other strand is degraded (Chen et al., 2005), would render this an inefficient food-uptake system and a metabolically costly behaviour.

Next, transformation is argued to provide advantages to the competent cells. The DNA is being taken up to repair damaged DNA and improve the population's adaptation to deleterious environmental changes. However, Redfield (Redfield, 1993) argues that the successful rate of repairing damaged DNA is low because only a very small fraction of extracellular DNA is being taken up. Additionally, bacteria cells do not usually uptake homologous DNAs that match the damaged sequence (Mongold, 1992).

The last argument seems to fit the purpose of transformation, whereby transformation promotes genetic diversity in a bacteria population (Tortosa and Dubnau, 1999, Morrison and Lee, 2000). Interestingly, a fratricide mechanism has been discovered in the *Streptococcus* species, whereby the streptococci not only uptake DNA from their