Crystallographic analysis of *Bacillus subtilis* signal peptide peptidase (SppA)

by

Sung-Eun Nam

B.Sc., Simon Fraser University, 2006

Thesis Submitted in Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Fall 2012

Approval

Name: **Sung-Eun Nam**

Degree: **Doctor of Philosophy**

Title of Thesis: **Crystallographic analysis of *Bacillus subtilis* signal peptide peptidase (SppA)**

Examining Committee:

Chair: Michael Silverman, Associate professor

Mark Paetzel
Senior Supervisor
Associate professor

Nicolas Harden
Supervisor
Professor

Edgar Young
Supervisor
Associate professor

Jack Chen
Internal Examiner
Associate professor
SFU/Department of Molecular Biology & Biochemistry

J. Thomas Beatty
External Examiner
Professor, Department of Microbiology and Immunology
University of British Columbia

Date Defended/Approved: December 6, 2012

Partial Copyright Licence

Abstract

Secreted proteins are initially synthesized in a precursor form that contains an N-terminal signal peptide for targeting the proteins to the cytoplasmic membrane. Upon translocation across the membrane via the Sec machinery, the signal peptide is cleaved off by signal peptidase. The remnant membrane embedded signal peptide is then cleaved within its hydrophobic core by signal peptide peptidase A (SppA). SppAs are membrane bound peptidases found in archaea, plant chloroplasts and bacteria. SppAs utilize a serine nucleophile and a lysine general base as catalytic residues. My PhD research has focused on the *Bacillus subtilis* SppA ($SppA_{BS}$).

During my PhD study, I solved the three-dimensional structure of $SppA_{BS}$ using X-ray crystallography, allowing us to study the difference and similarities between Gram-positive $SppA_{BS}$ and the previously solved Gram-negative *Escherichia coli* SppA ($SppA_{EC}$). Both proteins form similarly sized dome-shaped multi-subunit structures where the catalytic residues reside inside the concave portion of the dome. However, each subunit of $SppA_{BS}$ is half the size of the $SppA_{EC}$ subunit, thus $SppA_{BS}$ is an octameric complex while $SppA_{EC}$ is a tetramer. The structure revealed eight active sites in $SppA_{BS}$ where three of the protomers come together to form one complete active site whereas $SppA_{EC}$ contains only four active sites. Structural analysis on the substrate binding pockets S1 and S3, along with activity assays using a range of peptide substrates, revealed that $SppA_{BS}$ prefers substrates with leucine, arginine or tyrosine at the P1 position – a significantly different profile from $SppA_{EC}$ which prefers leucine but not arginine or tyrosine at the P1 position.

I also solved the crystal structure of $SppA_{BS}$ in complex with its own C-terminus bound within its active site. This structure has provided information on how the enzyme recognizes its substrates, confirms our previous $SppA_{BS}$ substrate preference analysis and has posed a new question “Why does $SppA_{BS}$ cleave its own C-terminus?” We show that $SppA_{BS}$ cleaves its own C-terminus in an intra-complex fashion and mutational analysis shows that Tyr331 is important for self-cleavage. I also showed that $SppA_{BS}$ is able to digest folded proteins which suggests that $SppA_{BS}$ possibly has a membrane quality control function.

Keywords: Signal peptide peptidase A; self-compartmentalized protease; Ser/Lys protease;

To my husband, Younghoon
& my daughter, Heebin

Acknowledgements

I would like to express my deep gratitude to Dr. Mark Paetzel for taking me under his wing. The knowledge, experiences and accomplishments I have achieved would have not happened without him. I am grateful for your enthusiasm, encouragement and patience.

I would like to thank my supervisory committee, Dr. Nick Harden and Dr. Edgar Young, for their advice and guidance over the years. Special thanks to Dr. Jack Chen for being my internal examiner, Dr. Michael Silverman for being the chairman and Dr. Tom Beatty for being my external examiner.

Deidre, what can I say? You have been my friend, mother, teacher, advisor, listener, proof reader and lastly the best lab manager ever! I don't think I could have survived my PhD career without you. Hopefully you will still be there forever for me.

Many thanks go to my second family (the Paetzel lab members): Kelly Kim – You have been a wonderful lab sister, thanks; Apollos Kim – Thank you for sharing your passion for the research; Linda Zhang & Suraaj Aulakh – You guys are the best bus buddies ever; Dan Chiang – Thanks for all your entertainment in the lab; Yun Luo – Thank you for always being optimistic; Charlie Stevens – Thank you for your great computer skills; Ivy Cheung – You are great talker; and special thanks to Jae Lee who gives me such insightful advice. And finally, a warm welcome to our new members, Mike and Minfei. I hope for you all the best!

And most importantly to my first family
Thank you to my husband, Younghoon Oh, for supporting me and to my daughter, Heebin, for laughter at home. Thank you Mom and Dad for your endless love and support for me.

Table of Contents

List of Tables

List of Figures

Acronyms

ATP	Adenosine triphosphate. Also called ‘molecular unit of currency’. Used as energy source within the cell.
CM	Cytoplasmic Membrane
DDM	n-Dodecyl-β-D-maltoside, a non-ionic detergent which has CMC (critical micelle concentration) of 0.15 mM
DFP	Diisopropylfluorophosphate. Serine protease inhibitor.
IMP	Inner Membrane Protein
IPTG	Isopropyl-β-D-thiogalactopyranoside.
LPS	Lipopolysaccharide, found on the outer leaflet of Gram-negative bacterial outer membranes.
MALS	Multi-Angle Light Scattering
MPD	2-Methyl-1,3 propanediol
OM	Outer Membrane
OMP	Outer Membrane Protein
PAGE	Polyacrylamide Gel Electrophoresis
PCR	Polymerase Chain Reaction, a molecular biology technique to amplify DNA sequence of interest exponentially using oligonucleotides (primers) through three different steps in a cycle; denaturing, annealing and elongation
PDB	Protein Data Bank
P(n)	Peptide. n is an integer which denotes the position of the residue of the peptide substrate. Schechter and Berger nomenclature
RMSD	Root-mean-square deviation, a measure of the differences between two different atoms superimposed. In structural biology, RMSD is used to describe how well two or more structures superimpose with each other. Lower the RMSD value, higher the structural similarity is. It also used to indicate how well the geometry of final refined structure is within the theoretical bond angles and bond lengths.
PMSF	Phenylmethanesulfonylfluoride. Serine protease inhibitor.
SDS	Sodium Dodecyl Sulphate
SDS-PAGE	Sodium Dodecyl Sulphate – Polyacrylamide gel electrophoresis
SEC	Size Exclusion Column
S(n)	Subsite or Substrate specificity pocket. n is an integer which denotes the position of the substrate binding pocket of the enzyme. Schechter and Berger nomenclature

Glossary

Å	Ångströms (10^{-10} m)
Apo-structure	Apo – ligand-free or co-factor free. Here, apo-structure is enzyme structure without substrates or inhibitors bound.
Asymmetric unit	The smallest unit of the crystal that can arrange into a unit cell with symmetry operation.
B-factor	Also known as 'temperature-factor'. A factor that applied to a degree of how much an atom oscillates or vibrate around the coordinates in the model, explained by equation, $B_i=8\pi^2U_i^2$ where U_i^2 is the mean square displacement of atom i.
Boc-Gly-ONp	tert-Butyloxycarbonyl-Glycine-O-Nitrophenyl ester. Colorimetric substrate
Cbz-Leu-ONp	Carboxylbenzyl-Leucine-O-Nitrophenyl ester. Colorimetric substrate.
Completeness	Expressed in percentage. The number of crystallographic measured in a X-ray diffraction data set in the total number of reflections present at the specified resolution.
Crystal	A regular repeat of molecules
Crystallographic refinement	A cyclic process of improving the model by monitoring $R_{crystal}$ and R_{free} values. (i.e. agreement between the molecular model and the diffracted data set) by adjustment of the model to obtain highly precise structural model that matches the measured data.
in vitro	Latin for "within glass" It refers to experimentation within test tube.
in vivo	Latin for "within the living". It refers to experimentation using a normal, living organism.
k_{cat}	Also known as "Turn over number", has unit in sec^{-1}. The maximum number of substrate molecules converted to products per active site per unit time.
kDa	Kilodalton, non-SI unit for molecular mass
K_i	Inhibition constant. Dissociation constant of inhibitor from the enzyme-inhibitor complex in competitive inhibition. Unit in M (molar concentration)
K_M	Also known as "Michaelis menten constant". Unit in M (molar concentration). The substrate concentration at which V_{max} is half.

Lipoprotein	A protein that has cysteine at the beginning of mature protein sequence which is lipid modified so that it can anchor to the membrane surface.
Matthews coefficient	Also known as specific volume (Vm). It is the crystal volume per unit of protein molecular mass and thus has a unit of $Å^3Da^{-1}$. It is calculated using formula: Vm=V/(n*M), where V is volume of the unit cell, n is the number of asymmetric units and M is molecular mass of contents of the asymmetric unit.
Mean I/ σ I	The average intensity of a group of reflections divided by the average standard deviation (sigma) of the same group of reflections. Indication of how strong the group of reflection.
Molecular replacement	A method for deriving initial phases by placing known homologous structure or the homology model built using known structure as template in the unit cell of unknown structure diffraction data set.
Peptidoglycan	A polymer made up of sugars and amino acids that forms a mesh-like layer in outside the cytoplasmic membrane of bacteria.
Periplasmic space	Space in Gram-negative bacteria between the outer and the cytoplasmic membranes.
Redundancy	There are multiple measurements of the same reflection due to 1. The symmetry of the crystal and 2. The measurement of overlapping of regions of the diffraction pattern. Redundancy is a calculation of number of measured reflection over the number of unique reflection.
$R_{crystal}$ and R_{free}	A measure of agreement between the crystal structure model and the original X-ray diffraction data. $R_{cryst} = \sum_{hkl}\left\|\|F_{obs}\| - \|F_{calc}\|\right\| / \sum_{hkl}\|F_{obs}\|$, where F_{obs} and F_{calc} are the observed and calculated structure-factor amplitudes, respectively. R_{free} is calculated using 5% of the reflections randomly excluded from refinement.
R_{merge}	A measure of agreement among multiple measurements of the same reflections. $R_{merge} = \sum_{hkl}\sum_i\|I_i(hkl) - \langle I(hkl)\rangle\| / \sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of an individual reflection and $\langle I(hkl)\rangle$ is the mean intensity of that reflection.

Pre-lipoprotein	Lipoprotein which is linked to signal peptide sequence at N-terminus but has not been lipid modified.
Pre-protein	Protein that is linked to signal peptide sequence at N-terminus.
Space Group	A mathematic description of a crystal lattice with a certain type of symmetry and a unit cell. Symmetry type is defined by a set of crystallographic symmetry operations, which may include rotation, translation, and screw axis that characterize a protein crystal. For example, space group $P2_12_12_1$. P means the unit cell is primitive. Symmetry type $2_12_12_1$ indicates that in a unit cell there is a two-fold screw axis along x-, y- and z-axis.
$SppA_{BS}$	Protease resistance fragment of *Bacillus subtilis* SppA protein which resulted in X-ray structure
$SppA_{EC}$	*Escherichia coli* SppA
Tris	Tris (hydroxymethyl) aminomethane, widely used as a component of buffer solutions. It has a pK_a value of 8.06 at 25 °C.
Unit cell	The smallest translational repeating unit of crystal.
V_{max}	The maximum reaction velocity of an enzyme at infinite substrate concentration.

1. Introduction

Bacterial signal peptide peptidase A (SppA) is a membrane bound protease that functions to cleave its membrane embedded substrate, the signal peptide (Hussain et al., 1982b). This chapter describes the environment where SppA and signal peptides are located. The background story of SppA's discovery and what was known about SppA before this thesis research started are also outlined.

1.1. Bacteria cell envelope

A bacterial cell envelope plays many important roles in the functioning of the cell. It is involved in diverse processes such as nutrition intake, waste outtake, and maintenance of the cell shape, as well as providing protection from the external environment. It is also used as a basis to divide bacteria into two large families; the Gram-positive and the Gram-negative. The way each family is distinguished from the other is their reaction to the crystal violet stain based on their cell envelope composition (Beveridge and Davies, 1983). The Gram-positive bacterial cell envelope consists of the cytoplasmic membrane and a peptidoglycan layer, also called the cell wall. The Gram-negative bacterial cell envelope is more complex, having a cytoplasmic membrane (CM) (also known as the inner membrane) and an outer membrane (OM), with a periplasmic space in between that contains a thin peptidoglycan layer (Figure 1.1). The periplasmic space has no energy sources, such as ATP (*a*denosine *t*ri*p*hosphate), but is populated by secreted proteins, lipoproteins and enzymes required for cell survival (Silhavy et al., 2010). The most extensively studied Gram-positive bacterium is *Bacillus subtilis* (*B. subtilis*), while *Escherichia coli* (*E. coli*) is the model Gram-negative bacterium.

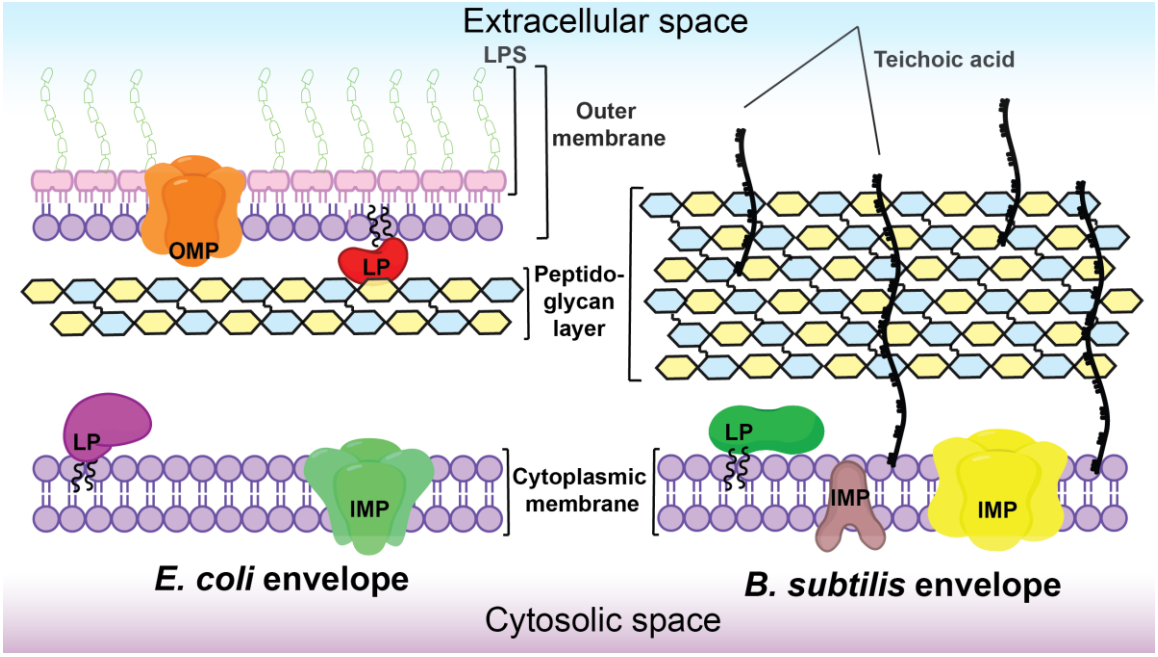


Figure 1.1. Bacterial cell envelopes.

Schematic diagram of the *E. coli* (left) and *B. subtilis* envelopes (right). The *E. coli* cell envelope has three layers, the outer membrane (OM), the peptidoglycan layer and the cytoplasmic membrane (CM). The OM is composed of phospholipid in the inner leaflet and lipopolysaccharide (LPS) in the outer leaflet. Outer membrane protein (OMP) and lipoprotein (LP) are found in the OM. LP and proteins that make channels are found in both the *E. coli* and *B. subtilis* CM, as is the inner membrane protein (IMP). *B. subtilis* does not have the OM but has a thick peptidoglycan layer. Teichoic acids are either linked to the peptidoglycan layer or the cytoplasmic membrane.

1.1.1. Outer membrane and peptidoglycan layer

The *E. coli* outer membrane is composed of an inner leaflet containing phospholipids and an outer leaflet containing glycolipids known as lipopolysaccharides (LPS) which can cause inflammation and septic shock in humans (Figure 1.1) (Raetz and Whitfield, 2002; Silhavy et al., 2010). The OM also consists of multitudes of proteins called outer membrane proteins (OMPs) as well as various enzymes (Bos et al., 2007; Fairman et al., 2011). OMPs are β-barrel proteins, many of whose functions are still unknown, although some create pores for the passive diffusion of small molecules and amino acids into and out of the cell (Nikaido, 2003). Being the outermost membrane in the *E. coli* cell, the OM acts as a barrier, providing protection to the cell from the exterior environment. *B. subtilis* lacks this OM but has a thick peptidoglycan layer that acts in a similar manner while maintaining cell structure integrity (Gan et al., 2008; Silhavy et al., 2010; Stewart, 2005). While *E. coli* has only a nanometer thick layer of peptidoglycan, in

B. subtilis, the peptidoglycan layer is 30-100 nm thick. Studies of the Gram-positive bacteria *Staphylococcus aureus* (*S. aureus*) peptidoglycan layer have shown that it contains surface proteins involved in adhesion, immune system evasion, internalization and phage binding (Silhavy et al., 2010). Since *B. subtilis* does not have an OM, it also lacks the periplasmic space.

1.1.2. Cytoplasmic membrane

The cytoplasmic membrane is a phospholipid bilayer filled with many membrane spanning proteins such as channels, proteases and chaperones (Raetz and Dowhan, 1990; Silhavy et al., 2010). The CM is where the heavy trafficking occurs in the cell, because lipids, peptidoglycans, membrane integral proteins and secreted proteins are synthesized in the cytoplasmic space and have to be transferred across the CM to various destinations including the CM, the periplasmic space, the peptidoglycan layer, the OM and the extracellular space. The CM is equipped with channels that translocate proteins, proteases that degrade misfolded proteins in the membrane, cleave or hydrolyze signal peptides, and proteins that send signals according to specific environmental changes. Most of these proteins are transported and translocated to the cytoplasmic membrane by the Sec-pathway (Dalbey et al., 2012; Tjalsma et al., 2004).

1.2. Protein secretion

In both Gram-negative and Gram-positive bacteria, secreted proteins on membrane surfaces play important roles in adherence, virulence, immunogenicity, nutrition uptake and host-pathogen interactions. In order for these secretory proteins to be delivered to the membrane surface, they have to travel through the cytoplasmic membrane. These secretory proteins contain a signal peptide at their N-terminus that facilitates the transport to and translocation across the cytoplasmic membrane by different pathways (von Heijne, 1990; Yuan et al., 2010). Once translocated, the signal peptide is cleaved off by signal peptidase and mature protein is released. The remnant membrane embedded signal peptide is subsequently cleaved by signal peptide peptidase A (Figure 1.4) (Hussain et al., 1982b).

1.2.1. Signal peptide

In the Gram-positive bacterial cell, proteins synthesized in the cytosolic space can travel to the cytoplasmic membrane, the cell wall or the extracellular space while in Gram-negative cells, proteins can move from the cytoplasmic space to the cytoplasmic membrane and beyond to the periplasmic space, to the outer membrane or the extracellular space. Proteins that are destined to locations other than the cytosolic space contain short peptide sequences called signal peptides at their N-terminus (Tjalsma et al., 2000; von Heijne, 1990; Watson, 1984).

Each signal peptide can be divided into three different regions; the N-region or N-terminal region which is positively charged with mostly lysine and arginine amino acids, the H-region or hydrophobic-region which adopts an α-helix conformation and the C-region or C-terminus region where signal peptidase recognition and cleavage sites exist (Figure 1.2) (Briggs et al., 1986; Emr and Silhavy, 1983; von Heijne, 1990). The signal peptides' amino acid composition and function are conserved in both Gram-negative bacteria and Gram-positive bacteria, although Gram-positive bacterial signal peptides are significantly longer than Gram-negative's at the hydrophobic-region (Watson, 1984).

Signal peptides can be classified into three different types based on the signal peptidase cleavage site and the export pathways. These are: 1. The Sec-SRP-dependent signal peptide; 2. Tat-dependent signal peptide; and 3. Lipoprotein signal peptide. The signal peptides are targeted by one of two major signal peptidases, Signal peptidase I (also known as leader peptidase, SPase I) and Signal peptidase II (Lipoprotein signal peptidase, SPase II) (Date and Wickner, 1981; Hussain et al., 1982a; Paetzel et al., 2002b; Tokunaga et al., 1982).

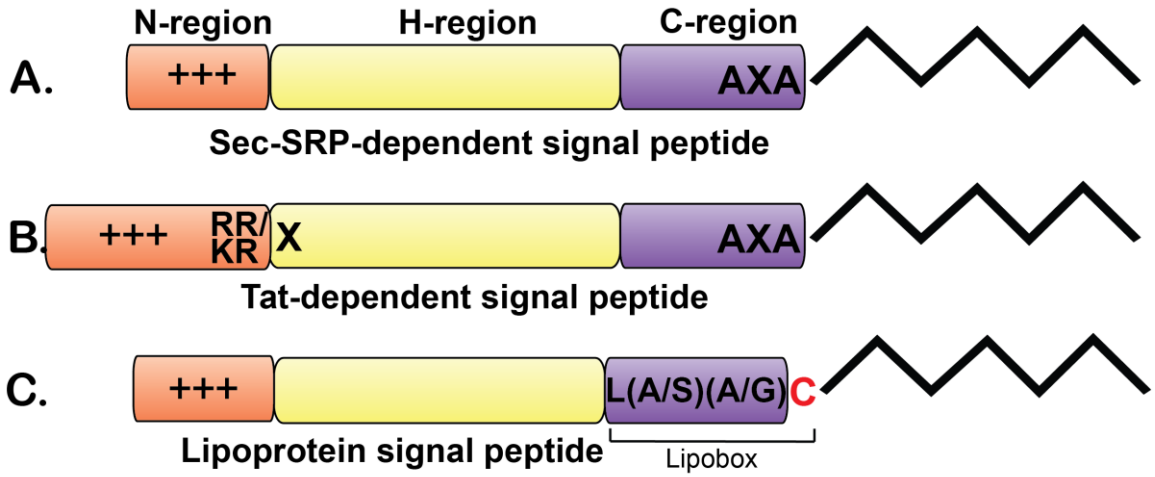


Figure 1.2. Three different types of signal peptides.

The signal peptide has three domains; the positively charged N-region (pink), the hydrophobic H-region (yellow), and the signal peptidase recognition C-region (purple). The black zigzag line represents the mature protein. Signal peptides taking either the Sec-SRP pathway **A.** or the Tat-pathway **B.** can be cleaved by type I signal peptidase, which recognizes the consensus sequence AXA. RR/KR-X is Tat-motif. **C.** Lipoprotein also has N-, H-, and C-regions as well as a type II cleavage site (lipobox is in purple with red C). Red C represent cysteine residue which gets lipid modified.

The Sec-SRP-dependent signal peptides containing the SPase I cleavage site and taking the Sec-pathway are the most abundant type of signal peptides. Signal peptides released by SPase I have the consensus sequence, AXA, at the C-region. Small, aliphatic amino acid residues, usually alanine but also valine or glycine, are preferred at the P3 (-3) and the P1 (-1) positions while the P2 (-2) position (where the X is) can accommodate almost any amino acid (Figure 1.2 & Figure 1.3) (Fikes et al., 1990; Paetzel et al., 2002b; Tjalsma et al., 2000). In *B. subtilis*, Sec-dependent signal peptides are predicted to be, on average, 28 residues in length, with 5-11 residues in the N-region and 17-18 residues in the H-region (Choo et al., 2005; Tjalsma et al., 2000).

Tat-dependent signal peptide is another type of signal peptide that is cleaved by SPase I but takes different secretion pathway, the Twin-arginine translocation (Tat) pathway (Tjalsma et al., 2000; Yuan et al., 2010). The Tat signal peptides have N-regions that are twice the length of the Sec-dependent signal peptides, with RR/KR-X##-motifs in their N-regions, where X is any amino acid and # are hydrophobic amino acids. There is no significant difference in the H-region found between the Sec- and Tat-

dependent signal peptides (Cristobal et al., 1999; Tjalsma et al., 2000). However, the H-region of the *Bacillus* signal peptides is longer and less hydrophobic than that of the Gram-negative bacteria (Tjalsma et al., 2000).

Figure 1.3. B. subtilis signal peptide examples.

Signal peptide sequences are randomly selected from (Tjalsma et al., 2000). Eight signal peptide sequences of each of three types of signal peptide are shown. Names of the protein that signal peptides belong to are on far left. Yellow highlighted regions are H-regions. **A.** Sec-SRP-dependent signal peptides. The C-region cleaved by signal peptidase I is underlined in red. **B.** Tat-dependent signal peptides. Tat-motifs are highlighted in blue. **C.** Lipoprotein signal peptides. Lipoboxes are between the red letter L and C.

The third type of signal peptide is the lipoprotein signal peptide which has a recognition site called a lipobox in the C-region (Figure 1.2 & Figure 1.3). The lipobox has a consensus sequence L(A/S)-(A/G)-C, where C is located at the +1, P1' position (first residue of the mature lipoprotein) (Braun, 1975; Sankaran and Wu, 1994). The cysteine residue is important for lipid modification before and after SPase II cleavage. The lipoprotein signal peptide is shorter than the SPase I cleavable signal peptide, with an average of four residues in the N-region and 12 in the H-region (Tjalsma et al., 2000). The examples of each type of signal peptide sequence are shown in Figure 1.3.

A previous study has shown that synthetic signal peptide is able to inhibit the translocation of precursor protein in membrane vesicles *in vitro* (Chen et al., 1987). In addition, another group has reported that the over-expression of signal peptide in *E. coli* cells result in cell 'lysis' (van der Wal et al., 1992). These findings suggest that it is beneficial for the cell to remove the remnant signal peptides from the membrane.

1.2.2. Protein secretion pathway

The pre-protein and pre-lipoprotein secretion pathway is very well conserved in both Gram-positive and Gram-negative bacteria. In *B. subtilis*, as well as in *E. coli*, the major protein secretion pathway is the Sec-SRP-dependent pathway (Driessen and Nouwen, 2008; Tjalsma et al., 2000; Yuan et al., 2010). The cytoplasmic chaperone known as signal recognition particle (SRP) is responsible for transferring the pre-proteins and pre-lipoproteins to the Sec-translocase, located in the cytoplasmic membrane. Shortly after the signal peptide exits from the ribosome, SRP binds to the nascent chains and transports the entire complex to the cytoplasmic membrane in a co-translational manner. The SRP-pre-protein complex is then recognized and the pre-protein translocated by Sec-translocase. Sec-translocase is composed of SecA (motor), a heterotrimer SecYEG (channel) and a heterodimer SecDF. SecA facilitates the translocation of the pre-protein through the channel. While SRP delivers the pre-protein during translation, there is another cytosolic chaperone, CsaA, which may be involved in targeting the posttranslated pre-protein to SecA in *B. subtilis* (Driessen and Nouwen, 2008; Muller et al., 1992). An equivalent chaperone of CsaA in *E. coli* is SecB (Bechtluft et al., 2010; Kumamoto, 1989).

Whereas the Sec-SRP-pathway exports unfolded or partially folded pre-protein, fully folded pre-proteins are transported and translocated to the cytoplasmic membrane via the Tat pathway. Only a few substrates are known to use the Tat pathway in *B. subtilis* and *E. coli.* Like the Sec-pathway, most studies on the Tat pathway have been done on *E. coli* proteins. In *E. coli*, there are three major proteins that form the Tat complex; TatA, TatB and TatC. TatA is the membrane spanning protein that forms the channel through which the folded pre-protein is translocated (Gohlke et al., 2005; Yuan et al., 2010). In *B. subtilis*, only TatA and TatC are found (Monteferrante et al., 2012). There is speculation that *B. subtilis* TatA has a dual function, playing the roles of both TatA and TatB in *E. coli* (Barnett et al., 2008). The exact function and mechanism of *B. subtilis'* Tat pathway is still under investigation.

1.2.3. Signal peptide released by Type I and II signal peptidase

Once pre-protein or pre-lipoprotein is translocated across the cytoplasmic membrane, the N-terminus of the precursor is exposed to the cytoplasmic space while the mature protein faces the exterior of the cytoplasmic membrane. This follows the "positive inside rule" where the N-region of the signal peptide is positively charged, therefore remaining in the cytosolic space (von Heijne, 1992). The signal peptide is then removed by a signal peptidases and this releases the secreted protein into the periplasm (in Gram-negative bacteria) or the extracellular space (in Gram-positive bacteria).

The signal peptide of the secreted pre-protein is cleaved at the C-region of the signal peptide by SPase I (Figure 1.2.A). *E. coli* SPase I is a membrane-integral protease with two N-terminal transmembrane segments spanning the cytoplasmic membrane and a soluble domain exposed to the periplasmic space (Paetzel et al., 2002a; Paetzel et al., 2002b). The soluble domain of SPase I has been solved using X-ray crystallography revealing that SPase I has a serine as the nucleophile and a lysine as the general base (Paetzel et al., 2002a). Unlike *E. coli* where only one gene encoding SPase I is found, *B. subtilis* has five SPase I – encoding genes in the chromosome; SipS, SipT, SipU, SipV and SipW. All five SPase Is appear to have overlapping activity, however, their preference for type of substrate and their efficiencies are different (van Roosmalen et al., 2004). A number of substrates used in each of Sip proteins have been

identified but structural information and the reason for multiple SPase Is in *B. subtilis* remains unknown.

For the pre-lipoprotein, lipid modification at the cysteine residue in position P1' is necessary for SPase II to release the signal peptide (Figure 1.2.B) (Sankaran and Wu, 1994). A diacylglyceryl is added to the cysteine residue of the pre-lipoprotein at the +1 position by prolipoprotein diacylglyceryl transferase (Lgt). This step is required before SPase II can cleave off the signal peptide. The mature lipoprotein is then further modified by lipoprotein aminoacyl transferase (Lnt), which adds an amide-linked fatty acid onto the same cysteine residue that was previously modified by Lgt (Vidal-Ingigliardi et al., 2007).

SPase II is a membrane-bound protease with four transmembrane segments. It has been proposed to have an Asp/Asp catalytic dyad with the two catalytic aspartic acids, located in the transmembrane segments, sitting on the exterior of the cytoplasmic membrane (Hussain et al., 1982a; Tjalsma et al., 1999). The exact mechanism and structure of SPase II still needs to be elucidated.

The residual signal peptides that are cleaved off from pre-protein or pre-lipoprotein by SPase I or SPase II and still present in the lipid bilayer are further degraded by Signal peptide peptidase A in bacteria (Figure 1.4.C) (Hussain et al., 1982b).

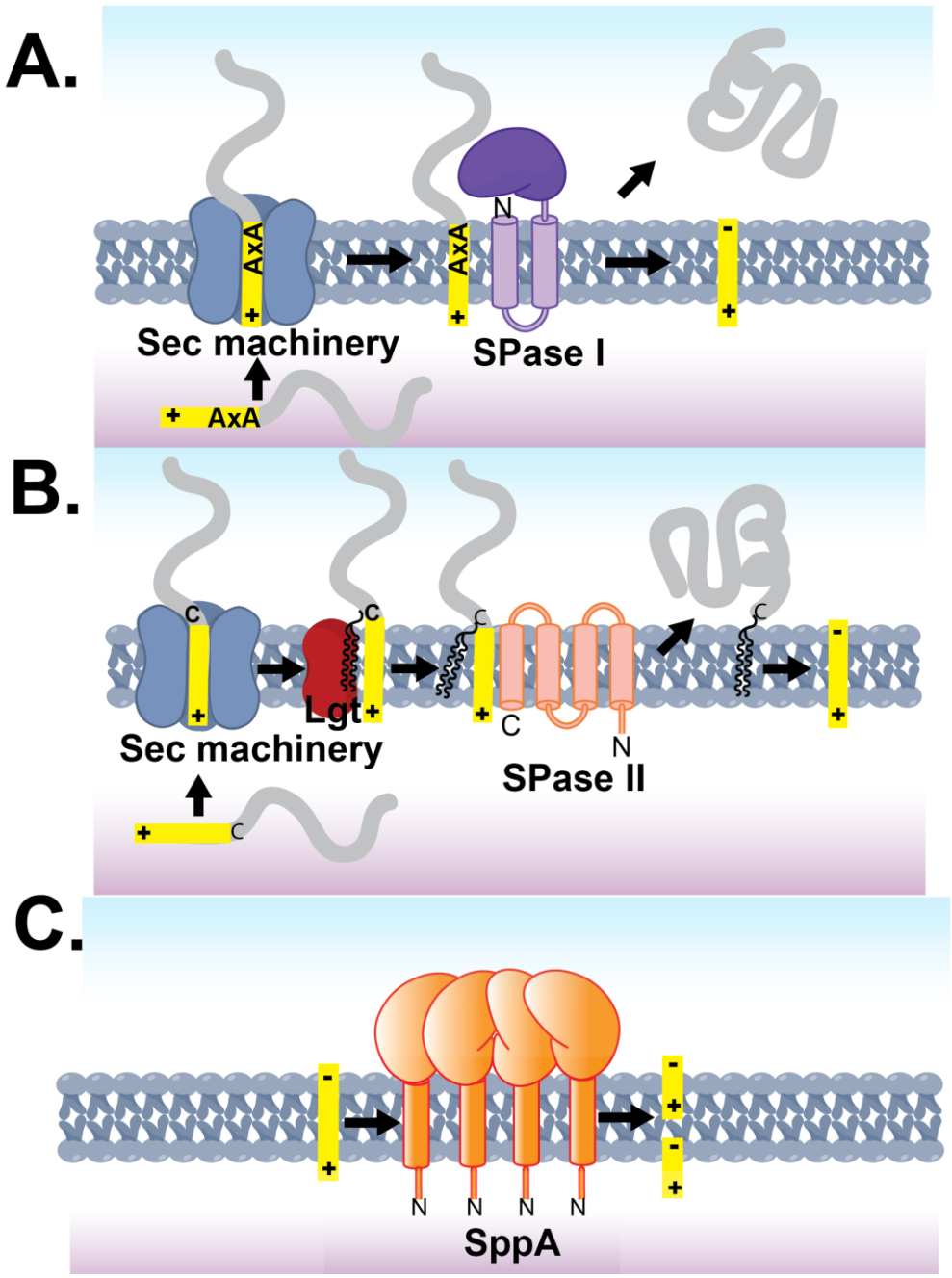


Figure 1.4. Schematic diagram of pre-protein and pre-lipoprotein translocation and signal peptide hydrolysis.

Schematic diagram of pre-protein and pre-lipoprotein translocation in the cytoplasmic membrane by Sec machinery (blue tunnel). Signal peptide (in yellow) targets pre-protein and pre-lipoprotein to the Sec machinery for translocation. Signal peptide has positive charge at N-terminus. Mature protein is in grey curved line. **A.** Pre-protein signal peptide is cleaved by SPase I (Purple). AXA in yellow box is SPase I recognition site. **B**. Pre-lipoprotein is lipid modified at the cysteine residue (Represented as "C" in part of the mature lipoprotein (grey)) by Lgt (red) and then the signal peptide is cleaved by SPase II (pink). **C.** The embedded signal peptide, which now has positive charge at N-terminus and negative charge at C-terminus, is cleaved by SppA (orange).

1.2.4. Signal peptide hydrolysis

1.2.4.1. *Escherichia coli* signal peptide peptidase A

In 1981, Protease IV, a protease from the outer membrane of *E. coli* reported to be 23,500 Da in size, was discovered to cleave αS1 Casein using endoproteolytic activity (Régnier, 1981). The following year, another group claimed that Protease IV was located in the cytosolic membrane of *E. coli* and that it was 34,000 Da in size (Pacaud, 1982). However, both studies agreed that Protease IV digests casein, albeit at a slow rate. Protease IV was also observed to cleave small synthetic substrates such as Boc-Ala-ONp and Cbz-Leu-ONp (Pacaud, 1982; Régnier, 1981).

Also in 1982, an unexpected discovery was made during the study of signal peptide cleavage from pre-lipoprotein by SPase II (Hussain et al., 1982a; Hussain et al., 1982b). A membrane extract, treated with globomycin (an SPase II inhibitor) prior to isolation, was washed with buffer to remove the globomycin. This resulted in the release of the signal peptide from the pre-lipoprotein by SPase II; however, when run on a gel, only the mature lipoprotein was seen, not the signal peptide, suggesting that another protease digests the signal peptide. This new protease was tentatively named signal peptide peptidase (SppA) (Hussain et al., 1982b).

A similar study demonstrated that the released signal peptide from pre-lipoprotein is digested by the new protease SppA, and that this reaction is inhibited by antipain, leupeptin, chymostatin, and elastinal in concentration ranges of 2-5 mM (Ichihara et al., 1984). Soon after, an SppA gene was cloned, sequenced, and the protein purified as a 67,241 Da protease (Ichihara et al., 1986). A cross linking experiment was carried out to confirm that SppA forms a tetramer. Ichihara *et al*. speculated that Protease IV is SppA because they share the same properties, and suggested the discrepancy in molecular mass could be possibly due to insufficient amounts of Protease IV purified along with larger amounts of a small molecular mass protein contaminant (Ichihara et al., 1986).

Despite its important role in digesting remnant signal peptides in the membrane, SppA is not essential for *E. coli* cell viability (Suzuki et al., 1987). Further studies on SppA determined that it digests signal peptides into fragments as small as three

peptides long and prefers to cleave substrates that have hydrophobic amino acids at the primary position with adjacent hydrophobic amino acids (Novak and Dev, 1988).

Twenty years after the discovery of SppA, solution of its structure using X-ray crystallography and mutagenesis studies of *E. coli* SppA confirmed that SppA is a Ser/Lys catalytic dyad protease, which forms a dome-shaped structure by assembling into a tetramer (Kim et al., 2008; Wang et al., 2008). Mutagenesis experiments are consistent with *E. coli* SppA utilizing Ser409 as the nucleophile and Lys209 as the general base (Wang et al., 2008). *E. coli* SppA contains two domains which are structurally a tandem repeat, although the sequence identity between the N-terminal and C-terminal domains is only 18 %. There are four separate active sites, each located at the interface between the N- and C- terminal domains. The catalytic serine arrives from the C-terminal domain and the catalytic lysine arrives from the N-terminal domain of the same molecule (Kim et al., 2008). The soluble domain of SppA exists in the periplasmic space of *E. coli*'s cell envelope, with the wide opening of the dome facing towards the inner membrane. Wang *et al*. reported that SppA has activity towards the synthetic peptide, Cbz-Val-ONp, which agrees with previous studies that SppA cleaves hydrophobic residue containing substrate. However, to date, *in vivo* studies have not been performed to support the function of SppA as a signal peptide peptidase (Wang et al., 2008).

1.2.4.2. *Bacillus subtilis* signal peptide peptidase A

Two proteins from *B. subtilis*, SppA and TepA (translocation-enhancing protein A), have been found to have similar functions to the *E. coli* SppA, and have sequence identities of 25% and 16% to *E. coli* SppA, respectively. They are not an essential genes since SppA or TepA knockout strains of *B. subtilis* are viable and grow normally. Transcription studies indicate that SppA is growth phase-dependent, meaning that the SppA transcription level increases after exponential growth phase is reached, and continues to increase once the post-exponential growth phase is attained. However, the TepA gene is continuously transcribed throughout the different growth phases (Bolhuis et al., 1999).

B. subtilis strains which lack either SppA or TepA genes showed an increased accumulation of pre-protein compared to the control strain which had both genes present.

This experiment shows that SppA and TepA are required for the processing of signal peptides from pre-proteins. *B. subtilis* SppA is a membrane bound protein like *E. coli* SppA and has the same Ser/Lys catalytic dyad, with serine acting as the nucleophile and lysine as the general base (Bolhuis et al., 1999). Based on topology predictions and analogy with *E. coli* SppA, the catalytic domain of *B. subtilis* SppA is expected to be located on the extracellular-side of the cytoplasmic membrane with a single N-terminal transmembrane segment anchoring it to the cytoplasmic membrane (Wang et al., 2008). *B. subtilis* SppA is approximately one half the size of *E. coli* SppA and alignes to the C-terminal domain than to the N-terminal domain of *E. coli* SppA (Figure 1.5). Being half the size of the *E. coli* SppA, it was speculated that *B. subtilis* SppA may form a similar dome-shaped structure with eight molecules instead of the four in *E. coli* SppA (Kim et al., 2008).

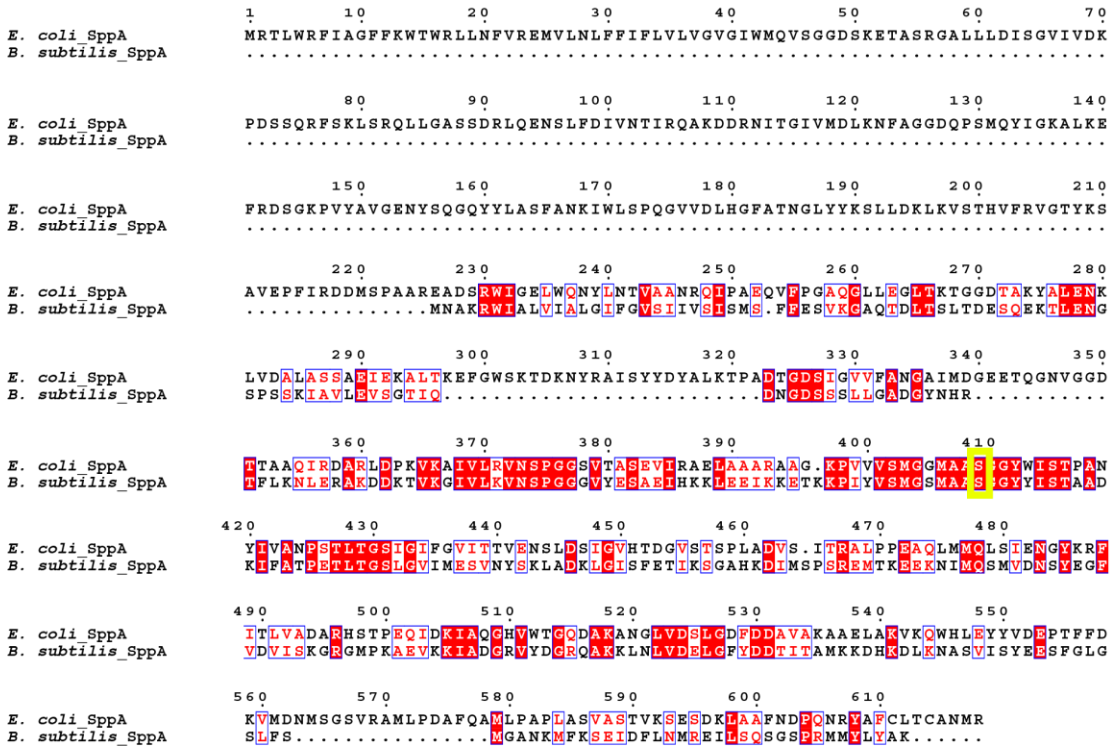


Figure 1.5. B. subtilis SppA and E. coli SppA sequence alignment.

B. subtilis SppA (Accession number: O34525) and *E. coli* (Accession number: P08395) SppA sequence alignment is shown. Red highlights indicate identical residues. Nucleophile serine is indicated in yellow box.

Unlike SppA, TepA is a cytosolic protease that is only found in Gram-positive bacteria. Based on the protein sequence alignment, Bolhuis *et al*. (1999) claimed that TepA is likely to utilize the Ser/His/Asp catalytic triad like ClpP, because despite the fact that the sequence identity is only 17%, the Ser/His/Asp triad is conserved (Figure 1.6). ClpP is a self-compartmentalized protease (two-homo-heptamer) found in both *B. subtilis* and *E. coli* which functions to degrade misfolded proteins in the cytoplasmic space (Sauer and Baker, 2011). If TepA is like ClpP, it could also assemble into a homo-multi subunit complex and may have functions other than processing signal peptides.

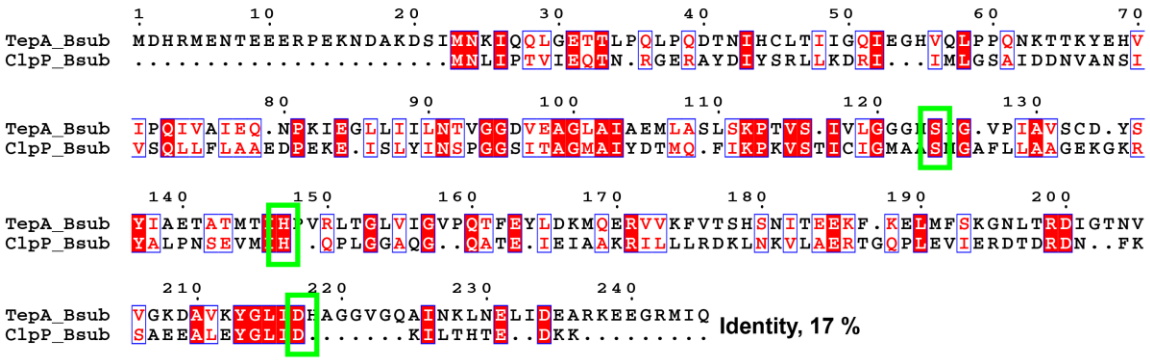


Figure 1.6. B. subtilis TepA and ClpP protein sequence alignment.

Red highlights indicate identical residues. The catalytic residues are boxed in green. The sequence alignment of TepA (UniProt accession number: Q99171) and ClpP (UniProt accession number: P80244) shows conserved Ser/His/Asp. Sequence identity is shown at the end of the alignment.

1.3. Other self-compartmentalized proteases in bacteria

As mentioned previously, *E. coli* SppA assembles into a self-compartmentalized tetramer with active sites buried inside the dome. SppA's only known function is to cleave signal peptides; however, other self-compartmentalized proteases found in bacteria are primarily involved in the degradation of abnormal proteins. These proteases include Lon proteases, ClpP, HslVU, DegP, FtsH and DppA (Table 1.1). These proteins are found in both Gram-negative and Gram-positive bacteria, playing important quality control roles in the cytoplasm, periplasm and cytoplasmic membrane (Gur et al., 2012; Sauer and Baker, 2011). Lon proteases, ClpP, FtsH and HslVU are ATP-dependent proteases that either form a complex with a subunit of ATP-chaperone or have an ATPase domain built within the protease domain. The ATP subunit and domain work as substrate recognition and unfolding units prior to digestion by the protease components of the complexes. Lon proteases, ClpP and HslVU degrade misfolded and unfolded proteins in the cytoplasm whereas FtsH is responsible for digesting misfolded and damaged membrane proteins on the cytoplasmic membrane (Langklotz et al., 2012).

Table 1.1. List of compartmentalized proteases in bacteria.

Protease	Gram -	Gram +	AAA	Oligomeric state	substrate	Localization	mechanism
Lon (La)[a]	Yes	Yes	Domain linked to protease domain	Hexamer	Misfolded, unfolded, mislocalized, damaged proteins	cytoplasm	Ser/Lys
ClpP (caseinolytic protease)[b]	Yes	Yes	Hexamer ClpA/X (-) ClpC/X (+)	2Xheptamer	Misfolded, unfolded, mislocalized, damaged protein with SsrA (11 amino acid tag)	Cytoplasm	Ser/His/Asp
HslVU(ClpQY) Heat shock locus[c]	Yes	Yes	Hexamers (HslU)	2Xhexamer	Misfolded, unfolded, mislocalized, damaged proteins	Cytoplasm	Threonine
HtrA (DegP) (High temperature requirement A)[d]	Yes	No	No AAA but PDZ domain	Hexamer as chaperon 12 to 24 when protease	Misfolded, unfolded, mislocalized, damaged proteins	Periplasm	Ser/His/Asp
FtsH (filamentation temperature sensitive H)[e]	Yes	Yes	Domain linked to protease domain	Hexamer	Intergral membrane proteins, cytoplasm proteins with SsrA	Membrane bound and large cytoplasmic domain	Zn^{2+}
DppA (D-aminopeptidase)[f]	No	Yes	No	Homo decamer	D-Ala-Ala D-Ala-Gly-Gly	cytoplasm	Zn^{2+}

[a], Gur *et al*. Protein science. (Gur et al., 2012)
[b], Butler *et al*. Molecular microbiology. (Butler et al., 2006)
[c], Butler *et al*. Molecular microbiology. (Butler et al., 2006)
[d], Krojer *et al*. Nature. (Krojer et al., 2002)
[e], Langklotz *et al*. Biochimica et Biophysica Acta. (Langklotz et al., 2012)
[f], Remaut *et al*. Nature structural biology. (Remaut et al., 2001)

FtsH is an essential gene for viability in *E. coli* and is similar to SppA in that it is a membrane integral protease with a large catalytic soluble domain that self-compartmentalizes (Langklotz et al., 2012; Tomoyasu et al., 1993). It has two transmembrane segments at the N-terminus and soluble domain is located on the cytosolic side of the cytoplasmic membrane. FtsH is a single polypeptide protein with an ATPase domain linked to the protease (catalytic) domain. FtsH forms homo-hexameric structure where the catalytic domain is stacked underneath the ATPase domain, creating a chamber between the ATPase disc and compartmentalized catalytic domain (Dalbey et al., 2012; Suno et al., 2006). The ATPase domain is further divided into two domains: a smaller and larger domain with the ATP binding site located on the interface between the two. (Suno et al., 2006). ATPase activity allows the substrate to be unfolded and translocated, through the small axial pore created by the hexamer, into the protease chamber. Subsequently, the substrate is digested into smaller peptides, 6-25 amino acids in length, by the protease domain where catalysis depends on Zn^{2+} ions (Langklotz et al., 2012).

Another quality control, compartmentalized protease that does not depend on the ATP-adapter for recognition is the periplasm DegP (HtrA, *H*igh *t*emperature *r*equirement *A*) protease which has a PDZ domain instead of an ATPase domain (Iwanczyk et al., 2007; Krojer et al., 2002). DegP has a dual function; chaperone and protease. As a chaperone, it binds outer membrane proteins in the periplasm to prevent aggregation, and as a protease it degrades abnormal proteins. DppA's (D-aminopeptidase's) function is to hydrolyze the N-terminal end of D-amino acid-containing peptides (Remaut et al., 2001). Interestingly, SppA has neither an ATP-adapter nor a PDZ domain. The active sites of all compartmentalized proteases are located in the interiors of the oligomers, but the catalytic mechanism varies from the Ser/His/Asp catalytic triad, to the Ser/Lys catalytic dyad, to metal-based catalysis (Raju et al., 2012).

1.4. SppA utilizes a Ser/Lys catalytic dyad mechanism

1.4.1. Ser/Lys dyad protease

Protease functions include degrading misfolded proteins, cleaving precursors to produce signaling molecules, digesting extracellular toxins and more. Proteases can be classified into four different protease groups depending on what is used for the nucleophile in the catalytic mechanism; metallo-proteases, serine proteases, aspartic acid proteases and cysteine proteases (Rawlings and Barrett, 2000). Some serine proteases use Ser/His/Asp as their catalytic triad (Ekici et al., 2008). Trypsin, subtilisin, chymotrypsin, ClpP and DegP are examples of Ser/His/Asp utilizing proteases. Serine acts as the nucleophile to attach the carbonyl of the scissile bond, histidine is the general base that removes a proton from the hydroxyl group of the serine, therefore activating the serine. Aspartic acid orients the histidine and neutralizes the general base throughout the transition state. Serine proteases that utilize Ser/Lys as their catalytic dyad are different from Ser/His/Asp protease that Ser/Lys proteases are not inhibited by the well-known serine protease inhibitors PMSF or DFP (Paetzel and Strynadka, 1999; Tschantz and Dalbey, 1994). PMSF and DFP irreversibly inhibit proteases that utilize serine as nucleophile (James, 1978; Jansen and Balls, 1952).

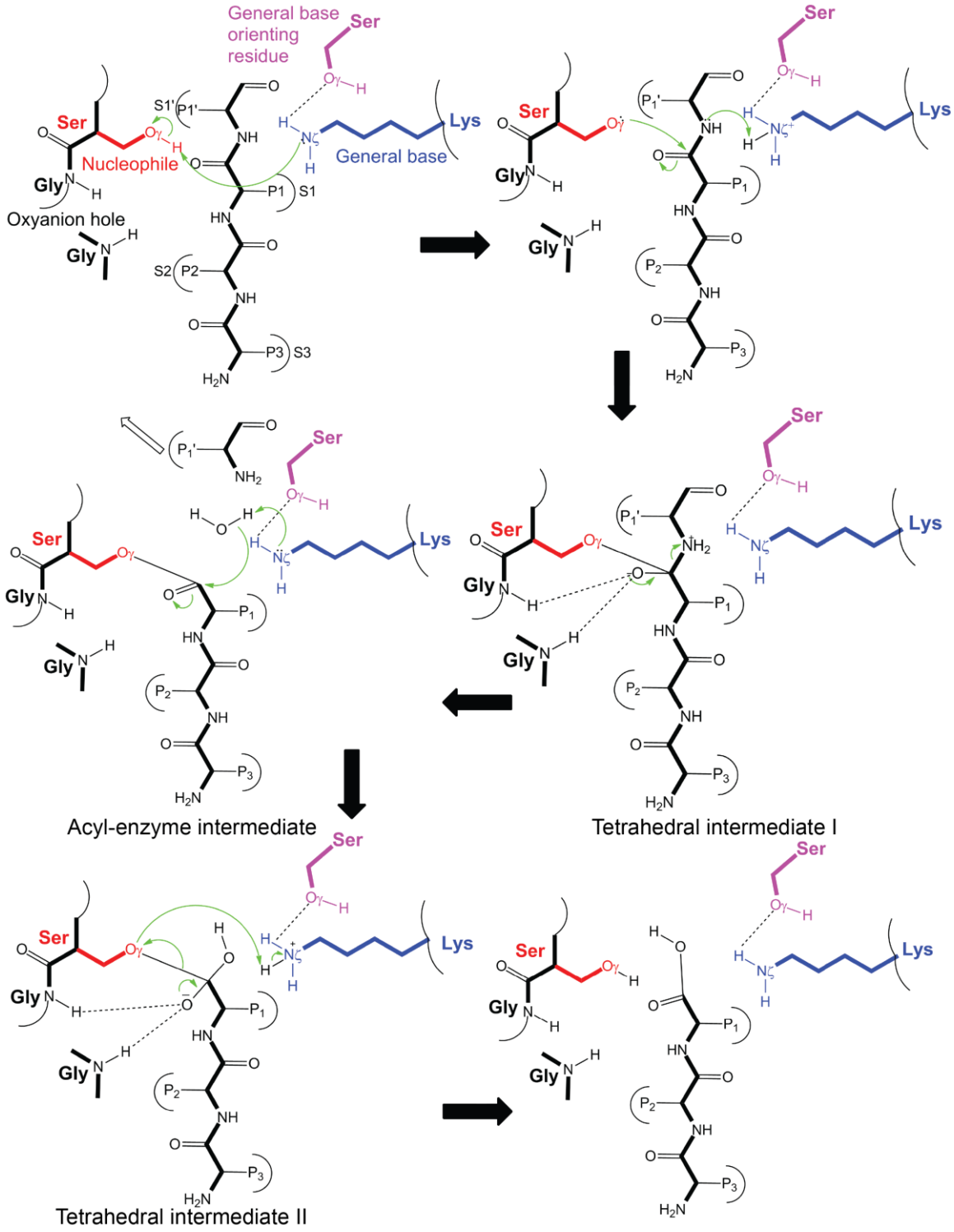


Figure 1.7. (Figure legend on next page)

Figure 1.7. Proposed Ser/Lys catalytic dyad mechanism.

Serine (red) is the nucleophile, lysine (blue) is the general base and serine (purple) is general base orienting residue. The green arrow represents electron movement and dashed lines represent hydrogen bonds. Substrate specificity pocket is represented in arch around P1, P2 and P3 position of the substrate. Nζ of lysine takes a proton from Oγ of serine which then attacks the carbonyl carbon. Tetrahedral intermediate I is formed and negatively charged oxygen is stabilized by hydrogen bonds formed from main chain NH group of two glycine (oxyanion hole). After the C-terminal group of substrate leaves and acyl-enzyme intermediate forms. The deacylating water comes in and gets deprotonated by Nζ of lysine and becomes hydroxide (now nucleophile). With hydroxide, tetrahedral intermediate II is formed. The proton from Nζ is then donated to Oγ of serine to return the proton to its native form.

Proteases that have a Ser/Lys catalytic dyad are SPase I, UmuD, Lon, LexA, cIλ, VP4 and SppA (Botos et al., 2004; Feldman et al., 2006; Kim et al., 2008; Luo et al., 2001; Paetzel and Dalbey, 1997; Paetzel and Strynadka, 1999). Studying solved structures has revealed that the substrate of Ser/Lys dyad proteases, SPase I and UmuD, interacts with the protease binding site in the anti-parallel β-sheet and the active serine attacks a carbonyl carbon to the *si*-face of the scissile amide bond (Paetzel and Strynadka, 1999). P1 and P3 residues are accommodated via S1 and S3 of the substrate specificity pocket, respectively, and P2 and P4 of the substrate face the solvent (Schechter & Berger nomenclature (Schechter and Berger, 1967)). The Ser (nucleophile) Oγ is within hydrogen bonding distance of the Lys (general base) Nζ (Figure 1.7) Lysine normally has a pK_a value of 10.5 in the polypeptide but in Ser/Lys dyad proteases, it is buried in the hydrophobic environment of the active site which lowers the pK_a value (Paetzel et al., 1997). As a result, it is able to activate the serine residue under physiological conditions. In addition to the catalytic residues serine and lysine, a hydroxyl group – containing residue (either threonine or serine) is found also coordinating the general base lysine (Ekici et al., 2008). During the hydrolysis of the substrate, at transition state, negatively charged oxygen of tetrahedral intermediate I & II, is thought to be stabilized by the oxyanion hole created by NH group of the main chain (Figure 1.7) (Paetzel et al., 2002b).

1.4.2. S49 family of proteases

MEROPS is the peptidase database where proteases are classified according to their amino acid sequence and structure similarities (Rawlings and Barrett, 1999; Rawlings and Barrett, 2000). Proteases' amino acid sequence similarities and type of catalytic residues divide them into different families. For example, bacterial SppA utilizes Ser/Lys as its catalytic dyad, so it belongs to the S49 family while human Spp, which has an Asp/Asp catalytic dyad belongs to the A22 family. Each family has subfamilies that share an ancient divergence. For the S49 family, there are three subfamilies; S49A, S49B and S49C. Gram-negative SppA, Gram-positive SppA, Chloroplast SppA (*Arabidopsis*-type), and Archaeal SppA2 and *E. coli* SohB are members of the S49A subfamily. Gram-negative SppA and Chloroplast SppA are both approximately 60-70 KDa (~610 amino acids) (Figure 1.8). In contrast, Gram-positive SppA, Gram-negative SohB and *Thermococcus kodakaraensis* (Archaea) SppA2 are half the size, ~33 kDa (~330 amino acids) (Figure 1.9) (Bolhuis et al., 1999; Matsumi et al., 2005; Wetzel et al., 2009). Some Gram-negative bacterial SppAs are also half the size of Gram-negative SppA. SohB is the suppressor of HtrA (DegP), meaning that bacteria which are heat sensitive due to the mutation in DegP are able to survive when a plasmid containing SohB is present in the cell (Baird et al., 1991). As mentioned previously, *E. coli* SppA and *B. subtilis* SppA show processing of signal peptides, however the functions of Chloroplast SppA, SohB, and Archaeal SppA2 remain unknown. The S49B subfamily includes bacteriophage lambda protein C (UniProt accession number: P03711), a prohead protease involved in the cleavage of structural proteins during phage capsid assembly (Liu and Mushegian, 2004). S49C subfamily consists of another archaea SppA (*Pyrococcus horikoshii*), SppA1 (UniProt accession number: O59279), a stomatin-like protein which in humans, appears to function as a negative regulator of univalent cation channels (Yokoyama et al., 2006). However, its function in archaea is undetermined. The function, structure, oligomeric states and localization of these proteases are still a mystery but it is thought that they all have the same catalytic mechanism, with serine acting as the nucleophile and lysine acting as the general base.

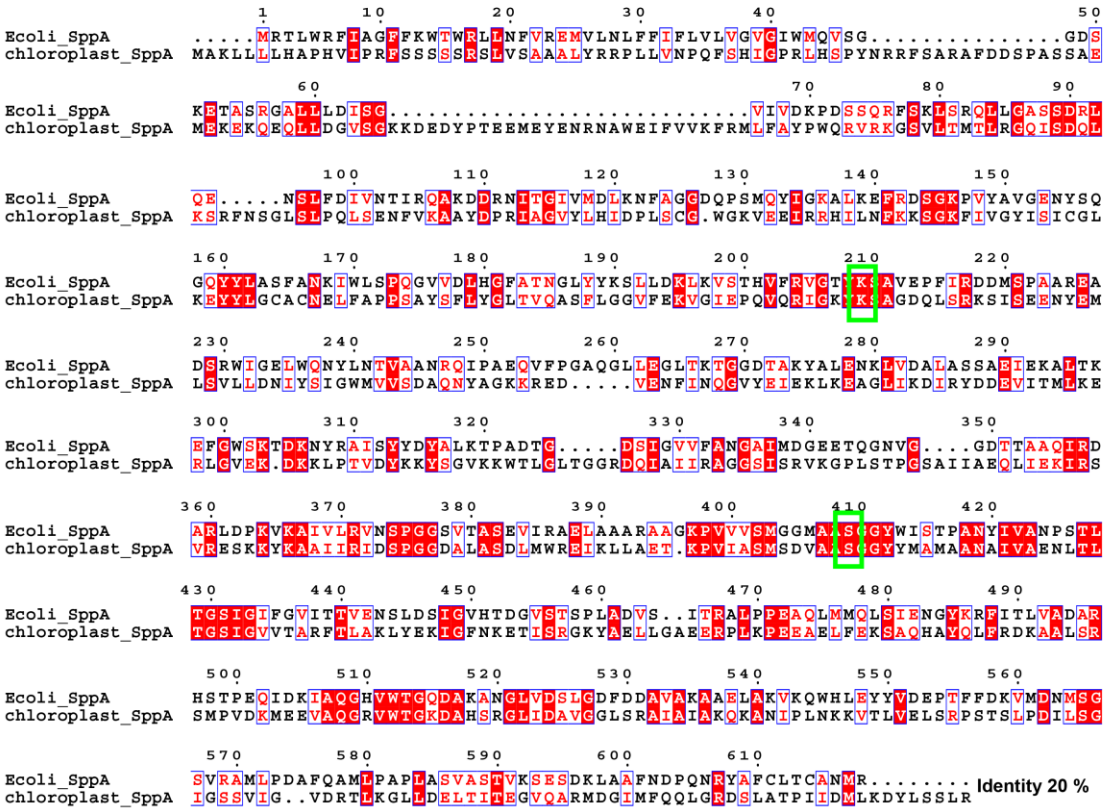


Figure 1.8. Sequence alignment of S49A subfamily that have longer sequence.

Red highlights indicate identical residues. The catalytic residues are boxed in green. The sequence alignment of chloroplast of *Arabidopsis* SppA (UniProt accession number: Q9SEL8) and *E. coli* SppA (UniProt accession number: P08395). Sequence identity is shown at the end of alignment.

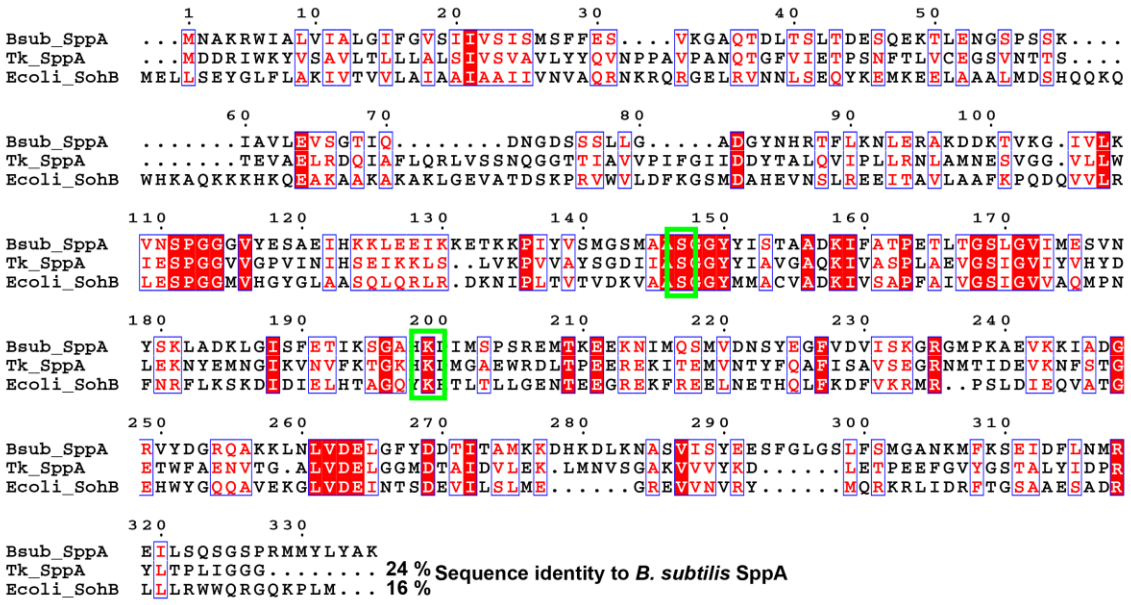


Figure 1.9. Sequence alignment of S49A subfamily that have shorter sequence.

Red highlights indicate identical residues. The catalytic residues are boxed in green. The sequence alignment of *B. subtilis* SppA (UniProt accession number: O34525), *Thermococcus kodakaraensis* SppA (UniProt accession number: Q5JE91) and *E. coli* SohB (UniProt accession number: P0AG14). Sequence identity to *B. subtilis* SppA is shown at the end of alignment.

1.5. Research objectives

The molecular structure and biochemical properties of *B. subtilis* SppA were unknown at the beginning of this PhD project. The following research questions were posed and have been answered:

A. What is the three-dimensional molecular structure of *B. subtilis* SppA?

1. What is the oligomeric nature of *B. subtilis* SppA?

2. How is the active site of *B. subtilis* SppA assembled?

3. How do the active sites and substrate binding pockets of *B. subtilis* SppA differ from those seen in *E. coli* SppA?

B. What are the substrate preferences of *B. subtilis* SppA?

2. Crystal Structure of *Bacillus subtilis* SppA

Note regarding contributions:

Portions of this chapter were published previously in the Journal of Molecular Biology.

Nam, S.E., Kim, A.C. & Paetzel, M. (2012). Crystal structure of *Bacillus subtilis* signal peptide peptidase A. *J. Mol. Biol.* **419**, 347-358

Escherichia coli SppA kinetic data in Section 2.3.11 were generated by Apollos Kim.

2.1. Overview

In order to study the structure and activity of *Bacillus subtilis* SppA, the SppA gene was PCR amplified and subcloned into a protein expression vector with an intact N-terminus six-histidine tag for Ni^{2+}-NTA affinity purification. Three different constructs were created. Full length SppA and the predicted transmembrane segment deleted (residues 1-25) SppA did not over-express resulting in inadequate amounts of protein to carry out further experiments such as enzymatic assays or crystallography. However, a further truncated construct with residues 2-54 deleted, purified enough to test for activity and set up initial crystallization screens. When the general base Lys199 was mutated to Ala (K199A) in the above three constructs (full length, Δ1-25, and Δ2-54), the expression levels increased dramatically, resulting in larger amounts of protein compared to the wild type (Δ2-54). Initial hits from the crystallization of the K199A constructs were obtained but they were low resolution diffracting crystals. In order to improve the crystallization so that the crystal would diffract to higher resolution, SppA proteins were subjected to limited proteolysis using different proteases. Limited proteolysis removes flexible regions such as the N- and C-terminus ends or the loop, leaving the protein with only the rigid domain which is likely to crystallize in a more orderly manner. The SppA proteins were digested by thermolysin and separated by size-exclusion chromatography, resulting in protease resistant fragments which produced an initial hit. The initial hit was optimized yielding a plate shaped crystal which diffracted to 2.4 Å resolution. The diffraction data were collected at the synchrotron (Canadian light source, CLS). The homologous structure of SppA from *Escherichia coli* had been solved previously, so the C-terminal domain of Chain A from *E. coli* SppA was used as a template to build the homology model with *B. subtilis* SppA's protein sequence. The molecular replacement method was used to obtain initial phases.

In this chapter, we report the first X-ray crystal structure of a Gram-positive bacterial SppA. The thorough structural analysis of *B. subtilis* SppA and the structural similarities and differences between the *B. subtilis* and *E. coli* SppAs are analyzed

extensively from the overall structure to the active sites and the substrate specificity pockets.

In addition to the structural analysis, the enzymatic activity of *B. subtilis* and *E. coli* SppAs have been studied. Fluorogenic substrates were utilized to identify the substrate preferences and the fluorogenic substrate used in these experiment has a short protein peptide link to the fluorescent moiety, *M*ethyl *C*umaryl *A*mide (MCA). Increases in the concentration of MCA were measured fluorometically to determine how much of peptide-MCA is cleaved, by the enzyme, between the peptide and MCA. The same range of peptide-MCA substrates were screened to determine the substrate preference for *B. subtilis* and *E. coli* SppA, and the results were consistent with the S1 and S3 substrate specificity pockets identified from the structures. Furthermore, the SppAs' proteinase activity was tested by incubating SppAs with lipoproteins from *E. coli.*

2.2. Materials and methods

2.2.1. Cloning and mutagenesis

The *B. subtilis* sppA gene, lacking the nucleotides that code for residues 1 to 25 ($SppA_{BS}^{\Delta 1-25}$), was amplified using PCR. The oligonucleotides used to amplify this construct were; forward primer 5'-gccatatgagtttctttgaaagcgtcaaaggc-3' and reverse primer 5'-ctcgagctacttcgcatagagatacatcattct-3'. The $SppA_{BS}^{\Delta 1-25}$ construct was cloned into the expression vector pET28b+ (Novagen) using the NdeI and XhoI restriction sites. Using the above construct as a template, the codon for the general base residue (Lys199) was mutated to a codon for an alanine residue (underlined) following the Quick-change site-directed mutagenesis procedure. The oligonucleotides used to perform the site-directed mutagenesis were; forward primer 5'-agcggggcccatgcggacattatgtct-3' and reverse primer 5'-agacataatgtccgcatgggccccgct-3'. The native and K199A mutant genes were confirmed by DNA sequencing (Macrogen). The expressed protein has a hexa-histidine affinity tag followed by a thrombin cleavage site such that the N-terminus of the expressed protein has the following sequence preceding the 26th residue of the *B. subtilis* sppA gene (MGSSHHHHHHSSGLVPRGSH). The expressed protein (including 6xHis tag and

linker/thrombin site) has a calculated molecular mass of 36,182 Da and a theoretical isoelectric point (pI) of 6.9.

The full length $SppA_{BS}$ and $SppA_{BS}^{\Delta 2\text{-}54}$ construct was cloned into the expression vector pET28a+ (Novagen) using the NdeI and XhoI restriction sites. The oligonucleotides used to amplify the gene are; forward primer 5'-catatgaatgcaaaaagatggattgca-3' for full length SppA and forward primer 5'-catatgagtccctcaagtaaaattgccg-3' for $SppA_{BS}^{\Delta 2\text{-}54}$. The reverse primer was 5'-ctcgagctacttcgcatagagatacatcattct-3' for both constructs. The construct sequence was confirmed by DNA sequencing. The expressed protein has a hexa-histidine affinity tag followed by a thrombin cleavage site such that the amino-terminus of the expressed protein has the following sequence preceding the 1^{st} or 55^{th} residue of the *Bacillus subtilis* sppA gene (MGSSHHHHHHSSGLVPRGSH).

2.2.2. Expression and purification

E. coli Tuner (DE3) competent cells were transformed with the pET28b+/$SppA_{BS}^{\Delta 1\text{-}25}$K199A plasmid described above. The cells were grown in LB media containing kanamycin (0.05 mg/ml) at 37^{o}C to an OD600 of 0.6 and induced with 0.5 mM IPTG at 25^{o}C for 1 hour. The cells were harvested by centrifugation (6,000 g for 6 minutes) and then resuspended in buffer A (20 mM Tris-HCl pH 8.0 and 150 mM NaCl). The cells were lysed by sonication with three 15 second pulses at 30% amplitude with a 30 second rest between each of the pulses (Fisher Scientific Sonic Dismembrator Model 500). Homogenization was then carried out at 20,000-25,000 psi for 3 minutes using an Avestin Emulsiflex-3C cell homogenizer. The cell lysate was centrifuged at 30,000 g for 35 minutes and the supernatant was applied to a buffer A pre-equilibrated Ni^{2+}-NTA affinity chromatography column. The column was washed with 50 mM and 75 mM imidazole in buffer A. The $SppA_{BS}^{\Delta 1\text{-}25}$K199A was eluted using a step-wise gradient, 100 mM – 600 mM imidazole in buffer A (Figure 2.1.A). Fractions containing $SppA_{BS}^{\Delta 1\text{-}25}$K199A were pooled and dialyzed against buffer A overnight at 4^{o}C to remove imidazole. Dialyzed protein was then concentrated to 10 mg/ml using an Amicon ultra-centrifugal filter device (Millipore) with a 10 kDa cut off. The protein concentration was measured using a NanoDrop ND-100 spectrophotometer. The extinction coefficient (17880 $M^{-1}cm^{-1}$)

was calculated based on the $SppA_{BS}^{\Delta 1-25}$K199A amino acid sequence using ProtParam (Gasteiger et al., 2005).

2.2.3. Limited proteolysis

$SppA_{BS}^{\Delta 1-25}$K199A (10 mg/ml) was digested with thermolysin (Sigma) (500:1 molar ratio) overnight at room temperature. The sample was then applied to a Superdex 200 size-exclusion chromatography column, equilibrated with buffer A, on an Amersham ÄKTA FPLC system running at a flow rate of 0.5 ml/min. The fractions containing the stable proteolytic fragment of $SppA_{BS}^{\Delta 1-25}$K199A were pooled and concentrated to ~18 mg/ml.

2.2.4. Amino-terminal sequencing analysis

To determine the N-terminus created by thermolysin digestion, crystals of the thermolysin digested $SppA_{BS}^{\Delta 1-25}$K199A were dissolved in buffer A, run on 13.5% SDS-PAGE and blotted electrophoretically onto a PDVF membrane (Millipore). The membrane was stained with Coomassie Brilliant Blue R-250 and sent to the Iowa State University Protein Facility. Automated Edman sequencing with a model 494 Procise Protein Sequencer/140C Analyzer (Applied Biosystems, Inc.) was used to determine the sequence for the first six amino acid residues.

2.2.5. Analytical size exclusion chromatography (SEC) and multi-angle light scattering (MALS) analysis

To determine the oligomeric state of $SppA_{BS}$ in solution, thermolysin treated $SppA_{BS}^{\Delta 1-25}$K199A (10 mg/ml) was applied to a Superdex200 (GE Healthcare) size-exclusion column equilibrated with buffer A and run at a flow rate of 0.4 ml/min. The column was connected in line to a Dawn 18-angle light-scattering detector coupled to an Optilab rEX interferometric refractometer and a Quasi-Elastic Light-Scattering (QELS) instrument (Wyatt Technologies). The molecular mass was calculated with ASTRA v5.1 software (Wyatt Technologies, Inc.) using the Zimm fit method (Zimm, 1948) with a refractive index increment, dn/dc = 0.185 mL/g.

2.2.6. Crystallization

The thermolysin digested $SppA_{BS}^{\Delta 1-25}$K199A crystals were grown using the sitting-drop vapor diffusion method. The detergent n-dodecyl-β-maltoside (DDM, 0.1% final concentration) was added to the protein before setting up crystallization drops. The drop contained 1 μl of protein and 1 μl of reservoir solution. The refined reservoir condition was: 23 % tert-butanol, 0.1 M Tris-HCl pH 8.5 and 5 % 2-methyl-2,4-pentanediol (MPD). The drop was equilibrated against 1 ml of reservoir solution at 18°C. The cryo solution contained 20 % MPD, 23 % tert-butanol and 0.1 M Tris-HCl pH 8.5. The crystal was transferred to the cryo solution and flash-cooled in liquid nitrogen.

2.2.7. Data collection

Diffraction images were collected on beamline 08B1-1 at the Canadian Macromolecular Crystallography Facility (CMCF) of the Canadian Light Source (CLS), using a Rayonix MX300HE x-ray detector. A total of 360 images were collected at wavelength 0.9795 Å with 0.5° oscillations and each image was exposed for 0.5 seconds. The crystal to detector distance was 250 mm. *HKL2000* was used to process the diffraction images (Otwinowski et al., 1993). The crystal belongs to space group $P2_12_12_1$ with unit cell dimensions of 87.8 x 131.1 x 207.3 Å. There are 8 molecules in the asymmetric unit with a Matthews coefficient of 2.67 $Å^3$ Da^{-1} (54.0% solvent) which was calculated with the Matthews Probabilities calculator (Kantardjieff and Rupp, 2003) using the $SppA_{BS}$ molecular mass after limited proteolysis (28,000 Da). See Table 2.1 for crystal parameters and data collection statistics.

2.2.8. Structure determination and refinement

Phase estimates were obtained by molecular replacement using the program *Phaser* (McCoy et al., 2005). A search model was built using the C-terminal domain of $SppA_{EC}$ (pdb: 3BFO, chain A) as a template. The homology model was built using the program *Chainsaw* (Stein, 2008). The side chains of conserved residues were included and non-conserved side chains were truncated to the β-carbon atom. The program *Autobuild* within *PHENIX* version 1.6.4 (Adams et al., 2010) was used to build the side chains and refine the initial structure. Restrained refinement was performed using the program *REFMAC5* (Murshudov et al., 2011; Winn et al., 2003) and manual adjustments

and manipulations were executed with the program *Coot* (Emsley and Cowtan, 2004). The final round of refinement included TLS and restrained refinement with 6 TLS groups for each chain. Input files were obtained by the TLS motion determination server (Painter and Merritt, 2006; Winn et al., 2003). See Table 2.1 for refinement statistics.

2.2.9. Structure analysis

The figures were created using *PyMOL* (DeLano, 2002). *ClustalW* (Thompson et al., 1994) and *ESPript* v.2.2 (Gouet et al., 1999) were utilized for the sequence alignment. Signal peptide sequences of *B. subtilis* and *E. coli* were obtained from Signal Peptide Resource (Choo et al., 2005). The protein-protein interface between protomers was analyzed using *PISA* (Krissinel and Henrick, 2007) and *PROTORP* (Reynolds et al., 2009). The conserved residues among Gram-negative and Gram-positive SppA were mapped onto the structure using *ConSurf* (Ashkenazy et al., 2010; Glaser et al., 2003; Landau et al., 2005). The substrate specificity pockets were analyzed using *CASTp* (Liang et al., 1998) while the stereochemistry of the structure was analyzed with the program *PROCHECK* (Laskowski, 2001). *PROMOTIF* (Hutchinson and Thornton, 1996) was used to identify and analyze the secondary structure and motifs within in the protein.

Table 2.1. Data collection and refinement statistics

Crystal Parameters	
Space group	**$P2_1\,2_1\,2_1$**
a,b,c (Å)	**87.8, 131.1, 207.3**
Data Collection Statistics	
Wavelength (Å)	**0.9795**
Resolution (Å)	**50.0 – 2.4 (2.5 – 2.4)[a]**
Total Reflections	**709425**
Unique reflections	**96535 (9520)**
R_{merge} [b]	**0.085 (0.294)**
Mean (I)/σ (I)	**52.7 (7.8)**
Completeness (%)	**99.8 (99.8)**
Redundancy	**7.4 (6.9)**
Refinement Statistics	
Protein molecules (chains) in A.U.	**8**
Residues	**1803**
Water molecules	**257**
Total number of atoms	**13957**
R_{cryst} [c] / R_{free} [d] (%)	**20.6/ 24.1**
Average B-factor (Å^2) (all atoms)	**49.0**
r.m.s.d. on angles (°)	**1.135**
r.m.s.d. on bonds (Å)	**0.010**

[a] The data collection statistics in brackets are the values for the highest resolution shell.

[b] $R_{merge} = \sum_{hkl}\sum_i|I_i(hkl) - \langle I(hkl)\rangle| \,/\, \sum_{hkl}\sum_i I_i\,(hkl)$, where $I_i(hkl)$ is the intensity of an individual reflection and $\langle I(hkl)\rangle$ is the mean intensity of that reflection.

[c] $R_{cryst} = \sum_{hkl}\big||F_{obs}| - |F_{calc}|\big| \,/ \sum_{hkl}|F_{obs}|$, where F_{obs} and F_{calc} are the observed and calculated structure-factor amplitudes, respectively.

[d] R_{free} is calculated using 5% of the reflections randomly excluded from refinement.

2.2.10. SppA activity assay via fluorogenic peptide substrates

The reactions were carried out in buffer A at 37°C using a Cary Eclipse fluorescence spectrophotometer (Varian). A series of peptide-MCA (4-*M*ethyl 7-*C*umaryl *A*mide) fluorogenic substrates were prepared at 10 mM stock solution by dissolving them in 100% dimethyl sulfoxide (DMSO). The substrates: Z-AAF-MCA, Suc-AAA-MCA and Z-LLL-MCA were purchased from The Peptide Institute. The substrates: Suc-LLVY-MCA and Z-LLE-MCA were purchased from Calbiochem. The substrate Z-GGL-MCA was purchased from Bachem. The substrate Boc-LRR-MCA was purchased from Peptides International. The abbreviations for the N-terminal modifications are: Z- is Benzyloxycarbonyl; Suc- is Succinyl; Boc- is tert-Butyloxycarbonyl. For comparison of the relative activity between $SppA_{BS}^{\Delta 2\text{-}54}$ and $SppA_{EC}^{\Delta 2\text{-}46}$ a single substrate concentration was utilized (10 μM for $SppA_{BS}^{\Delta 2\text{-}54}$, 2 μM for $SppA_{EC}^{\Delta 2\text{-}46}$) in each 500 μl reaction containing 40 nM of enzyme in a 1 cm x 1 cm cuvette. The final DMSO concentration was 2%. The excitation and emission wavelengths used were 380 nm and 460 nm, respectively.

2.2.11. SppA activity assay via protein digestion

$SppA_{EC}^{\Delta 2\text{-}46}$ (3 μg) or $SppA_{BS}^{\Delta 2\text{-}54}$ (4 μg) was added to 100 μg of substrate either *E. coli* BamB or *E. coli* BamD and then incubated for 0, 1, 2, 3 and 17 hours at 37 ºC, in buffer A (20 mM TrisHCl pH8 and 150 mM NaCl). The *E. coli* periplamic lipoprotein BamB (UniProt accession number: P77774) was over-expressed and purified as previously described (Kim and Paetzel, 2011). The *E. coli* periplamic lipoprotein BamD (UniProt accession number: P0AC02) was over-expressed and purified as previously described (Kim et al., 2011). The total reaction volume was 60 μl and at the end of each incubation period 10 μl of the reaction solution was stopped by addition of an equal volume of 2X SDS-PAGE sample buffer and then heated for 8 min at 100ºC. In the case of the reaction with $SppA_{EC}^{\Delta 2\text{-}46}$K209A, or $SppA_{BS}^{\Delta 2\text{-}54}$K199A, 4 μg of these mutant enzymes were used. The quenched reactions were run on a 13.5 % SDS-PAGE gel and stained with PageBlue stain. A negative control was run for each gel were the substrate was incubated for 17 hours at 37ºC, in buffer A without SppA. $SppA_{EC}$ was prepared as previously described (Kim et al., 2008).

2.3. Results

2.3.1. $SppA_{BS}$ purification, crystallization and structure solution

We have observed that expressing the soluble domain of $SppA_{BS}$ ($SppA_{BS}\Delta^{1-25}$) in the cytoplasm of *E. coli* results in slow cell growth and a limited protein expression level. Sequence alignments to $SppA_{EC}$ suggest that Ser147 serves as the nucleophile and Lys199 served as the general base in the $SppA_{BS}$ catalytic mechanism, therefore we designed the mutants K199A and S147A to test this hypothesis and to observe their effect on *B. subtilis* SppA expression level. Consistent with their proposed role as active site residues, our activity assays revealed that these mutant enzymes are inactive. In addition, the K199A mutant was highly over-expressed in the cytoplasm of *E. coli* and was easily purified in milligram quantities (Figure 2.1). This protein was subjected to limited proteolysis using thermolysin in order to improve crystallization. The resulting proteolytically resistant fragment has a molecular mass of approximately 28 kDa, approximately 8 kDa smaller than the originally purified protein. N-terminal sequencing analysis has revealed that the proteolytically resistant fragment starts at Leu51 (Figure 2.2.A). The alcohols tert-butanol and 2-methyl-2,4-pentanediol and the detergent n-dodecyl-β-maltoside were used to produce conditions that led to highly ordered plate shaped crystals (Figure 2.1.B). The stable proteolytic fragment of $SppA_{BS}{}^{\Delta 1-25}$K199A, which resulted in diffraction quality crystal followed by data collection and structure solution, will be called $SppA_{BS}$ from now on. The 2.4 Å resolution structure reveals clear electron density for residues 57 to 295, with the exception of a loop region between residues 72 to 82.

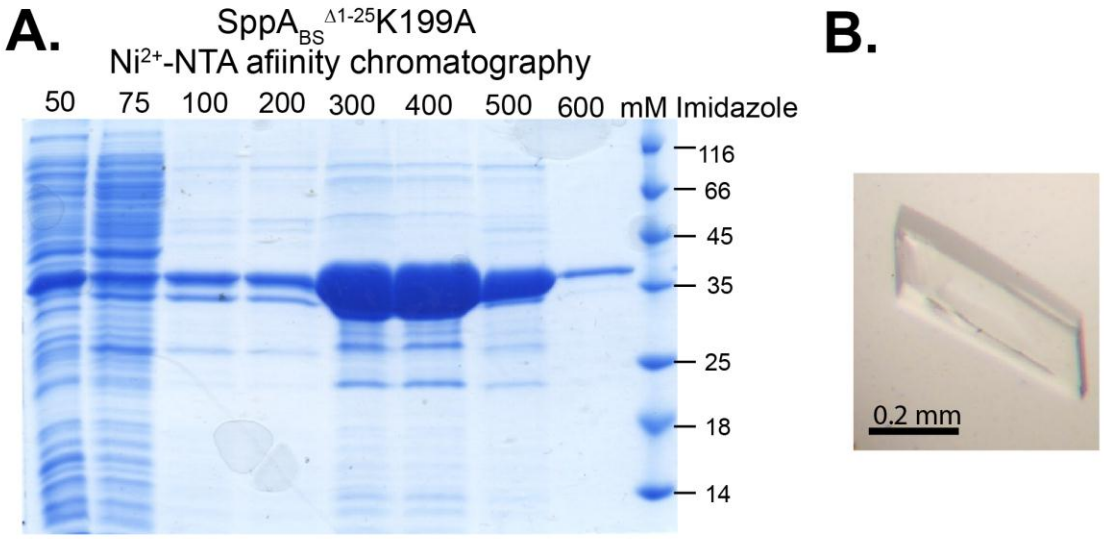


Figure 2.1. Ni^{2+}-NTA affinity chromatography of $SppA_{BS}^{\Delta 1-25}K199A$ and picture of $SppA_{BS}$ crystal.

A. Elution fractions of $SppA_{BS}^{\Delta 1-25}K199A$ from Ni^{2+}-NTA affinity chromatography are ran on 13.5 % SDS-PAGE gel. The imidazole concentrations used to elute the sample are shown on top of each lane. Molecular marker is shown on far right. **B.** Picture of $SppA_{BS}$ crystal. Themolysin resistance fragment of $SppA_{BS}^{\Delta 1-25}K199A$ which resulted X-ray diffraction data at 2.4 Å resolution

2.3.2. The $SppA_{BS}$ protein fold

The protomeric unit of the $SppA_{BS}$ catalytic domain (residues: 57 to 295) has an α/β protein fold consisting of seven β-strands, eight α-helices and one 3_{10} helix (Figure 2.2). The $SppA_{BS}$ protomer has two regions; a globular region and an extension region. The globular region includes β-strands 1-4 and 7 which are arranged in a parallel β-sheet, surrounded by α-helices 1-3 and 6-8. The extension region consists of α-helices 4 and 5 and β-strands 5 and 6. Each protomer contains a nucleophilic residue: Ser147, and a general base residue: Lys199 (Figure 2.2). Ser147 is located at the turn before α-helix 3 of the globular region while Lys199 is located on the loop between β-strand 6 and α-helix 5 of the extension region. The distance from the Ser147 Oγ to the Lys199 Cβ is approximately 29 Å, suggesting that a monomeric $SppA_{BS}$ would not be capable of catalysis. Eight molecules are in the asymmetric unit of this crystal. Analysis of the asymmetric unit reveals that the active site catalytic dyad is created by Ser147 from one protomer and Lys199 from the neighboring protomer (Figure 2.3.B). Eight $SppA_{BS}$ protomers come together to form a dome shaped structure with eight separate active sites (Figure 2.3.A).

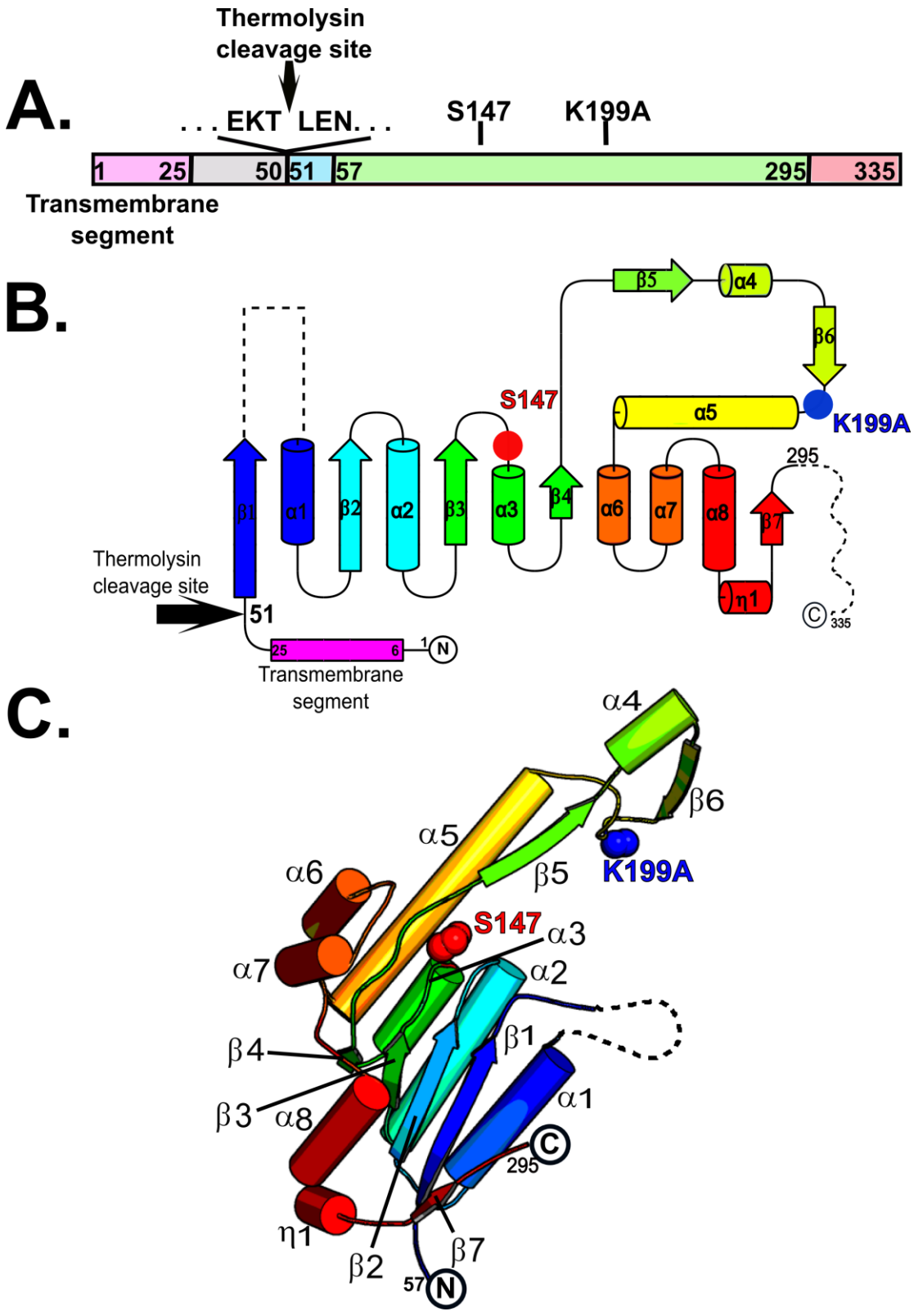


Figure 2.2. (Figure legend on next page)

Figure 2.2. The $SppA_{BS}$ protein folds.

A. A schematic diagram showing the full length $SppA_{BS}$ with its predicted transmembrane segment (left pink) and the confirmed N-terminal thermolysin cleavage site. The light green region is what is observed in the electron density. **B.** A topology diagram of $SppA_{BS}$ showing the full length protein with β-strand as arrows and α-helices as cylinders. The protein is colored as a gradient from N-terminus (blue) to C-terminus (red). The regions not seen in the electron density (residues 72-82) are shown as dashed lines. **C.** A cartoon diagram showing the tertiary structure of the $SppA_{BS}$ protomer. The β-strands are shown as arrows and the α-helices as cylinders. The color scheme is the same as in B. The nucleophile Ser147 (red) and general base K199A (blue) are shown as spheres.

2.3.3. $SppA_{BS}$ is octameric in solution

To confirm the existence of the octameric state of $SppA_{BS}$ in solution, we analyzed its size by analytical size exclusion chromatography (SEC) and multi-angle light scattering (MALS). We observed that $SppA_{BS}$ has an average molecular mass of 225,400 ± 4,508 g/mol, consistent with the $SppA_{BS}$ complex (proteolytically resistant fragment) being in an octameric arrangement in solution (Figure 2.3.C).

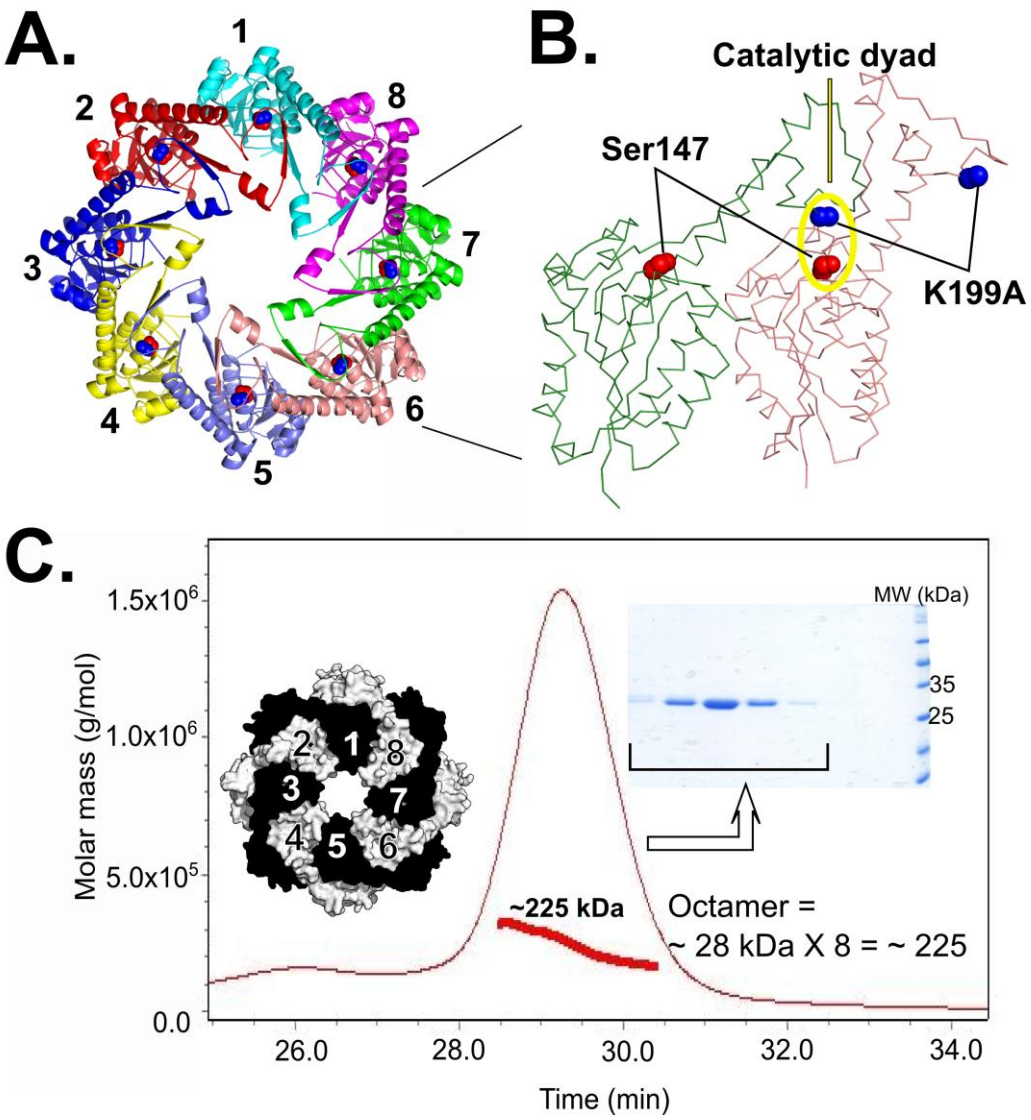


Figure 2.3. Octameric arrangement of $SppA_{BS}$.

A. The $SppA_{BS}$ octamer as shown in a cartoon diagram, viewed from the top (closest to the extension region that forms the smaller opening). Each protomer is represented in a different color. The serine nucleophile (red) and lysine general base mutated to alanine (blue) are shown as spheres, revealing eight active site catalytic dyads. **B.** A Cα trace of two neighboring protomers coming together to form one of the eight active sites; the serine (red sphere) comes from one protomer and the lysine (blue sphere) arrives from the adjacent protomer. **C.** A multi-angle light-scattering (MALS) analysis of purified $SppA_{BS}$ proteolytic resistant fragment is represented by a molar mass (g/mol) versus time (min) plot overlaid with an analytical size-exclusion elution profile. A molecular mass of 225,000 Da is approximately equivalent to eight $SppA_{BS}$ protomers. The SDS-PAGE gel stained with PageBlue shows the fractions collected from the SEC column. The molecular mass markers are shown on the right. Next to the peak is a top view of the $SppA_{BS}$ octamer as shown in molecular surface representation, each protomer is colored alternating black and white.

2.3.4. The $SppA_{BS}$ dimensions and surface properties

The octameric $SppA_{BS}$ catalytic domain is dome-shaped with a wide opening created by the globular regions and a narrower opening made by the extension regions (Figure 2.4). Based on membrane topology predictions and analogy to $SppA_{EC}$, whose membrane topology has been experimentally determined, the wider opening of $SppA_{BS}$ is predicted to face the outer leaflet of the cytoplasmic membrane (Wang et al., 2008) (Figure 2.4.A). The outside of the dome has a relatively polar surface made up of both positively and negatively charged patches. The interior of the dome is predominantly hydrophobic with some patches of negative charge. The positively charged rim at the narrower opening of the dome is created by eight lysines, Lys185 from each of the $SppA_{BS}$ protomers. The narrow opening is ~22 Å in diameter. A cross section of the dome shows a concave groove where the active sites and substrate specificity binding pockets reside (Figure 2.4.E). At this point, the diameter is approximately ~73 Å, while the narrowest region of the $SppA_{BS}$ dome interior is approximately 50 Å in diameter. The height of the dome is approximately 50 Å. The base of the dome, 86 Å in diameter at the opening, is primarily positively charged. The inner cavity has a solvent accessible surface area of 17,220 $Å^2$ and a volume of 56,456 $Å^3$.

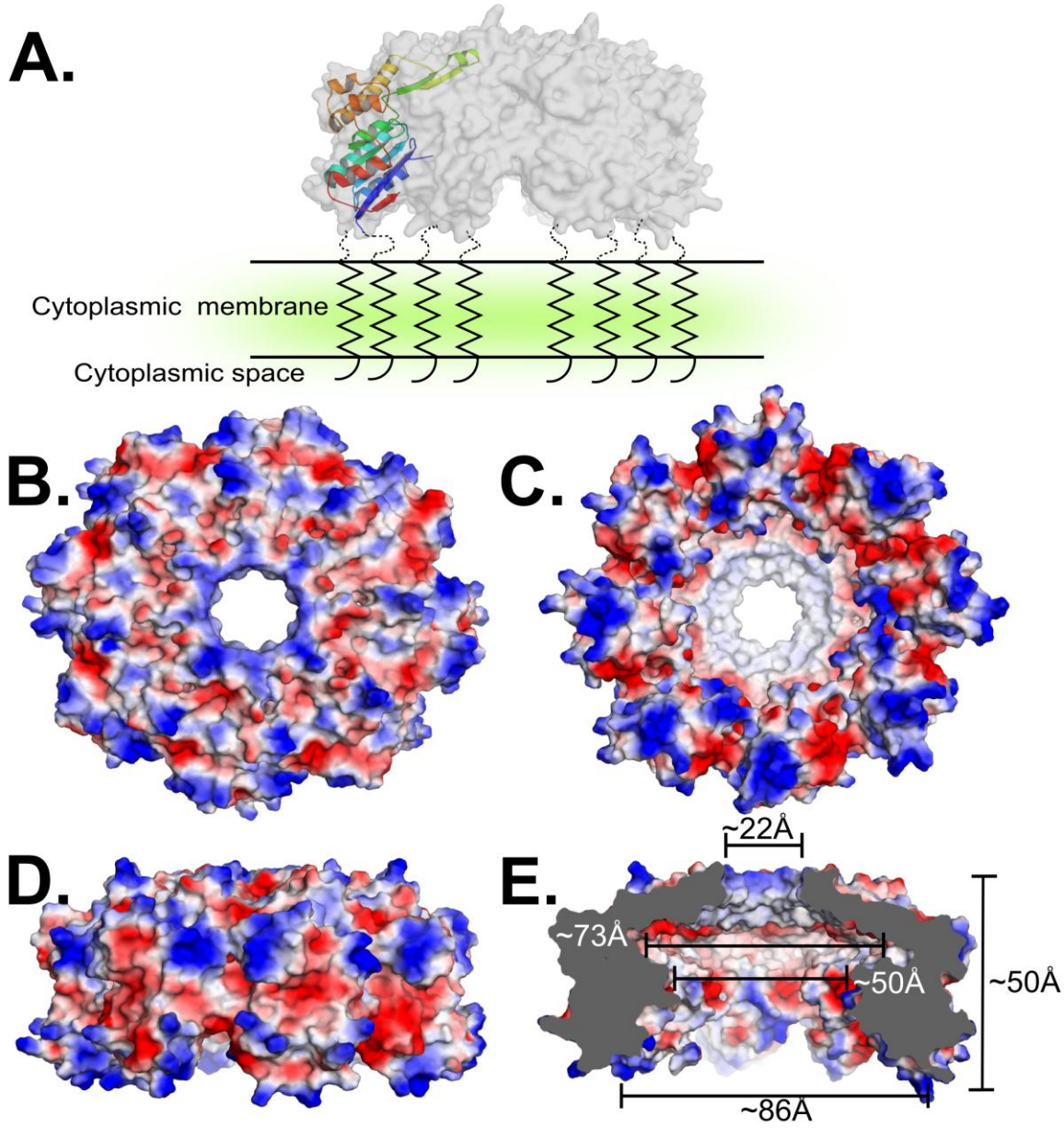


Figure 2.4. Surface properties and dimensions of $SppA_{BS}$.

A. A semitransparent surface representation of the $SppA_{BS}$ octamer, with one protomer shown in cartoon diagram. The widest opening in $SppA_{BS}$, formed from the globular region ("bottom"), is facing the cytoplasmic membrane and the transmembrane segment is schematically drawn. The electrostatic properties of $SppA_{BS}$ are shown from the top **B.**, bottom **C.** and side **D.** (blue; positive, red; negative). **E.** A cross-section view of the $SppA_{BS}$ octamer, with dimensions given.

2.3.5. Protein-protein interactions involved in the $SppA_{BS}$ octameric assembly

The two long edges of each monomer contribute to the eight equivalent protein-protein interfaces within the $SppA_{BS}$ octameric complex. Each interface buries, on average, 1,384 $Å^2$ of solvent accessible surface area on one side of the protomer, and 1,514 $Å^2$ of solvent accessible surface area on the other side. There are, on average, 22 hydrogen bonds and 7 salt bridges between the protein chains at each interface. The residues involved at the interface are 35 % charged, 28 % polar, and 37 % non-polar. One can describe the interface between monomers by looking at three separate regions of the protein; the extension region, the globular region, and the middle region (Figure 2.5). The protein-protein interface in the extension region is a two stranded anti-parallel β-sheet created by β-strand 5 from one protomer and β-strand 6 from the neighboring protomer (Figure 2.5.B). The interface between the two globular regions is created by residues from α-helix 1 and 2 in one monomer interacting with residues on β-strand 1, 2, 3 and α-helix 7 of its neighboring molecule (Figure 2.5.C). The middle of the interface consists primarily of interactions between residues in α-helix 5 from one protomer and the loop between α-helix 6 and 7 of the neighboring molecule (Figure 2.5.D).

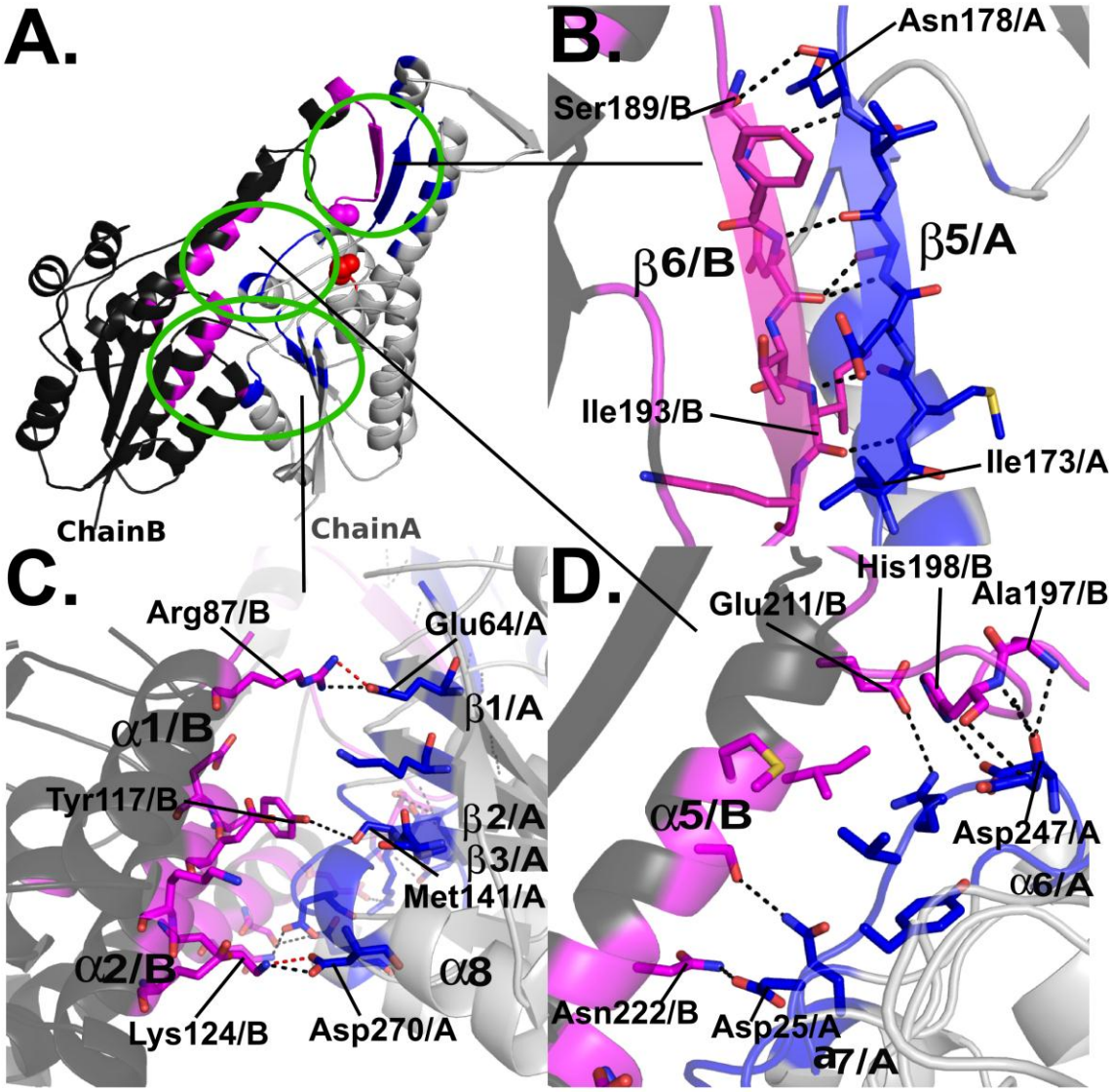


Figure 2.5. Protein-protein interactions at the interface between protomers within the octameric $SppA_{BS}$.

A. Residues involved in interactions between two molecules of $SppA_{BS}$ (black and grey) are highlighted in magenta and blue respectively. Close-up views of the top, middle and bottom regions of the interface are depicted in panels **B., C.** and **D.** Hydrogen bonds are depicted by dotted lines. Residues are labeled: residue, residue number / Chain ID.

2.3.6. The $SppA_{BS}$ Ser/Lys catalytic dyad and other active site residues

The catalytic dyad of $SppA_{BS}$ is made up of the nucleophilic Ser147 Oγ from one protomer and the general base Lys199 Nζ from the adjacent protomer. Having eight

protomers in the octamer therefore creates eight separate active sites, each formed at the protein-protein interface (Figure 2.6). In order for $SppA_{BS}$ to have one completely functional unit, three neighboring $SppA_{BS}$ protomers are required to assemble. The nucleophile (Ser147), oxyanion hole residues (Gly114 and Gly148), the general base coordinating residue (Ser169) and S1 specificity pocket residues arrives from the central molecule (salmon in Figure 2.6), the lysine general base arrived from one neighboring protomer (green in Figure 2.6) and the S3 specificity pocket is partially formed from the other neighboring protomer (light blue in Figure 2.6).

Modeling the Lys199 side chain within this K199A mutant structure shows that the Ser147 $O\gamma$ would be expected to be within hydrogen bonding distance to the Lys199 $N\zeta$. The oxyanion hole is created by the hydrogen bond donor NH groups of Gly114 and Gly148 from the same molecule that contributes the nucleophilic Ser147 to the catalytic dyad (Figure 2.6). The N-terminal end of the α-3 helix dipole, where Ser147 is located, is also a possible contributor to oxyanion stabilization. This molecule also contributes the side chain hydroxyl group of Ser169, which in other Ser/Lys proteases similarly positioned hydroxyl groups have been proposed to help coordinate and orient the lysine general base epsilon-amino group (Feldman et al., 2006; Paetzel et al., 2002a) (Figure 2.6).

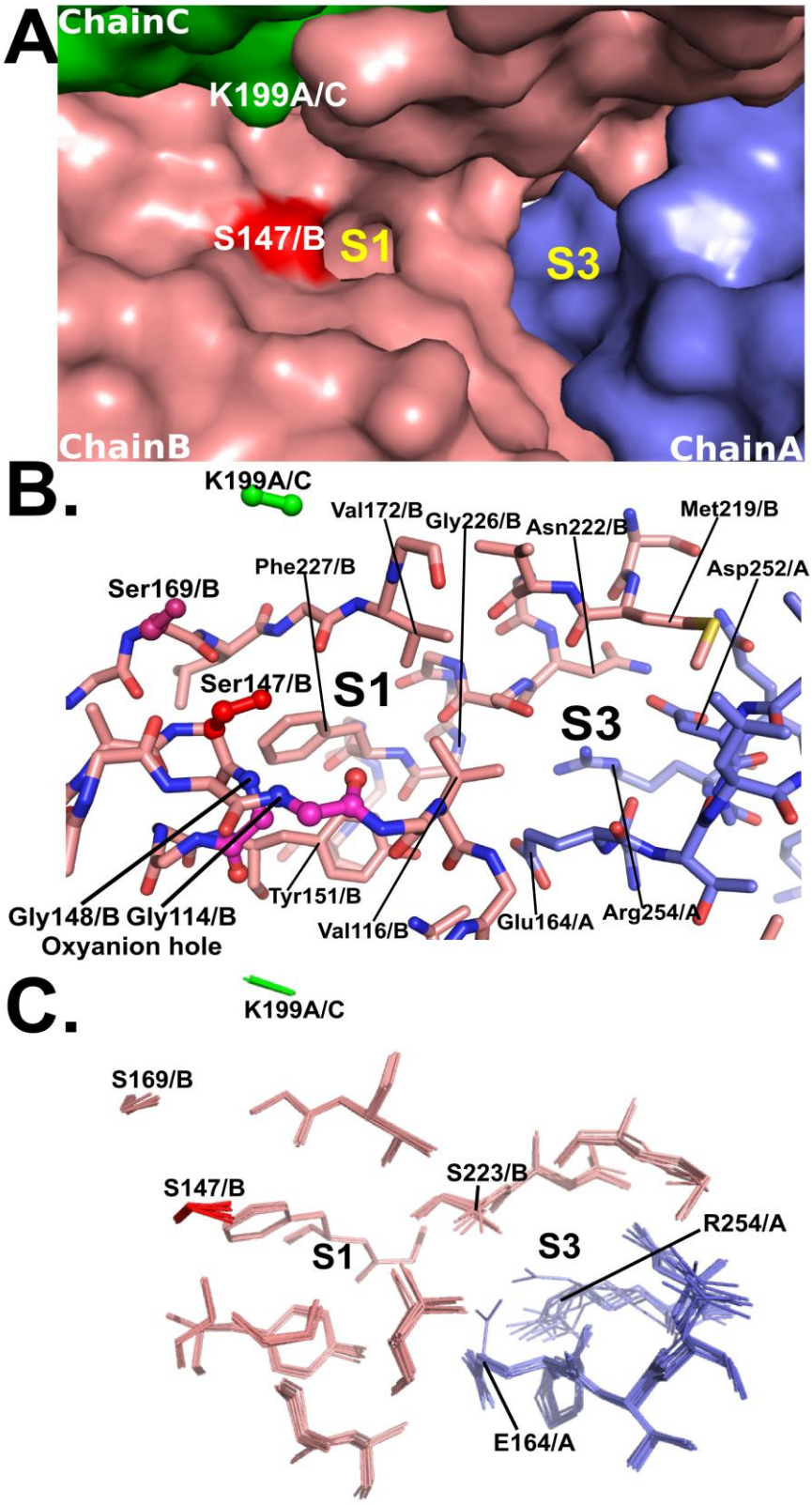


Figure 2.6. (Figure legend on next page)

Figure 2.6. $SppA_{BS}$ catalytic residues and substrate specificity pockets.

A. Three protomers of $SppA_{BS}$ create one complete catalytic core with the general base (K199A, green) arriving from the green colored protomer, the nucleophile (Ser147, red) and residues that make up the S1 substrate specificity pocket arriving from the salmon colored protomer, and the S3 pocket being formed from residues in the salmon and light blue colored protomers. **B.** Stick representation of residues involved in forming the active site and substrate specificity pockets. Ser147 is shown in red, K199A is shown in green and the proposed general base orienting residue Ser169 in maroon. Gly114 and Gly148 shown in magenta form the oxyanion hole. **C.** Superimposition of residues involved in forming each of the eight active sites and binding pockets.

2.3.7. The $SppA_{BS}$ substrate specificity binding pockets

The eight active sites and substrate binding grooves of $SppA_{BS}$ are located approximately midway up the interior of the octameric dome, creating a continuous concave surface that encircles the entire inner bowl (as seen in surface cross-section in Figure 2.4.E). Analysis of the grooves reveals only two clear specificity pockets for each of the eight active sites. Modeling an extended conformation for the substrate puts the P1 and P3 residues side chains facing the S1 and S3 specificity pockets and the P2 and P4 residue side chains facing towards the solvent, away from the binding groove. The S1 substrate specificity pocket of $SppA_{BS}$ is created by residues Gly114, Val116, Ser119, Gly148, Tyr 151, Gly171, Val172, Ser223, Gly226 and Phe227 from one protomer, and Glu164 from a neighboring protomer, while the S3 pocket is created by Val116, Met219, and Ser223 from one protomer and Pro163, Glu164, Thr165, Leu166, Asp252 and Arg254 from a third adjacent protomer (Figure 2.6.E). The S1 pocket is narrow and deep with hydrophobic walls (Phe227, Tyr151, Val116) and a more polar bottom (Glu164 $O\varepsilon1$, $O\varepsilon2$, and Tyr 151 $O\eta$) (Figure 2.13). The S3 pocket is not as deep as the S1 pocket but is significantly wider. The S3 entrance is hydrophobic, however deeper into the pocket there are polar characteristics formed by the main chain carbonyl oxygens of Pro163, Glu164, Met219 and side chains of Ser223, Asp252 and Arg254. Structural alignment of the residues involved in creating the S1 and S3 pockets in the eight binding sites of $SppA_{BS}$ showed that they superimpose quite well except for residues: Ser223, Glu164 and Arg254 (Figure 2.6.C). The Ser223 side chain rotamer, which is located at the S1 and S3 pocket boundary, varies among the eight chains. Alternate conformations are observed for the side chains of residues Glu164 and Arg254, which form part of the S3 pocket, this suggests that these residues may be dynamic and possibly be involved in an induced fit process upon substrate binding.

2.3.8. Sequence conservation between Gram-positive and Gram-negative SppA

$SppA_{BS}$ has a sequence identity of 16 % and 26 % to the N-terminal and the C-terminal domains of $SppA_{EC}$, respectively (Figure 2.7). The majority of the conserved residues shared by $SppA_{BS}$ and the C-terminal domain of $SppA_{EC}$ are randomly located on the protomer of $SppA_{BS}$, however, there is a patch of conserved residues clustered around the nucleophile Ser147 (Figure 2.8). Similar patterns of conserved residues are also observed when $SppA_{BS}$ is aligned with SppA sequences from other Gram-positive species or when $SppA_{EC}$ is aligned with SppA sequences from other Gram-negatives (Figure 2.9 & Figure 2.10).

SppAs of Gram-positive bacteria are located on the extracellular surface whereas SppAs of Gram-negative bacteria are located on the inner membrane facing the periplasmic space. Given the potentially great surface exposure of Gram-positive SppAs one might expect there to be less conservation on their exposed surfaces, as compared to the Gram-negative SppA. Mapping residue conservation onto the surface of the $SppA_{BS}$ and $SppA_{EC}$ structures, based on separate Gram-positive and Gram-negative SppA protein sequence alignments (Figure 2.9&Figure 2.10), reveals a distinctly different pattern of conservation which is consistent with the potential difference in surface exposure. There are four patches of mostly conserved residues (maroon regions) that show up on the top surface of Gram-negative SppAs, whereas a broader, more evenly distributed, circle of highly conserved residues are observed on the top surface of Gram-positive SppAs (Figure 2.11).

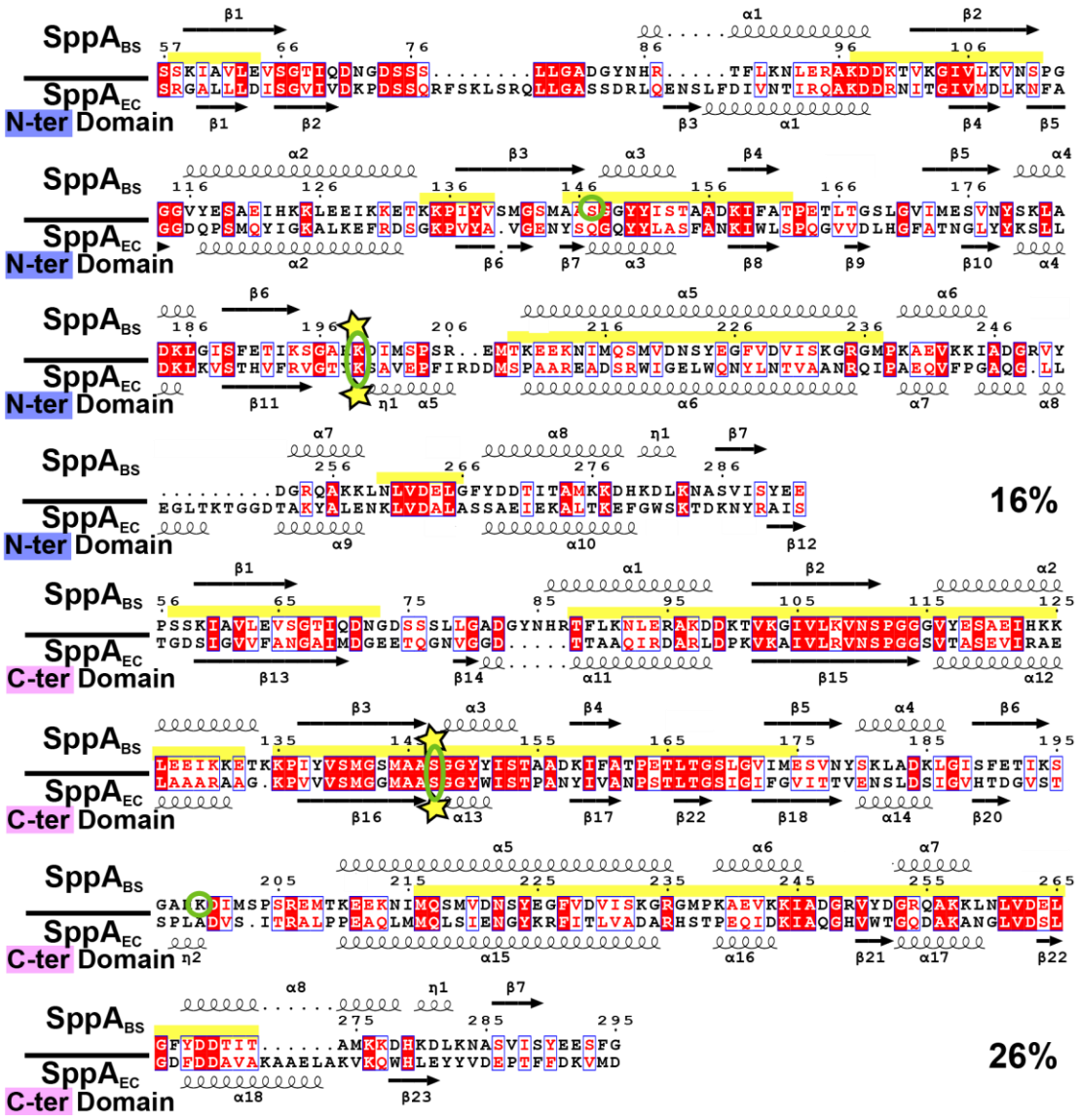


Figure 2.7. Sequence alignment of $SppA_{BS}$ to the N-terminal and C-terminal domains of $SppA_{EC}$.

Red highlights indicate identical residues. The catalytic residues are circled in green and stars indicate where the catalytic residues align. Secondary structural elements are shown ($SppA_{BS}$ above, $SppA_{EC}$ below). The residue numbers for $SppA_{BS}$ are shown above the alignment. The yellow box above the $SppA_{BS}$ sequence shows the structural alignment between $SppA_{BS}$ and $SppA_{EC}$. They have the same protein sequence register and the distance between the aligned Cα atoms of $SppA_{BS}$ and $SppA_{EC}$ ranges 0 - 2.8 Å. **A.** The sequence alignment of $SppA_{BS}$ (UniProt accession number: O34525) and the N-terminal domain of $SppA_{EC}$ (UniProt accession number: P08395) shows that the general base lysines align. **B.** The sequence alignment of $SppA_{BS}$ and the C-terminal domain of $SppA_{EC}$ shows that the nucleophilic serines align. The percent identity is shown on the right.

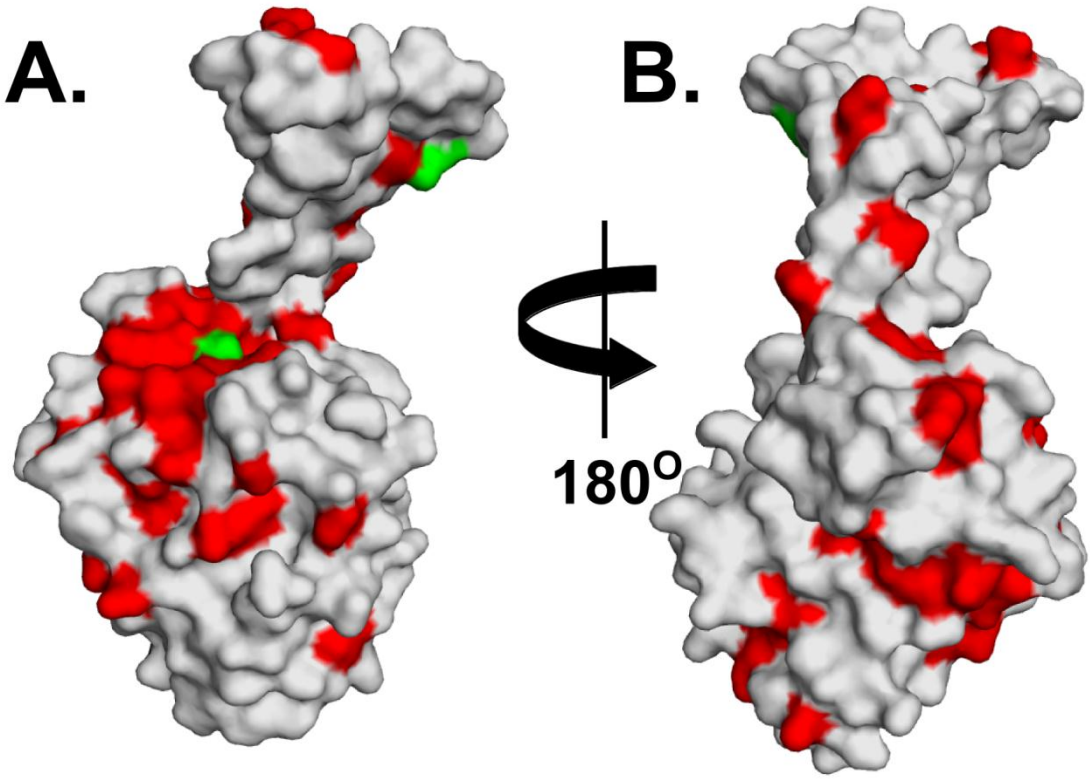


Figure 2.8. Mapping conserved surface residues between $SppA_{BS}$ and the $SppA_{EC}$ C-terminal domain onto the surface of a $SppA_{BS}$ protomer.

A. The location of conserved surface residues (red) between $SppA_{BS}$ and the $SppA_{EC}$ C-terminal domain are mapped on the $SppA_{BS}$ protomer surface, as seen from the interior of the $SppA_{BS}$ complex (left). The catalytic dyad is in green. **B.** The $SppA_{BS}$ protomer on the right shows the surface that corresponds to the exterior of the $SppA_{BS}$ complex.

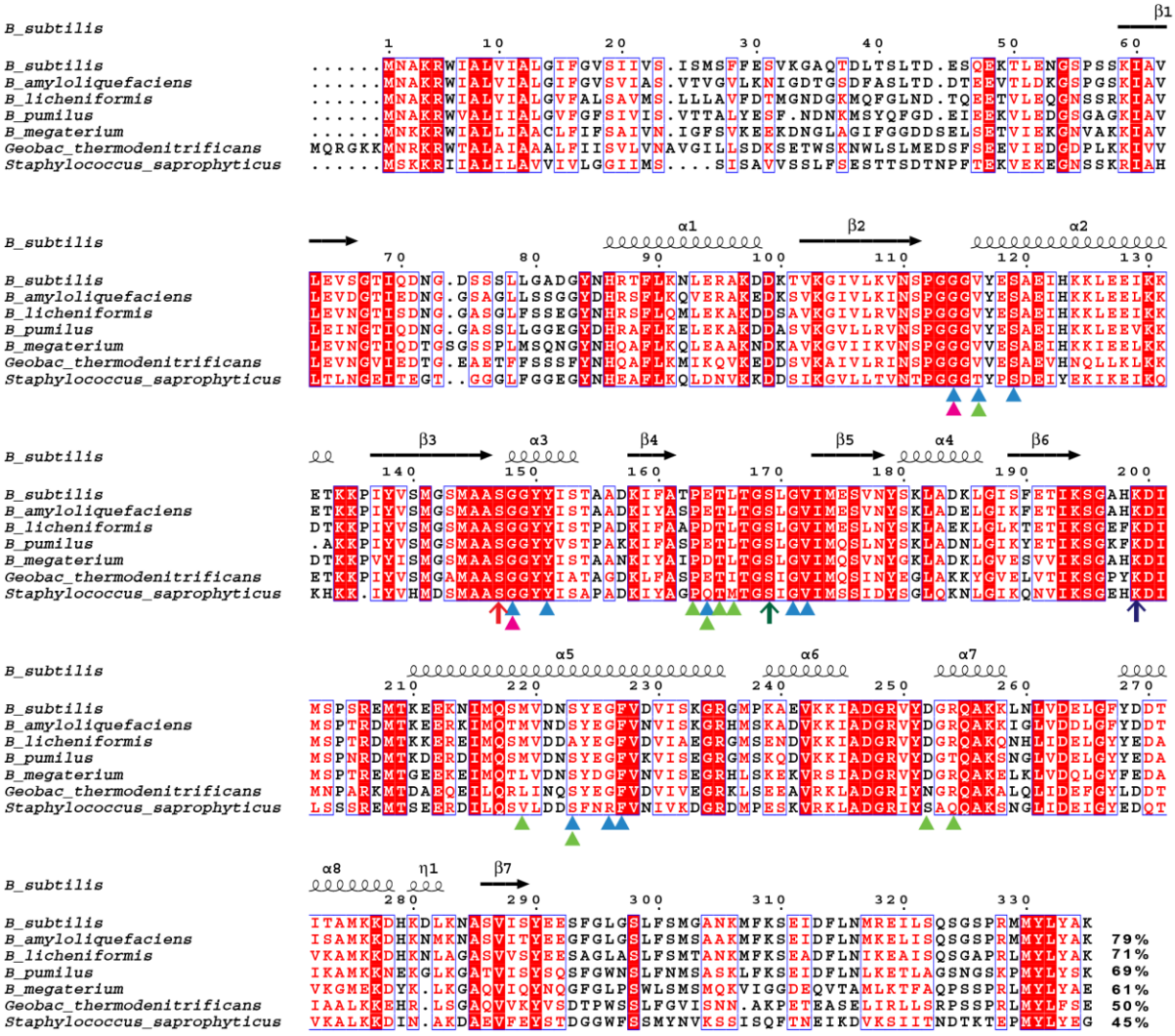
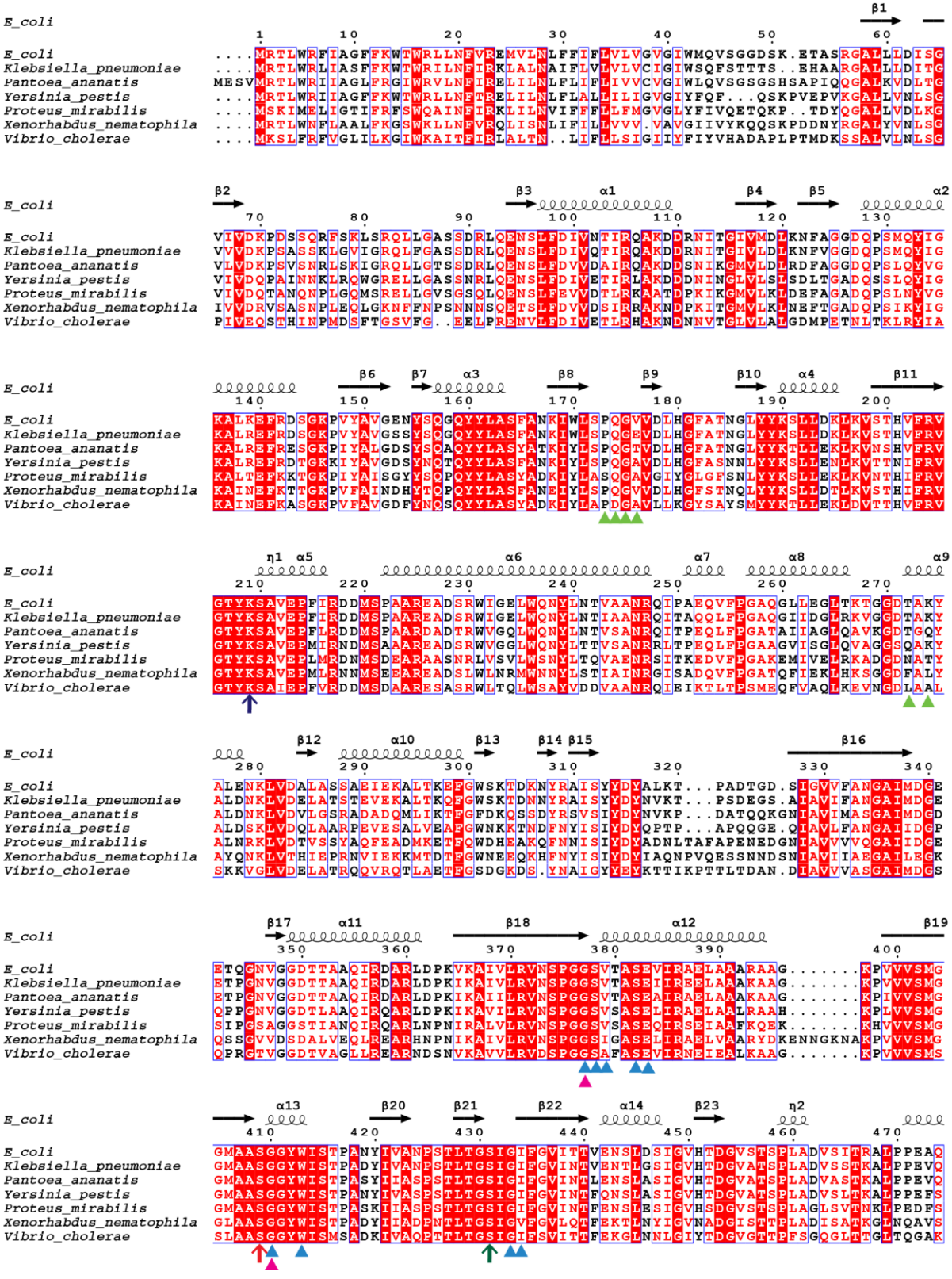


Figure 2.9. Sequence alignment of Gram-positive bacterial SppA.

Red highlights indicate identical residues. Secondary structure of $SppA_{BS}$ is shown on the top of the alignment. The residue numbers for $SppA_{BS}$ are shown above the alignment. The serine nucleophile (S147) is marked with a red arrow. The lysine general base (K199) is marked by dark blue arrow. The general base coordinating serine (S169) is marked by a dark green arrow. S1 substrate specificity pocket residues are marked by blue triangles. S3 substrate specificity pocket residues are marked by green triangles. Pink arrows mark the residues that form the oxyanion hole. The UniProt accession numbers for SppA sequences used are: *Bacillus subtilis* (O34525), *Bacillus amyloliquefaciens* (A7Z7M8), *Bacillus licheniformis* (Q65G59), *Bacillus pumilus* (A8FG77), *Bacillus megaterium* (D5DMZ1), *Geobacillus thermodenitrificans* (A4IRT5) and *Staphylococcus saprophyticus* (Q49YE6). Sequence identities relative to the *B. subtilis* sequence are shown at the end of the each alignment.


E_coli
Klebsiella_pneumoniae
Pantoea_ananatis
Yersinia_pestis
Proteus_mirabilis
Xenorhabdus_nematophila
Vibrio_cholerae

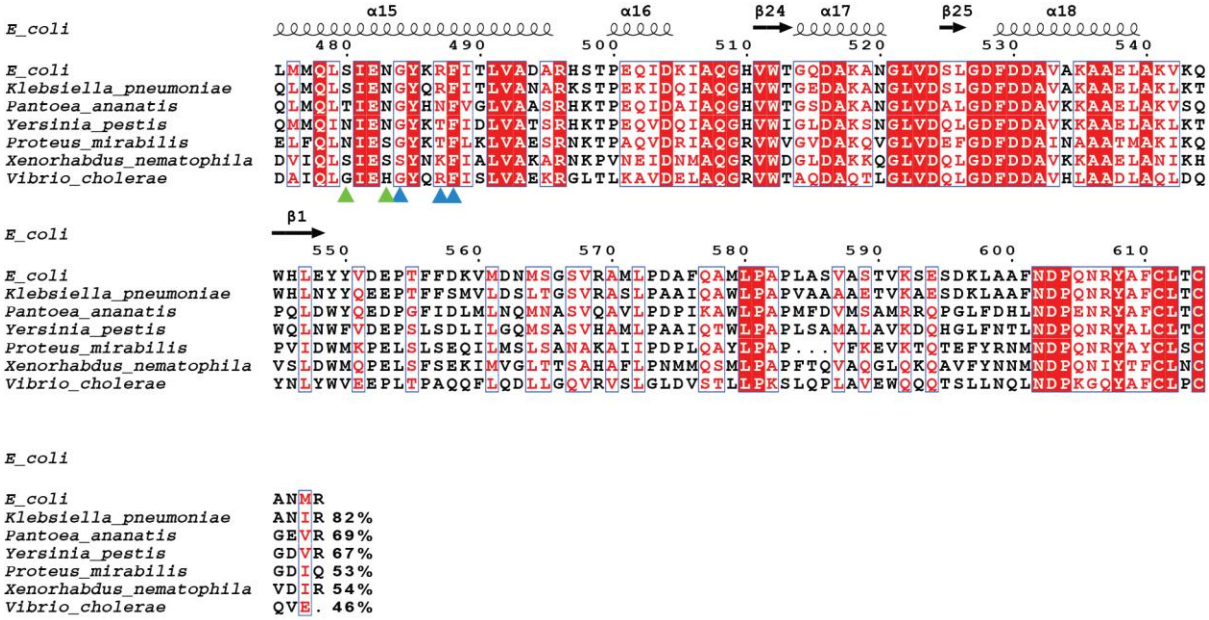


Figure 2.10. Sequence alignment of Gram-negative bacterial SppA.

Red highlights indicate identical residues.The serine nucleophile (S409) is marked with a red arrow. The lysine general base (K209) is marked by dark blue arrow. The general base coordinating serine (S431) is marked by a dark green arrow. S1 substrate specificity pocket residues are marked by blue triangles. S3 substrate specificity pocket residues are marked by green triangles. Pink arrows mark the residues that form the oxyanion hole. Secondary structure of $SppA_{EC}$ is shown on the top of the alignment. The residue numbers for $SppA_{EC}$ are shown above the alignment. The UniProt accession numbers for SppA sequences used are: *Escherichia coli* (P08395), *Klebsiella pneumonia* (B5XS66), *Pantoea ananatis* (D4GG20), *Yersinia pestis* (D1TZZ4), *Proteus mirabilis* (B4EXT2), *Xenorhabdus nematophila* (D3VHB2) and *Vibrio cholerae* (F9C891). The sequence identities relative to the *E. coli* sequence are shown at the end of the each alignment.

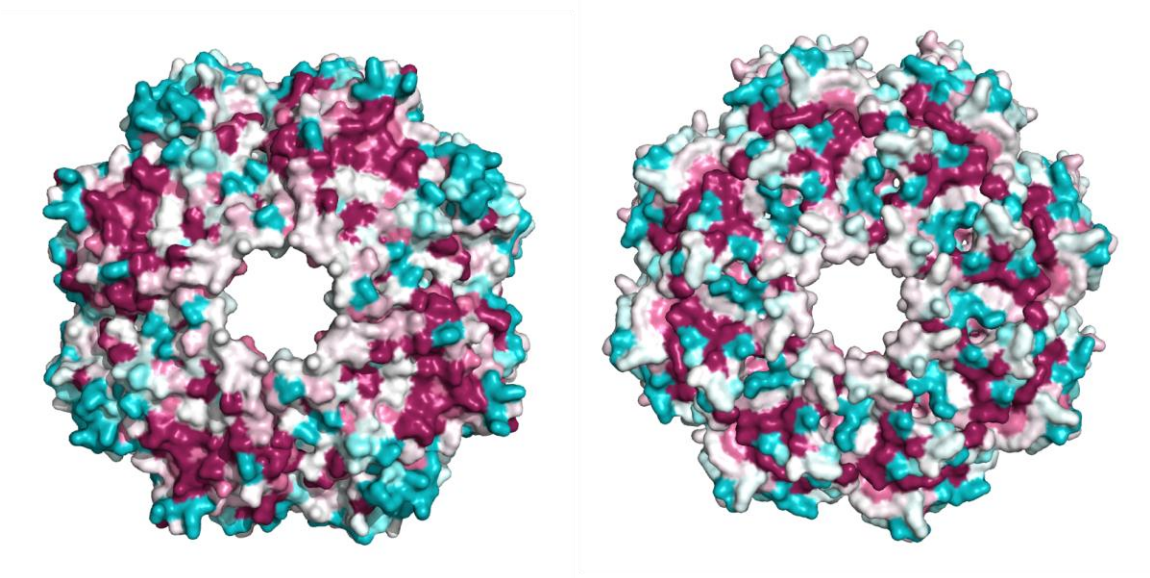


Figure 2.11. Surface residue conservation of Gram-negative SppA versus Gram-positive SppA.

Residue conservation of Gram-negative SppA and Gram-positive SppA are mapped onto the surface of the $SppA_{EC}$ (PDB:3BF0) and $SppA_{BS}$ (PDB:3RST) structures, as viewed from the top external surface. The conservation is based on separate Gram-positive and Gram-negative SppA sequence alignments (see Figure 2.9 & Figure 2.10). Completely conserved residues are shown in maroon, while highly variable residues are shown in cyan.

2.3.9. Octameric $SppA_{BS}$ versus tetrameric $SppA_{EC}$

Both the $SppA_{BS}$ and $SppA_{EC}$ form oligomers that assemble into dome-shaped structures (Kim et al., 2008). The $SppA_{BS}$ protomer is half the size of the $SppA_{EC}$ protomer. Therefore the $SppA_{BS}$ dome complex is created by eight protomers while only four protomers are required for $SppA_{EC}$ (Figure 2.12). $SppA_{EC}$'s N- and C-terminal domains are structurally tandem repeats (Figure 2.12.C). The $SppA_{BS}$ protomer superimposes onto the N-terminal domain of $SppA_{EC}$ with an r.m.s.d. value of 2.7 Å (Figure 2.12.E) while the $SppA_{BS}$ protomer superimposes onto the C-terminal domain of $SppA_{EC}$ with a much lower r.m.s.d. values of 1.1 Å (Figure 2.12.F). The most significant differences in the superimposition of the N-terminal domain of $SppA_{EC}$ with the $SppA_{BS}$ protomer are observed in the outer helices (α-helices 1 and 8) and the extension region. In addition, there is an extra helix found in the N-terminal domain of $SppA_{EC}$, between α-helix 6 and 7 of $SppA_{BS}$ (Figure 2.12.E).

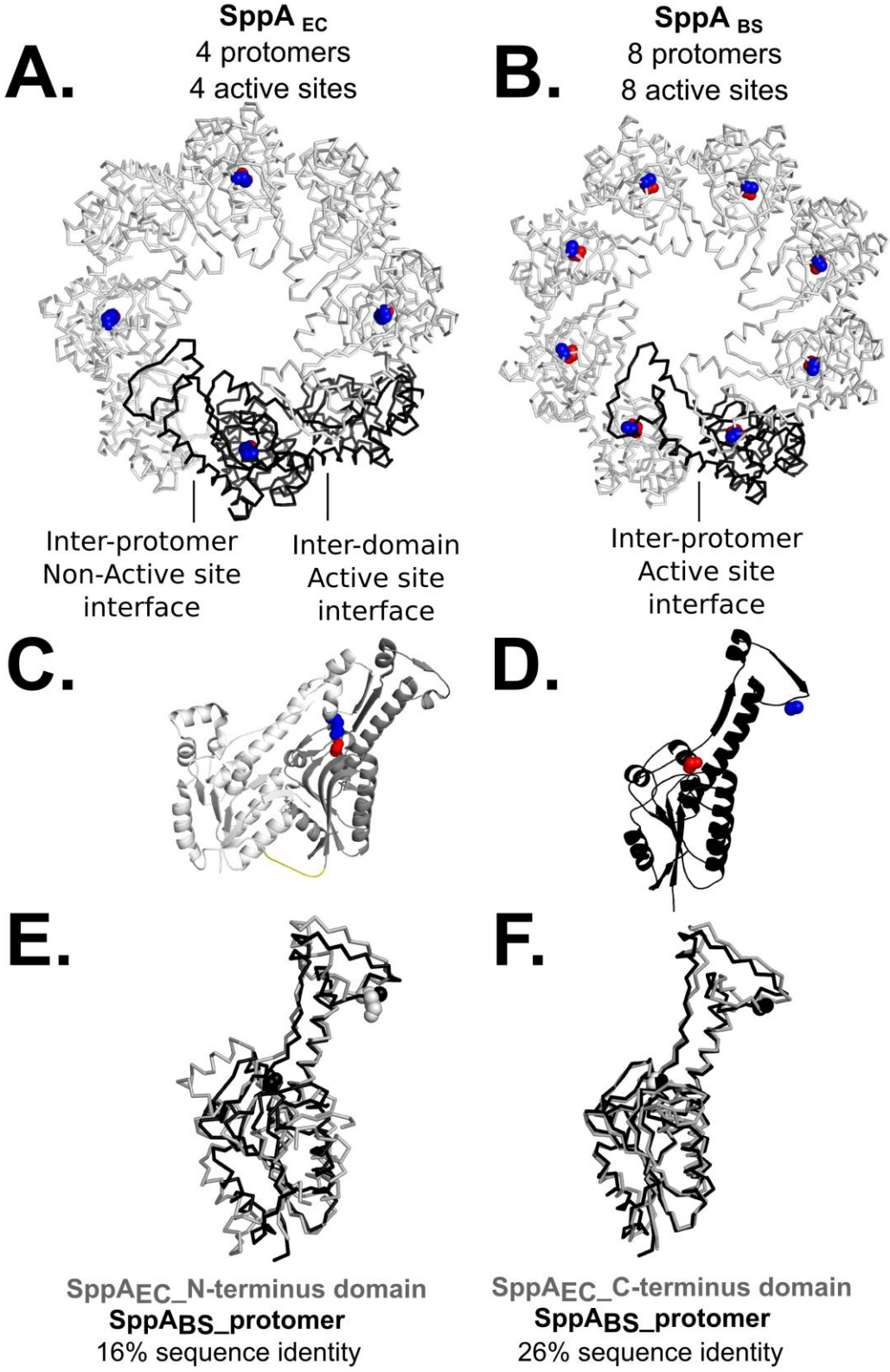


Figure 2.12. (Figure legend on next page)

Firgure 2.12. Comparison between octameric $SppA_{BS}$ and tetrameric $SppA_{EC}$.

A. The Cα trace of $SppA_{EC}$. **B.** The Cα trace of $SppA_{BS}$. The catalytic dyads are shown (general base; blue spheres, nucleophile; red spheres). One protomer in each oligomer is shown in black, in A. and B. **C.** A cartoon diagrams of a $SppA_{EC}$ protomer, the C-terminal domain is colored darker grey, the region linking the domains is colored yellow. **D.** A cartoon diagrams of a $SppA_{BS}$ protomer. **E.** A superposition of the $SppA_{BS}$ protomer (black) on the N-terminal domains of $SppA_{EC}$ (grey). **F.** A superposition of the $SppA_{BS}$ protomer (black) on the C-terminal domains of $SppA_{EC}$ (grey).

2.3.10. S1 substrate specificity binding pocket: $SppA_{BS}$ versus $SppA_{EC}$

A comparison of the substrate binding grooves of $SppA_{EC}$ and $SppA_{BS}$ reveals that the $SppA_{BS}$ S1 substrate specificity pocket is significantly deeper (Figure 2.13). Structural alignment of the $SppA_{BS}$ and $SppA_{EC}$ active site residues shows that the smaller S1 pocket seen in $SppA_{EC}$ results mostly from three residue substitutions: $SppA_{BS}$ Tyr151 to $SppA_{EC}$ Trp413, $SppA_{BS}$ Ala120 to $SppA_{EC}$ Glu383 and $SppA_{BS}$ Gly226 to $SppA_{EC}$ Arg487 (Figure 2.13.C).

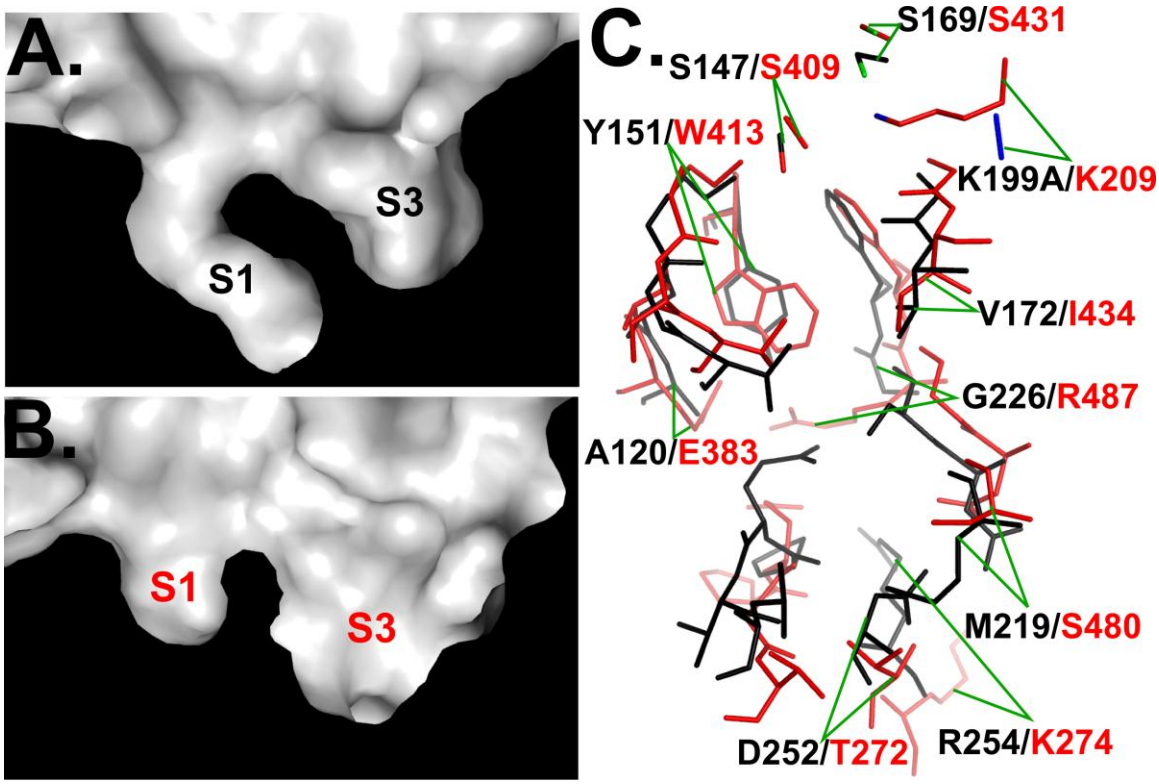


Figure 2.13. Substrate specificity pocket comparison between $SppA_{BS}$ and $SppA_{EC}$.

A. A cross-section of the substrate specificity groove surface in $SppA_{BS}$. **B.** A cross-section of the substrate specificity groove surface in $SppA_{EC}$. **C.** A superimposition of the active site residues in $SppA_{BS}$ (black) and $SppA_{EC}$ (red), rendered in stick.

2.3.11. Differences in the S1 subsite shape of $SppA_{BS}$ and $SppA_{EC}$ are consistent with the range of residues accommodated at the substrate P1 position

Using wild-type active site enzymes and a series of peptide-MCA (4-*M*ethyl 7-*C*umaryl *A*mide) fluorogenic substrates (where the C-terminal residue corresponds to the P1 residue) we observe that both $SppA_{EC}$ and $SppA_{BS}$ show a preference for the leucine tripeptide (LLL) substrate. This is consistent with the predominance of leucine residues in the H-region of signal peptides (Choo et al., 2005). Yet a clear difference can be observed between $SppA_{EC}$ and $SppA_{BS}$ when using the other peptide-MCA substrates (Figure 2.14). The second and third most effective substrate for $SppA_{BS}$ in this series are LRR and LLVY, respectively, whereas $SppA_{EC}$ shows close to no detectable activity using the same compounds. The ability of $SppA_{BS}$ to cleave the LRR and LLVY peptides, where arginine and tyrosine are the P1 residues, is consistent with its deeper and more

polar S1 pocket which could easily accommodate the tyrosine side chain (Figure 2.13). No detectable activity was observed for $SppA_{BS}$ active site mutants (S147A or K199A) or for $SppA_{EC}$ with the corresponding active site mutations (S409A, K209A).

The $SppA_{EC}$ and $SppA_{BS}$ S1 pocket shape differences along with the corresponding observed difference in the range of acceptable P1 residues are somewhat surprising given that the signal peptides (both predicted and experimentally verified) from *B. subtilis* and *E. coli* reveal approximately the same residue content. The only significant difference appears to be that the signal peptides of Gram-positive bacteria are in general longer than those in Gram-negative bacteria (Choo et al., 2005). Given the differences between the substrate specificity pockets, it is plausible that SppA may have substrates in addition to signal peptides.

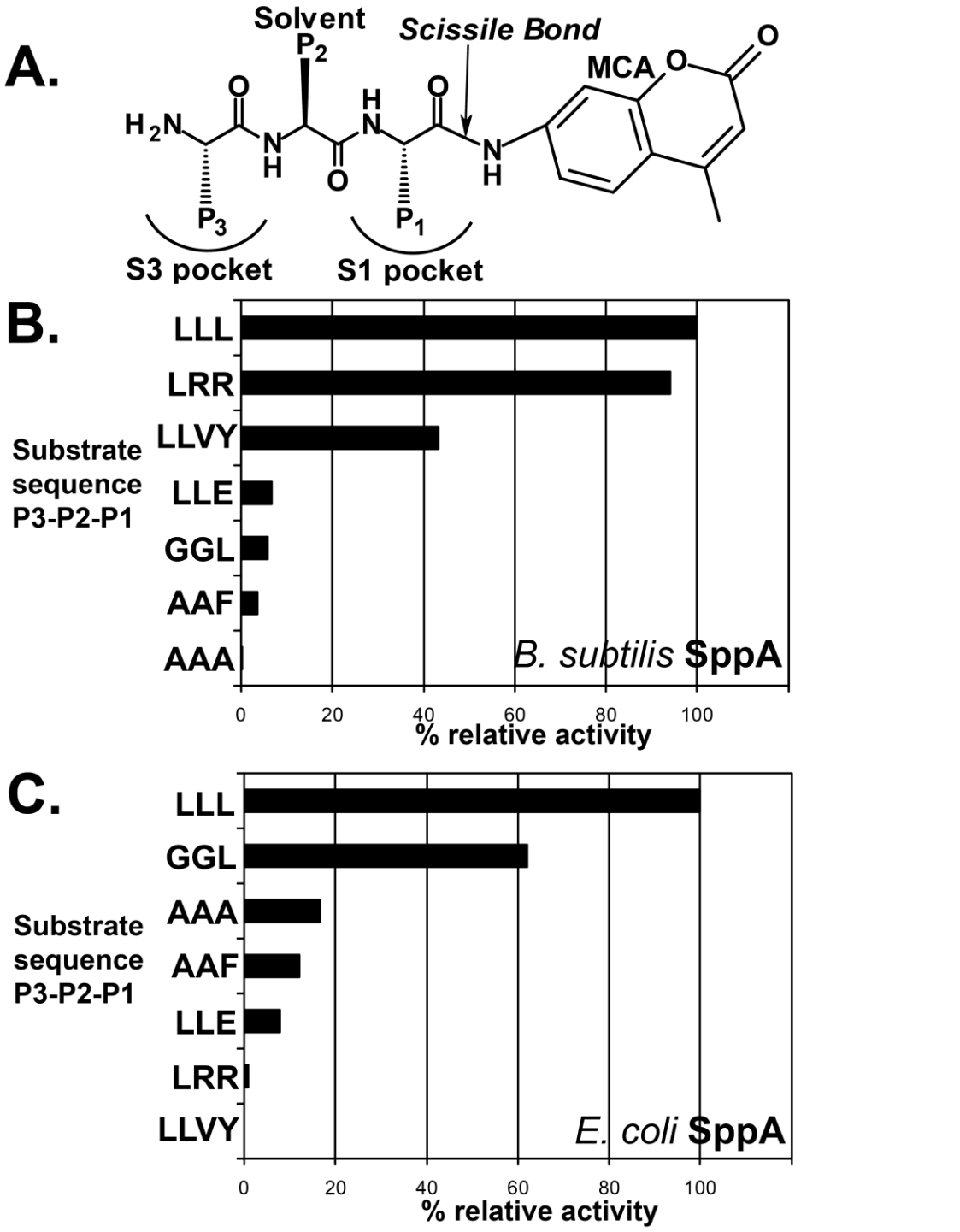


Figure 2.14. Activity profile of $SppA_{BS}$ verses $SppA_{EC}$ reveals a difference in the range of residues the enzymes will accommodate at the P1 position.

Six different fluorogenic peptide substrates were used to compare % relative activity. The C-terminal residue for each peptide corresponds to the P1 residue position. **A.** The activity profile for $SppA_{BS}$. **B.** The activity profile for $SppA_{EC}$.

2.3.12. SppA is capable of digesting folded proteins

Interestingly, we observed that the soluble domains of both $SppA_{BS}$ and $SppA_{EC}$ are capable of digesting fully folded proteins. For example both enzymes digest the *E. coli* lipoproteins BamB and BamD, periplasmic proteins (PDB: 3P1L and 3TGO, respectively) involved in outer membrane protein assembly, to near completion after a 17 hour incubation (Figure 2.15). The active site mutant forms of each enzyme showed no ability to digest.

2.4. Discussion

Using X-ray crystallography, we solved the first Gram-positive bacterial signal peptide peptidase A structure. With the structure solved we discovered the following characteristics; 1. $SppA_{BS}$ forms an octameric complex with eight separate active sites located within the interface between the protomers, 2. Three molecules of $SppA_{BS}$ are required to create an active site and substrate binding site. When the three molecules are aligned, the middle molecule of the three contributes the nucleophile serine and the substrate specificity pocket S1, one of the neighboring molecules provides the general base lysine, and the other neighboring molecule completes the S3 pocket with the middle molecule. 3. Both the S1 and S3 substrate specificity pockets have hydrophobic walls with polar base. The S1 pocket is narrower and deeper than the S3 pocket.

2.4.1. The active site residue Ser147 has unusual phi and psi angles

Interestingly, Ser147 has unusual phi and psi angles, 60° and 120° respectively, forming a *cis* peptide bond, which is also observed in the apo-structure of $SppA_{EC}$. This particular conformation of the catalytic serine residue seems to be unique to SppAs. Other Ser/Lys proteases such as SPase I, LexA, VP4, cIλ, Lon protease and UmuD have a *trans* peptide bond angle with phi and psi that are approximately -61° and -18°, respectively.

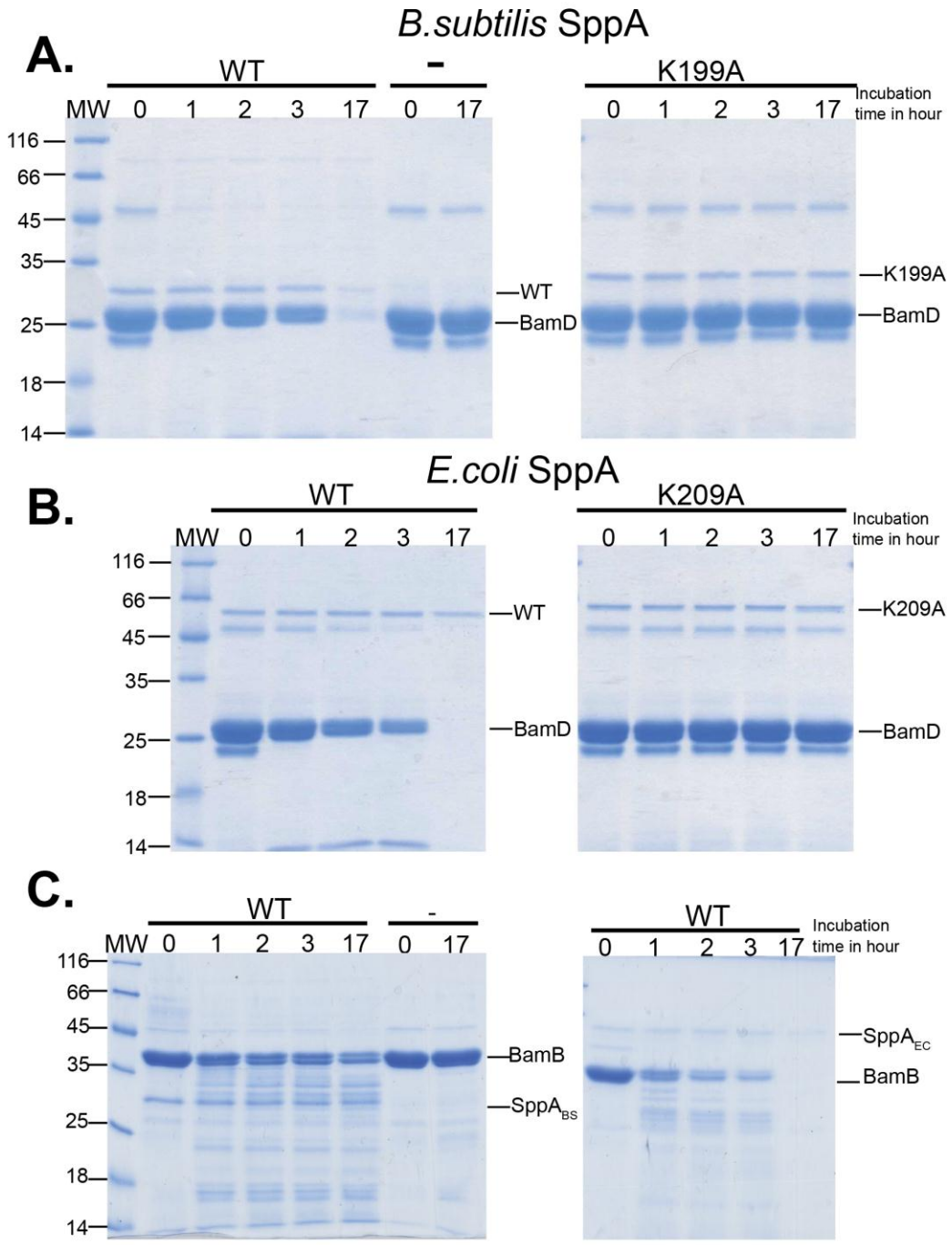


Figure 2.15. $SppA_{EC}$ and $SppA_{BS}$ (soluble domains) are capable of digesting folded proteins.

A. The digestion profile of the *E. coli* lipoprotein BamD by $SppA_{BS}^{\Delta 2\text{-}54}$ with wild-type active site (left) and active site mutant negative control (right) as well as a BamD only negative control. **B.** The digestion profile of the *E. coli* lipoprotein BamD by $SppA_{EC}^{\Delta 2\text{-}46}$ with wild-type active site (left) and active site mutant negative control (right). **C.** *E. coli* lipoprotein BamB digestion by $SppA_{BS}^{\Delta 2\text{-}54}$ and $SppA_{EC}^{\Delta 2\text{-}46}$.

2.4.2. Electron density is missing for a dynamic loop region located inside the cavity of both $SppA_{BS}$ and $SppA_{EC}$

Another interesting observation is the missing loop region (residue 72-82) mentioned earlier in this chapter (Figure 2.2). This loop is located midway between the base of the bowl and the substrate binding groove in $SppA_{BS}$ and is missing in both structure of $SppA_{BS}$ and $SppA_{EC}$. This region in $SppA_{BS}$ is protease resistant, as demonstrated by N-terminus sequencing which showed the N-terminal residue is Leu51, 20 residues before the missing loop. It is surprising that the loop is highly flexible (as the electron density is not observed in the structure solution) but also resistant to the protease at the same time. A possible explanation could be that the loop is located on the inner side of the dome structure where it is not accessible to protease. The function of the loop needs further investigation, but one could speculate that it may be involved in recognizing and bringing the substrate to the binding site, given that the loop is located between the substrate binding site and the wide opening of the bowl where it faces the cytoplasmic membrane where signal peptides are embedded.

2.4.3. What is the function of the positively charged rim at the top of the dome?

As shown in this chapter, $SppA_{BS}$ and $SppA_{EC}$ both form dome-shaped structures with hydrophobic interiors. An interesting feature of the domes is the positively charged rim around the narrow opening of the bowl created by Lys185 for $SppA_{BS}$ and Lys195 & Lsy197 for $SppA_{EC}$. It may be that the function of this 22 Å diameter rim is to form an exit for the digested peptides or amino acids, since a narrower opening is closer to the active site groove than the wider one which sits on the membrane. When Lys185 of the $SppA_{BS}$ is mutated to alanine, changing the property of one residue of the rim from charged to hydrophobic, the resulting mutant ($SppA_{BS}$ K185A) behaved the same as the native $SppA_{BS}$ in its expression and purification. The mutant also showed the same gel-filtration elution profile as the native $SppA_{BS}$, indicating that it forms an oligomer. However, the $SppA_{BS}$ K185A mutant protein precipitates during concentration and the highest concentration we were able to achieve was 3 mg/ml. Native $SppA_{BS}$ can be concentrated to 12 mg/ml. This indicates that K185A may be involved in the structural integrity of $SppA_{BS}$. Further studies with wild type $SppA_{BS}$ may reveal the rim's positive charge involvement in the enzyme's activity.

2.4.4. $SppA_{BS}$ and $SppA_{EC}$ have proteinase activity

In bacteria, there is a self-compartmentalized protease, FtsH, which functions to degrade misfolded and damaged proteins on the cytoplasmic side of the membrane (Dalbey et al., 2012). FtsH is a homo-hexameric membrane integral protease with a large soluble catalytic domain in the cytoplasm and an ATPase domain which unfolds the protein (Langklotz et al., 2012). Wild type $SppA_{BS}$ and $SppA_{EC}$ are capable of digesting folded lipoproteins from *E. coli*. Both SppAs show no preference for substrate secondary structure because the structures of BamB and BamD are solved and their secondary structure are either all α-helical or β-strands, respectively. The ability of SppA to digest folded proteins as well as to cleave peptides is consistent with SppA possibly having a membrane protein quality assurance role. This suggests that SppAs may have a protein quality control function like FtsH, but on the extracellular side of the membrane. Further investigation is required to understand SppA as a proteinase.

3. Crystal structure of *Bacillus subtilis* SppA in complex with a peptide from its own carboxy-terminus.

3.1. Overview

The soluble catalytic domain of *Bacillus subtilis* SppA was subjected to limited proteolysis and after crystal plating resulted in a crystal of the proteolytically resistant fragment of *B. subtilis* SppA. In order to shorten the experimental protocol from Ni^{2+}-NTA affinity chromatography, dialysis, concentration, limited proteolysis, gel-filtration, concentration and crystal plating, supernatant containing $SppA_{BS}^{\Delta 1\text{-}25}$K199A was applied to Ni^{2+}-NTA resin and instead of a step-wise elution, $SppA_{BS}$ protein was released from the column by limited proteolysis. The sample was then concentrated, loaded on to the gel-filtration column and crystal plated after concentration. To improve the diffraction quality and the size of the crystal, paraffin oil was used to cover the drop containing protein and reservoir. By covering the drop with the oil, it slowed down the water vapour diffusion and therefore slowed down the growth of the crystal, leading to a larger sized crystal over time (Chayen, 1997). With the combination of a different purification protocol and crystallizing condition, a new $SppA_{BS}$ crystal diffracting to 2.4 Å resolution was obtained. The molecular replacement method, with the same homology model from the previous chapter, was used to obtain the structure solution. Although the space group and overall structure (octameric complex) was the same as the apo-structure, unexpectedly, after the first cycle refinement a positive difference electron density map that circled inside the substrate binding groove of $SppA_{BS}$ was observed. The positive electron density map resembled a poly-peptide and in order to find out what could be bound in the active site, the same sample that was used to crystalize was also sent for positive ESI-mass spectrometry analysis. With the mass spectrometry data and electron density map, the poly-peptide bound in the substrate binding groove was identified as the $SppA_{BS}$ C-terminus. This posed the question; why is its own C-terminal peptide

bound in the substrate binding groove? Interestingly, wild type $SppA_{BS}$ migrates in an SDS-PAGE gel as a smaller size than $SppA_{BS}$ with inactivated catalytic residue. A series of experiments were carried out to investigate if the C-terminus is a propeptide that functions as an intra-molecular chaperone or as an inactivator of the mature protease. Propeptides have roles in either the proper folding of a protein's tertiary or quaternary structure or in the regulation of the proteolytic activity of a protease (Chen and Inouye, 2008; Khan and James, 1998). Once propeptides complete their function, they are then autoprocessed.

In this chapter, we will discuss in depth how the C-terminal peptide is interacting with the substrate binding groove and what the role of the $SppA_{BS}$ C-terminus could be.

3.2. Materials and methods

3.2.1. Cloning and mutagenesis

The protocol for cloning the $SppA_{BS}$ (UniProt accession number,O34525) construct which lacks residues 1-25 ($SppA_{BS}^{\Delta 1-25}$) and 2-54 ($SppA_{BS}^{\Delta 2-54}$) and contains an N-terminus 6 histidine (HisX6) tag as well as Lys199 mutated to alanine, has been previously described in chapter 2, section 2.2.1. The Quick-Change site-directed mutagenesis procedure (Stratagene) was carried out to mutate the Tyr331 codon to alanine (underlined) using the oligonucleotide 5'-ccgagaatgatggctctctatgcgaag-3' and 5'-cttcgcatagagagccatcattctcgg-3'. Different C-terminus truncated constructs were made using $SppA_{BS}^{\Delta 1-25}$K199A plasmid as a template with the following oligonucleotides mutating the amino acid residues, Met307, G295 and M329 to stop codons (underlined). $SppA_{BS}^{\Delta 1-25, \Delta 329-335}$K199A; 5'-ggcgcgaacaaatagtttaaaagtgaa-3' and 5'-ttcacttttaaactatttgttcgcgcc-3', $SppA_{BS}^{\Delta 1-25, \Delta 307-335}$K199A; 5'-gaggaaagcttctgattaggctcactg-3' and cagtgagcctaatcagaagctttcctc-3', $SppA_{BS}^{\Delta 1-25, \Delta 295-335}$K199A; 5'-ggttcgccgagatagatgtatctctat-3' and 5'-atagagatacatctatctcggcgaacc-3'. The Quick-Change site-directed mutagenesis procedure was utilized and the sequences were confirmed by DNA sequencing (Genewiz).

3.2.2. Expression, purification and limited proteolysis

The $SppA_{BS}^{\Delta 1\text{-}25}$K199A construct described above was expressed in cells and lysed as described previously in chapter 2, section 2.2.2. but was purified as described below. The cell lysate was centrifuged at 30,000 g for 35 minutes and the supernatant was applied to buffer A (20 mM Tris pH8, 150 mM NaCl) pre-equilibrated Ni^{2+}-NTA affinity resin. The supernatant with resin was rocked gently at 4^{o}C overnight and then washed in the following order: with buffer A, with 50 mM imidazole in buffer A, with 75 mM imidazole in buffer A and finally with buffer A alone. 0.5 mg of thermolysin (Sigma) in 25 ml of buffer A was then added to the resin and rocked gently at room temperature overnight. The flow through was collected and concentrated using an Amicon ultra-centrifugal filter device (Millipore) with a 10 kDa cut off. The sample was then applied to a Superdex 200 size-exclusion chromatography column, equilibrated with buffer A, on an Amersham ÄKTA FPLC system at a flow rate of 0.5 ml/min.

$SppA_{BS}^{\Delta 1\text{-}25,\ \Delta 329\text{-}335}$K199A, $SppA_{BS}^{\Delta 1\text{-}25,\ \Delta 307\text{-}335}$K199A, $SppA_{BS}^{\Delta 1\text{-}25,\ \Delta 295\text{-}335}$K199A, and $SppA_{BS}^{\Delta 2\text{-}54}$ were over-expressed and purified as previously described in chapter 2, section 2.2.2. with the exception of $SppA_{BS}^{\Delta 2\text{-}54}$ whose supernatant was applied to buffer A pre-equilibrated Ni^{2+}-NTA affinity resin and supernatant with resin were gently rocked overnight at 4^{o}C.The other protein supernatants were not rocked with resin overnight but applied to Ni^{2+}-NTA affinity chromatography column. The extinction coefficient, molecular mass and pI were calculated using ProtParam (Gasteiger et al., 2005).

3.2.3. Positive electrospray ionization (ESI) - mass spectrometry

A sample for positive ESI-mass spectrometry was prepared using ZipTipsC18 (Millipore). 6 M Guanidine-HCl and 0.1 % trifluoroacetic acid (TFA) was added to thermolysin treated $SppA_{BS}^{\Delta 1\text{-}25}$K199A followed by a superdex 200 and was aspirated and dispensed five times with 0.1 % TFA pre-equilibrated ZipTipsC18. The beads in the ZipTipsC18 were washed using 0.1 % TFA and 5 % methanol by aspirating and dispensing and were then eluted with 10 μl of 50 % acetonitrile (ACN) and 0.1 % TFA. The eluted sample was analyzed using an Agilent 6210 TOF LC/MS system (Chemistry Department in Simon Fraser University).

3.2.4. Crystallization

The sitting-drop vapor diffusion method was used to grow the thermolysin digested $SppA_{BS}^{\Delta 1-25}$K199A crystals. The drop contained 1 μl of protein and 1 μl of reservoir solution and was covered with paraffin oil. The refined reservoir condition was: 23 % tert-butanol and 0.1 M Tris-HCl, pH 8.5 and the drop was equilibrated against 1 ml of reservoir solution at room temperature. The cryo solution contained 20 % 2-methyl-2,4-pentanediol (MPD), 23 % tert-butanol and 0.1 M Tris-HCl, pH 8.5. The crystal was transferred to the cryo solution and flash-cooled in liquid nitrogen.

3.2.5. Data collection

Diffraction images were collected on beamline 08B1-1 at the Canadian Macromolecular Crystallography Facility (CMCF) of the Canadian Light Source (CLS), using a Rayonix MX300HE x-ray detector. Each image, collected using MxDC (Macromolecular Crystallography Data Collector software), was exposed for 0.8 seconds at wavelength 0.9795 Å. The oscillation angle used was 0.5° and a total of 360 images were collected at a detector distance of 305 mm. The images were processed using *HKL2000* (Otwinowski et al., 1993). The unit cell dimensions of the crystal was 87.8 x 131.0 x 207.1 Å and the space group was $P2_12_12_1$. Eight molecules are found in the asymmetric unit with a Matthews coefficient of 2.67 $Å^3$ Da^{-1} (51.9 % solvent). The $SppA_{BS}$ molecular mass after limited proteolysis (~ 28,000 Da) was used to calculate the Matthews coefficient (Kantardjieff and Rupp, 2003). See Table 3.1 for crystal parameters and data collection statistics.

3.2.6. Structure determination and refinement

The structure solution was obtained through molecular replacement using the program *Phaser* (McCoy et al., 2005) and the search model was created using *Chainsaw* (Stein, 2008). The $SppA_{BS}$ homology model was built using the C-terminus domain of $SppA_{EC}$ (Protein Data Bank ID: 3BF0 ChainA) as a template. Conserved side chain residues were kept and non-conserved residues were truncated to the Cβ side chain. The side chain was built and the initial structure was refined using program *Autobuild* within *PHENIX* version 1.6.4 (Adams et al., 2010). *Coot* (Emsley and Cowtan, 2004) was utilized to make manual adjustments to the structure. A final round of restraint

refinement including TLS was performed using *REFMAC5* (Murshudov et al., 2011; Winn et al., 2003) in *CCP4* (Winn et al., 2011). TLS was analyzed using TLS motion determination server (Painter and Merritt, 2006; Winn et al., 2001).

3.2.7. Structural analysis

PyMOL (DeLano, 2002) was used to make the figures. The stereochemistry of the structure was analyzed with the program *PROCHECK* (Laskowski, 2001). *PROMOTIF* (Hutchinson and Thornton, 1996) was used to identify and analyze the secondary structure and motifs within the protein. *AreaImol* (The_CCP4_Suite, 1994) in *CCP4* (Winn et al., 2011) and *PISA* (Krissinel and Henrick, 2007) were utilized for the accessible surface area calculation and *Coot* (Emsley and Cowtan, 2004) was used for measuring the distances between atoms.

3.2.8. SppA self-processing assay

To test intra- versus inter-complex processing of $SppA_{BS}$, 16 μg of $SppA_{BS}^{\Delta 2\text{-}54}$ or 8 μg of $SppA_{BS}^{\Delta 2\text{-}54}$K199A alone or together was incubated at 37 ^{o}C in buffer A. In a total reaction volume of 40 μl, 10 μl aliquots were taken out at time 0, 1 and 24 hours. In case of self-processing assay 15 μg of $SppA_{BS}^{\Delta 2\text{-}54}$K199A or 15 μg $SppA_{BS}^{\Delta 2\text{-}54}$K199A&Y331A alone was incubated at room temperature in buffer A. The total reaction volume was 50 μl and 10 μl aliquots were taken out at time 0, 4, 5 and 6 days. To each aliquot, equal volumes of 2XSDS-Loading dye were added and followed by boiling for 10 minutes. The samples were then run on a 13.5% SDS-PAGE gel and stained with PageBlue stain (Fermentas).

3.2.9. Limited proteolysis of C-terminus truncated constructs and gel filtration

The $SppA_{BS}^{\Delta 1\text{-}25}$K199A and each of the C-terminal truncated constructs; $SppA_{BS}^{\Delta 1\text{-}25,\ \Delta 329\text{-}335}$K199A, $SppA_{BS}^{\Delta 1\text{-}25,\ \Delta 307\text{-}335}$K199A, and $SppA_{BS}^{\Delta 1\text{-}25,\ \Delta 295\text{-}335}$K199A were incubated with thermolysin (500:1 molar ratio) overnight at room temperature. The samples were then applied to a buffer A pre-equilibrated a Superdex 200 column connected to an Amersham ÄKTA FPLC system. The flow rate was 0.5 ml/min.

3.2.10. Inclusion body isolation and attempts to refold $SppA_{BS}^{\Delta 2-54}$K199A and $SppA_{BS}^{\Delta 2-54,\Delta 329-335}$K199A

The constructs $SppA_{BS}^{\Delta 2-54}$K199A and $SppA_{BS}^{\Delta 2-54,\Delta 329-335}$K199A were over-expressed, the cell pellets harvested, and the cells lysed as described in Chapter 2, section 2.2.2. Lysed cell pellets were spun at 3444 g for 10 minutes and the pellets were resuspended in buffer A containing 0.5 % TritonX-100. Resuspended pellets were then centrifuged for 5 minutes at 3444 g. Resuspension and centrifugation steps were repeated three times, except in the final resuspension, buffer A was used (without triton X-100). The purified inclusion body pellets were then stored at -80^{o}C. The final purified inclusion bodies were clean white pellets.

To refold the $SppA_{BS}$ from inclusion bodies, 100 mg of the inclusion body pellet was dissolved in 100 ml of 4 M Guanidine-HCl in buffer B (20 mM HEPES pH7 and 150 mM NaCl) and subjected to centrifugation for 30 minutes at 30,000 g. Supernatant from the centrifugation step was applied to 2 ml of pre-equilibrated Ni^{2+}-NTA affinity resin and washed in the following sequence: 20x the volume of the resin, 3 M, 2 M, 1 M, 0.5 M Guanidine-HCl in buffer B and finally with buffer B alone. $SppA_{BS}^{\Delta 2-54}$K199A and $SppA_{BS}^{\Delta 2-54,\Delta 329-335}$K199A proteins were each eluted with 8 ml of 200 mM, 400 mM and 4 ml of 600 mM imidazole in buffer B. Eluted fractions 200 mM and 400 mM were pooled (16 ml), then dialyzed against buffer B (4 L) overnight at 4^{o}C to remove the imidazole and subjected to buffer B pre-equilibrated gel-filtration (Superdex200) at a flow rate 0.5 ml/min. 300 μl of a ~2 mg/mL protein sample (measured with was injected onto the Superdex200 column connected to ÄKTA FPLC system. The protein concentration was measured using a NanoDrop ND-100 spectrophotometer. The extinction coefficient, 17880 $M^{-1}cm^{-1}$ and 14900 $M^{-1}cm^{-1}$, were calculated based on the $SppA_{BS}^{\Delta 2-54}$K199A and $SppA_{BS}^{\Delta 2-54,\Delta 329-335}$K199A, amino acid sequence using ProtParam, respectively (Gasteiger et al., 2005).

3.2.11. Kinetic and inhibition analysis using a fluorometric peptide assay

The reaction was carried out at 23^{o}C in 100 μl total volume using the SpectraMax M5 Multi-Mode Microplate Reader (Molecular Devices). A black 96 well micro array plate with a clear base (Greiner Bio-One) was used for all the reactions. The fluorogenic

substrate, dodecanoyl-NGEVAKA-MCA (4-*M*ethyl 7-*C*umaryl *A*mide) and SPRMMYLYAK peptide were each dissolved in 100 % dimethyl sulfoxide (DMSO) to make 10 mM stocks. Each well contained 2 % DMSO and peptide concentration ranges from 0.625 μM to 60 μM for Dodecanoyl-NEGVAKA-MCA plus 45 nM of $SppA_{BS}^{\Delta 2\text{-}54}$ in buffer A. Dodecanoyl-NEGVAKA-MCA was synthesized by CanPeptide Inc. and SPRMMYLYAK was synthesized by Chinapeptide Inc. The excitation and emission wavelengths used were 380 nm and 460 nm, respectively. Data points, collected in triplicate, were fitted to the Michaelis-Menten equation to determine the k_{cat} and K_M values using the program Prism5 (GraphPad). In order to test inhibition using SPRMMYLYAK, $SppA_{BS}^{\Delta 2\text{-}54}$ was incubated with SPRMMYLYAK at concentrations of 625 nM, 1.25 μM, 2.5 μM, 5 μM, 10 μM and 20 μM for 30 minutes at room temperature in buffer A. To each SPRMMYLYAK concentration combined with $SppA_{BS}^{\Delta 2\text{-}54}$, Dodecanoyl-NEGVAKA-MCA concentration ranges from 0.625 μM to 60 μM were added and fluorescent intensity was measured. The data points were fitted, using the competitive inhibitor method to determine the K_i, values using the program Prism5 (GraphPad).

Table 3.1. Data collection and refinement statistics

Crystal Parameters	
Space group	$P2_1\,2_1\,2_1$
a,b,c **(Å)**	87.8, 131.0, 207.1
Vm (Å^3/Da)	2.56
% solvent	51.9%
Protein molecules (chains) in A.U.	8
Data Collection Statistics	
Source	CLS
Beamline	08ID-1
Wavelength (Å)	0.9795
Resolution (Å)	44.6-2.4 (2.5 – 2.4)[a]
Total Reflections	638197
Unique reflections	93912 (9085)
R_{merge} [b]	0.060 (0.302)
Mean (*I*)/σ (*I*)	50.3 (6.7)
Completeness (%)	99.2 (97.5)
Redundancy	6.8 (6.3)
Refinement Statistics	
Protein molecules (chains) in A.U.	8
Residues	1865
Water molecules	221
Total number of atoms (protein)	14102
R_{cryst} [c] **/ R_{free}** [d] **(%)**	20.6 /24.0
Average B-factor (Å^2) (all atoms protein)	51.0
Average B-factor (Å^2) (C-terminus peptide only)	58.0
Average B-factor (Å^2) (226 waters)	44.5
r.m.s.d. on angles (º)	0.008
r.m.s.d. on bonds (Å)	1.061

Table note on next page

[a] The data collection statistics in brackets are the values for the highest resolution shell.

[b] $R_{merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| \,/\, \sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of an individual reflection and $\langle I(hkl)\rangle$ is the mean intensity of that reflection.

[c] $R_{cryst} = \sum_{hkl} \big||F_{obs}| - |F_{calc}|\big| / \sum_{hkl} |F_{obs}|$, where F_{obs} and F_{calc} are the observed and calculated structure-factor amplitudes, respectively.

[d] R_{free} is calculated using 5% of the reflections randomly excluded from refinement.

3.3. Results

3.3.1. Self-processing of wild type $SppA_{BS}^{\Delta 2\text{-}54}$

When $SppA_{BS}^{\Delta 2\text{-}54}$, with a native active site, was purified alongside $SppA_{BS}^{\Delta 2\text{-}54}$K199A, with an active site mutation, and run on an SDS-PAGE gel, $SppA_{BS}^{\Delta 2\text{-}54}$ migrated as a smaller protein compared to $SppA_{BS}^{\Delta 2\text{-}54}$K199A (Figure 3.1). This suggests that $SppA_{BS}^{\Delta 2\text{-}54}$ is able to self-process (cleave itself). Furthermore, $SppA_{BS}^{\Delta 2\text{-}54}$, having an intact affinity tag at the N-terminus, was still able to be purified using affinity chromatography suggesting that processing occurs at the C-terminus.

In order to determine if this processing is occurring between octameric complexes (i.e. inter-complex) or within an octameric complex (i.e. intra-complex), wild type $SppA_{BS}^{\Delta 2\text{-}54}$ was incubated with $SppA_{BS}^{\Delta 2\text{-}54}$K199A. $SppA_{BS}^{\Delta 2\text{-}54}$K199A did not result in a smaller protein when it was incubated with $SppA_{BS}^{\Delta 2\text{-}54}$, indicating that the C-terminus processing was between the protomers within an intra-complex reaction (Figure 3.1).

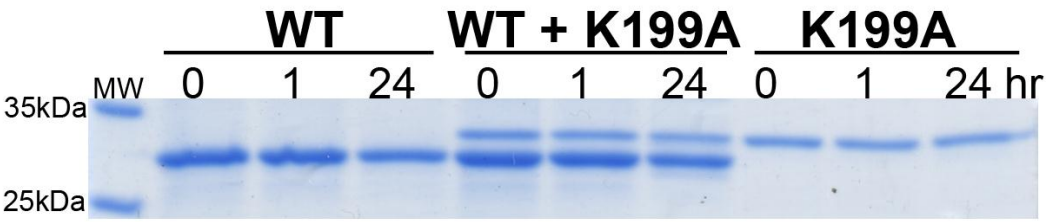


Figure 3.1. C-terminal self-cleavage occurs within the $SppA_{BS}$ complex.

The far left lane is the protein marker. The next three lanes are grouped together as $SppA_{BS}^{\Delta 2\text{-}54}$ (WT) alone, $SppA_{BS}^{\Delta 2\text{-}54}$K199A (K199A) and WT together and $SppA_{BS}^{\Delta 2\text{-}54}$K199A alone at 37°C. Samples were taken out at 0 hour, 1 hour and 24 hours. Loading dye was added to stop the reactions and the results were run on 13.5% SDS-PAGE gel, followed by staining with PageBlue stain.

3.3.2. Crystallization and structure solution of $SppA_{BS}^{\Delta 1\text{-}25}$K199A

Unlike $SppA_{BS}^{\Delta 1\text{-}25}$, the general base Lys199 mutated to alanine construct, $SppA_{BS}^{\Delta 1\text{-}25}$K199A produced adequate amounts of protein and it was subjected to limited proteolysis using thermolysin as mentioned in previous chapter. From here on, the protease resistant fragment of $SppA_{BS}^{\Delta 1\text{-}25}$K199A which resulted in the structure will be called $SppA_{BS}$. The purification protocol was modified from chapter 2; instead of subjecting $SppA_{BS}^{\Delta 1\text{-}25}$K199A to limited proteolysis after the purification, it was subjected to limited proteolysis while bound to the Ni^{2+}-NTA resin. The protease resistant fragment, 28kDa $SppA_{BS}$, was crystal plated using tert-butanol as the precipitant and the drop was covered by paraffin oil. Crystallizing under paraffin oil and without MPD and DDM was different from the crystallizing condition used to obtain the apo-structure of $SppA_{BS}$. The crystal resulting from this drop diffracted to 2.4 Å resolution. As mentioned in chapter 2, it was confirmed by N-terminal sequencing analysis that the starting residue of the themolysin treated $SppA_{BS}$ crystal is Leu51. The structure solution was obtaining utilizing molecular replacement method. The crystal has $P2_12_12_1$ space group and there are eight molecules in the asymmetric unit which forms a dome-shaped structure. Strong electron density was observed from residues 56 to 295, except in the loop region from 73 to 81. In addition to these, the electron density revealed a polypeptide region from the $SppA_{BS}$ C-terminus (residues 326-335) bound within the substrate binding groove (Figure 3.2).

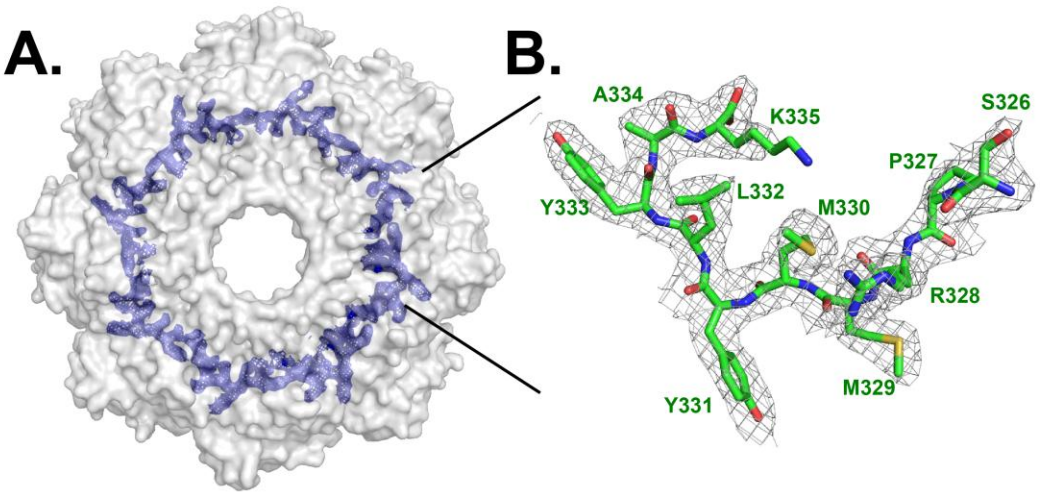


Figure 3.2. $2F_o$-F_c electron density map (1σ) for the $SppA_{BS}$ C-terminal peptide bound in the substrate binding groove of $SppA_{BS}$.

A. Semi-transparent surface representation of eight molecules of $SppA_{BS}$ as viewed from above and $2F_o$-F_c electron density map (1σ) of C-terminal peptides are shown in blue mesh. Black line shows the zoomed in region of the electron density map with the peptide on left on the left **B**. **B.** C-terminal peptide, in stick representation with carbon in green, oxygen in red, sulfur in yellow, and nitrogen in blue are shown together with $2F_o$-F_c electron density map (1σ) (grey mesh). S326 and K335 are N- and C-terminus end of the peptide, respectively. Residues are labeled in green with the residue number.

3.3.3. The electron density inside the binding groove of $SppA_{BS}$ is consistent with a bound peptide

After the initial refinement of the $SppA_{BS}$ structure, the electron density map revealed continuous positive density forming a continuous circle around the concave binding groove of $SppA_{BS}$ (Figure 3.2.A). Peptides were built into the density in each of the eight binding grooves. The longest peptide was ten residues in length and the shortest peptide was seven residues in length.

Since no substrate was added during the purification nor crystallization, the peptide seen inside the binding site was hypothesized to be the $SppA_{BS}$ fragment resulting from thermolysin digestion. After careful observation of the protein sequence at the N-terminus and C-terminus ends of $SppA_{BS}$, the C-terminus sequence ^{331}YLY^{333} was

found to be a good candidate for the P1 and P2' position where the P1 position residue of the peptide is accommodated by the S1 substrate specificity pocket of the enzyme and the S2' pocket accommodates the P2' position residue. Protein sequence alignment showed that Tyr331 is highly conserved among *Bacillus* species (Figure 3.5). To confirm that its own C-terminus was bound to the active site, the thermolysin treated sample, followed by gel-filtration chromatography, was sent for positive electrospray ionization (ESI)-mass spectrometry analysis. The major peak from the positive ESI-mass spectrometry analysis gave an experimental mass of 3355.67 m/z which matched the C-terminus sequence of $SppA_{BS}$ '^{307}(M)FKSEIDFLNMREILSQSGSPRMMYLYAK^{335}' having a theoretical mass of 3355.9 m/z (Figure 3.3). Themolysin cleaves between Met307 and Phe308 while Lys335 is the last residue of $SppA_{BS}$. Eight short peptides (underlined portion of peptide shown above) were built into the difference electron density map within each of the eight active sites of $SppA_{BS}$. The bound peptides revealed that the P1 position corresponds to Tyr331 and the P3 position is Met329 while the P2' position is Tyr333 (Figure 3.4). The electron density fit well with the hypothesized peptide and the S1, S3, and S2' substrate specificity pockets accommodated Tyr331, Met329, and Tyr333, respectively.

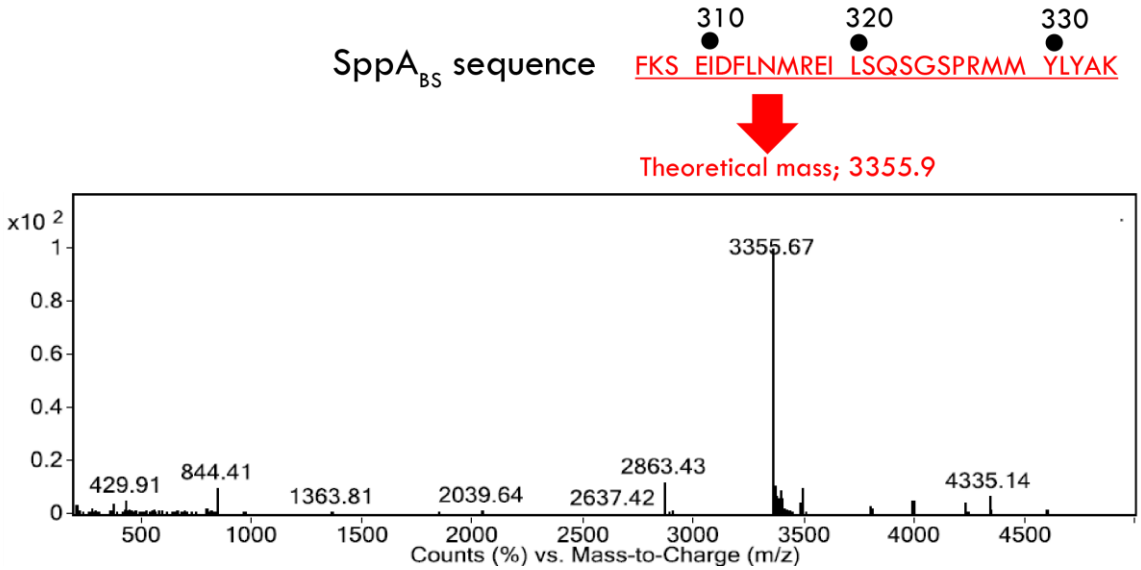


Figure 3.3. Positive electrospray ionization (ESI)-mass spectrometry of thermolysin treated SppA$_{BS}$.

The thermolysin treated SppA$_{BS}$, followed by gel-filtration, was analyzed using positive ESI mass spectrometry. The resulting data showed the peak with highest counts has a mass-to-charge (m/z) of 3355.67 which is almost identical to the theoretical molecular mass of the C-terminus peptide shown in red with theoretical mass of 3355.9 calculated from ProtParam.

3.3.4. Peptide complex reveals the S1, S3 and S2' substrate specificity pocket in SppA$_{BS}$

The S1 substrate specificity pocket accommodates the aromatic side chain of the P1 residue, Tyr331 with the hydrophobic wall and polar base created by residues; Ser118, Gly171, Phe227, Tyr151, Val172, Val116, Ser223 and Glu164 (Figure 3.4.A). The side chain of the P3 residue, Met329, is located within the S3 pocket formed by residues; Leu166, Val220, Met174, Asp252, Met219 and also Glu164, Ser223, Val116 and Val172 are also involved in the formation of the S1 pocket. There is a shallow hydrophobic groove, the S2' substrate specificity pocket, created by residues Ile201, Met202, Val250, Leu155, Met144, Ala146, and Ser169, which accommodates one side of the aromatic side chain of the P2' residue, Tyr333.

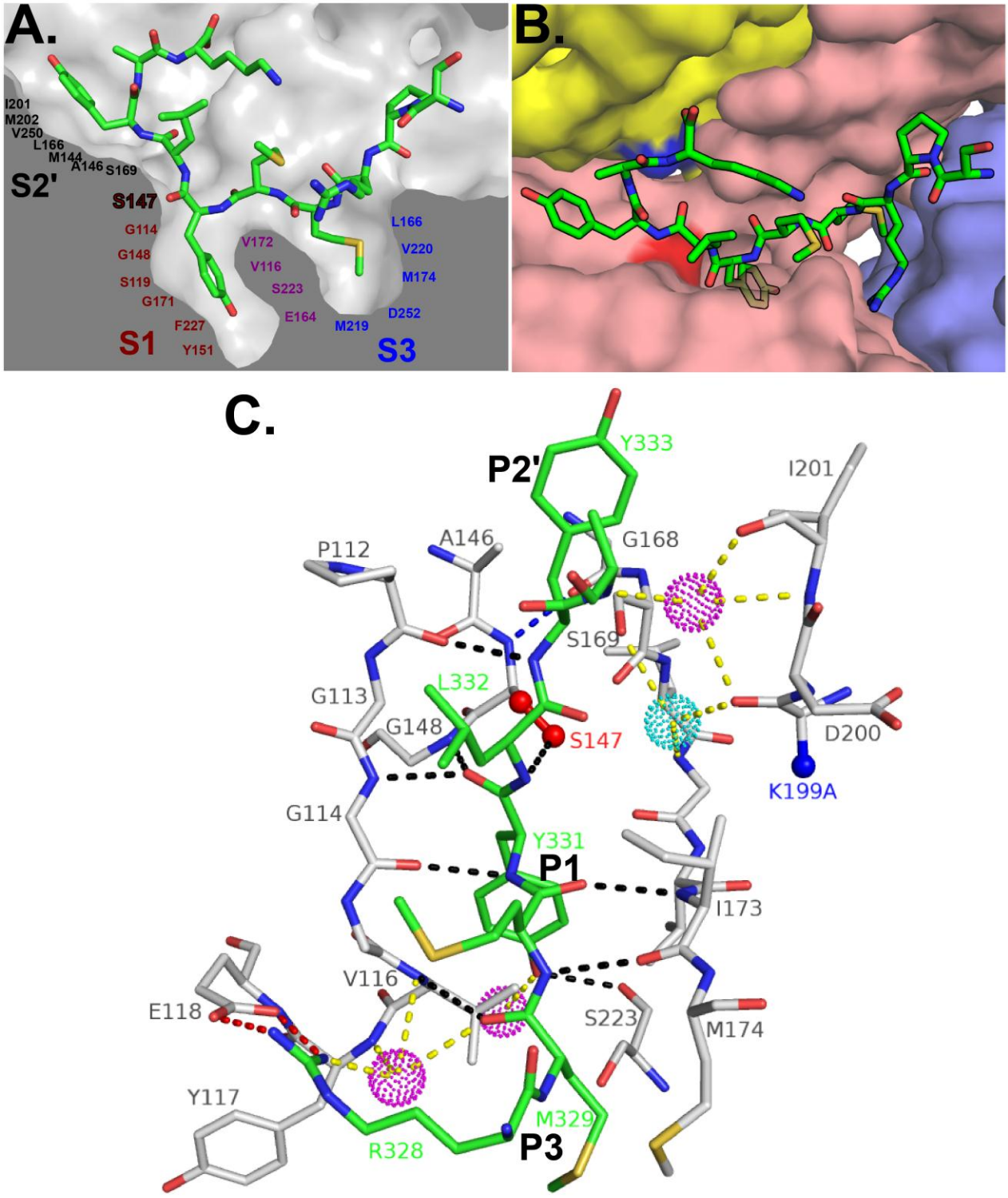


Figure 3.4. (Figure legend on next page)

Figure 3.4. Interactions between the bound C-terminal peptide and the substrate binding groove of $SppA_{BS}$.

C-terminal peptide is shown in stick (carbon in green, oxygen in red, sulfur in yellow, and nitrogen in blue). **A.** The C-terminus of $SppA_{BS}$ fit into the substrate specificity pockets. Residues involved in making the S2' pocket are shown in black, the S1 pocket in red, the S3 in blue and shared between the S1 and the S3 in purple. Cross-section view of binding pocket is shown in molecular surface representation. **B.** The C-terminal peptide bound to the substrate binding groove and active site created by three molecules of $SppA_{BS}$. The yellow colored protomer donates the general base, lysine (blue), the salmon colored protomer provides the S1 substrate specificity pocket with the nucleophile, serine (red), and the light blue and salmon come together to form the S3 substrate specificity pocket. **C.** The binding groove residues that interact with the peptide are shown in grey stick. Dashed lines represent following: Yellow – the hydrogen bond between water molecules and binding groove atoms, black – the hydrogen bond between peptide atom and binding groove atoms, blue – the hydrogen bond between binding groove atom and red – the salt bridge between the main chain of binding groove atom and peptide side chain atom. Catalytic residues are represented as ball and stick, (red for S147 and blue for K199A). Water is represented as dotted sphere (magenta interact with the peptide and binding groove main chain atom, and cyan interacts between binding groove atoms).

3.3.5. Substrate binding groove and C-terminal peptide interactions

Three protomers come together to form one active site within the $SppA_{BS}$ structure. The C-terminal peptides interact with all three protomers that make up each active site (Figure 3.4.B). The yellow protomer in Figure 3.4B, provides the general base (K199A), while the salmon protomer provides the S1, S2' substrate specificity pocket, the nucleophile Ser147 and most of the hydrogen bonding interactions. Lastly, the P3 and P4 residues of the peptide are located in the interface between the light blue and salmon protomer. The side chains of the P2 (Met330) and P1' (Leu332) residues face the solvent.

On average, the buried surface area of the interface between the peptide and the protomer is 667 $Å^2$, whereas the total accessible surface area of each is 1283 $Å^2$ and 8469 $Å^2$, respectively. The peptides are bound in an anti-parallel β-sheet fashion with respect to the residues that line the substrate binding groove (Figure 3.4.C). The $O\gamma$ of nucleophile Ser147 is located close to the carbonyl carbon of the P1 residue, Tyr331 (2.8 Å, average of eight) and the $O\gamma$ of nucleophile Ser147 is within the hydrogen bonding distance to the main chain NH groups of P1' (3.0 Å, average of eight). The carbonyl oxygen of Tyr331 is stabilized by the oxyanion hole created by the main chain

NH groups of Gly114 and Gly148. Gly148 is located after the serine nucleophile but Gly114 is located in the β-strand that forms an anti-parallel interaction with the peptide.

There are three conserved waters seen in all eight active sites. As seen in Figure 3.4.C, waters colored in magenta create a bridge between the C-terminal peptide and protomer, forming hydrogen bonds with both molecules and contributing to a stable interaction. There are seven hydrogen bonding interactions between the main chain and side chain atoms of the C-terminal peptide and protomer, including the oxyanion hole interaction and the Oγ of the nucleophile interaction with NH group of P1 (Table 3.2). There is a salt bridge formation between the P4 residue, Arg328 Nη1 & Nη2, of the C-terminal peptide and Glu118 Oε1 & Oε2 of the protomer. The Oη of Tyr331 is within hydrogen bonding distance of the carbonyl oxygen of Ser223. The Tyr333 Oη is within hydrogen bonding distance of the carbonyl oxygen of Pro329, belonging to the C-terminal peptide bound in the next binding groove.

Table 3.2. Hydrogen bonding between the C-terminus peptide and the substrate binding groove of $SppA_{BS}$.

C-terminus peptide atom	SppA atom	Distance*(Å)
Tyr 331 O	Gly 148 N	2.9
Tyr 331 O	Gly 114 N	2.9
Tyr 331 N	Gly 114 O	2.9
Tyr 331 Oη	Ser 223 O	2.8
Met 329 O	Val 116 N	3.3
Met 330 O	Ile 173 N	3.0
Met 330 N	Ile 173 O	3.0
Leu 332 N	Ser 147 Oγ	3.0
Tyr 333 N	Pro 112 O	3.0

*Average value from the eight peptides

3.3.6. *Potential role of Tyr331 in C-terminus recognition and SppA stability.*

As mentioned previously, Tyr331 is very well conserved among all *Bacillus* species. This residue was mutated in $SppA_{BS}^{\Delta 2\text{-}54}$ and $SppA_{BS}^{\Delta 2\text{-}54}$K199A to investigate if there were any structural or functional differences (Figure 3.5.A). $SppA_{BS}^{\Delta 2\text{-}54}$Y331A was found to still be active, however $SppA_{BS}^{\Delta 2\text{-}54}$K199A_Y331A behaved differently from $SppA_{BS}^{\Delta 2\text{-}54}$K199A. When $SppA_{BS}^{\Delta 2\text{-}54}$K199A was incubated at room temperature for a number of days, it began to degrade, even though the general base had been mutated to alanine (Figure 3.5.B). However, $SppA_{BS}^{\Delta 2\text{-}54}$K199A_Y331A did not show any sign of degradation, even after six days at room temperature.

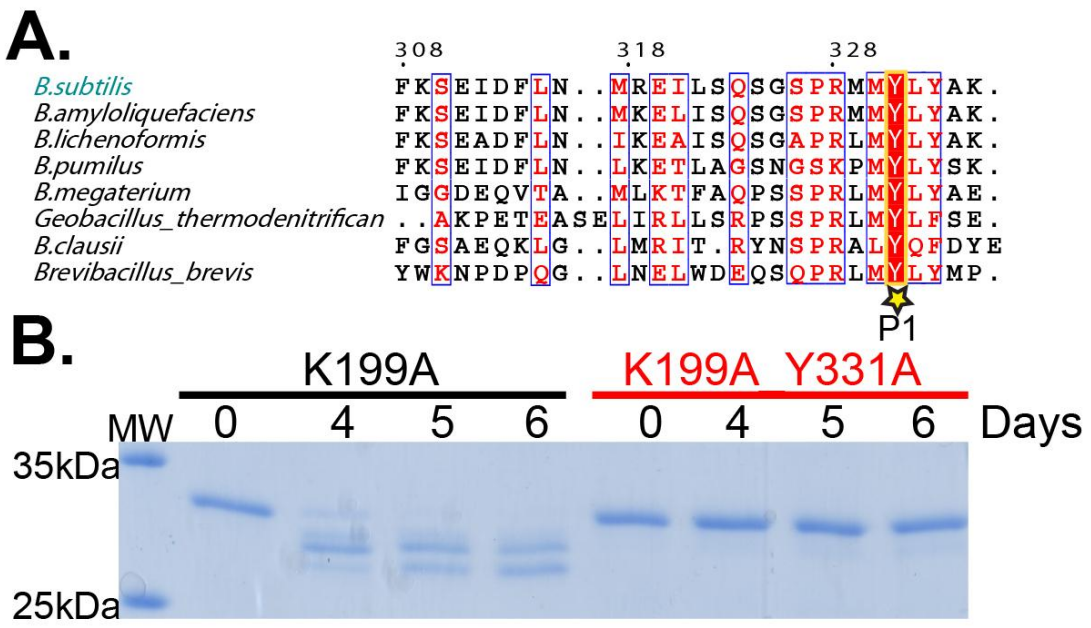


Figure 3.5. Mutating conserved residue Tyr331 to Ala increases stability of $SppA_{BS}$.

Alignment of the C-terminus of SppA from different *Bacillus* species. Tyr331 bound in the S1 pocket of $SppA_{BS}$ is conserved among all *Bacillus* species. UniProt accession numbers: *Bacillus subtilis* (O34525), *Bacillus amyloliquefaciens* (A7Z7M8), *Bacillus licheniformis* (Q65G59), *Bacillus pumilus* (A8FG77), *Bacillus megaterium* (D5DMZ1), *Geobacillus thermodenitrificans* (A4IRT5), *Bacillus clausii* (Q5WEC8), and *Brevibacillus brevis* (C0ZL24). Conserved Y331 (red box) is marked with star and labeled as P1.**B.** Far left lane is molecular marker. $SppA_{BS}^{\Delta 2\text{-}54}$K199A (K199A) and $SppA_{BS}^{\Delta 2\text{-}54}$K199A_Y331A (K199A_Y331A) proteins were each incubated at room temperature and samples were collected after 4, 5 and 6 days.

3.3.7. *$SppA_{BS}$ C-terminal peptide (^{326}SPRMMYLYAK^{335}) competes for its own substrate binding grooves*

Kinetic constants for already self-cleaved $SppA_{BS}^{\Delta 2\text{-}54}$ were determined using a fluorogenic substrate, dodecanoyl-NGEVAKA-MCA which measures intensity of free MCA resulted from enzyme cleavage. A synthetic peptide (NH3-SPRMMYLYAK-COOH) corresponding to the sequence of the C-terminal peptide bound in the $SppA_{BS}$ substrate binding groove was analyzed to see if $SppA_{BS}^{\Delta 2\text{-}54}$ has affinity for the C-terminal peptide. The k_{cat} and K_M values using the dodecanoyl-NGEVAKA-MCA substrate were obtained at different concentrations of the C-terminal peptide (Table 3.3). With these values, the K_i value was calculated using Equation 8.11 from (Copeland, 2000) listed in Prism5. Using six different concentrations of the C-terminal peptide, SPRMMYLYAK, the k_{cat} at each inhibitor concentration stayed approximately the same (Figure 3.6.A). However, the K_M values increased as the inhibitor concentration increased (Figure 3.6.B) and because K_M value increases, k_{cat}/K_M decreases with increase in C-terminal peptide concentration (Figure 3.6.C). The K_i value was determined to be 2.1 ± 0.4 μM. Kinetic analysis suggests that the peptide competes for the substrate binding groove in competitive manner (Figure 3.6.D).

Table 3.3. k_{cat}, K_M and k_{cat}/K_M values using the substrate, dodecanoyl-NGEVAKA-MCA, and in the presence of different concentrations of the $SppA_{BS}$ C-terminal peptide (^{326}SPRMMYLYAK^{335}).

	NH3-SPRMMYLYAK-COOH concentration (μM)						
	0	**0.625**	**1.25**	**2.5**	**5**	**10**	**20**
k_{cat} (s^{-1})	0.066	0.081	0.087	0.076	0.062	0.076	0.098
K_M (μM)	18.9	21.0	23.1	30.0	46.9	123.2	485.4
k_{cat}/K_M ($s^{-1}.M^{-1}$)	3539.3	3838	3785.4	2541.6	1331.0	614.4	202.7

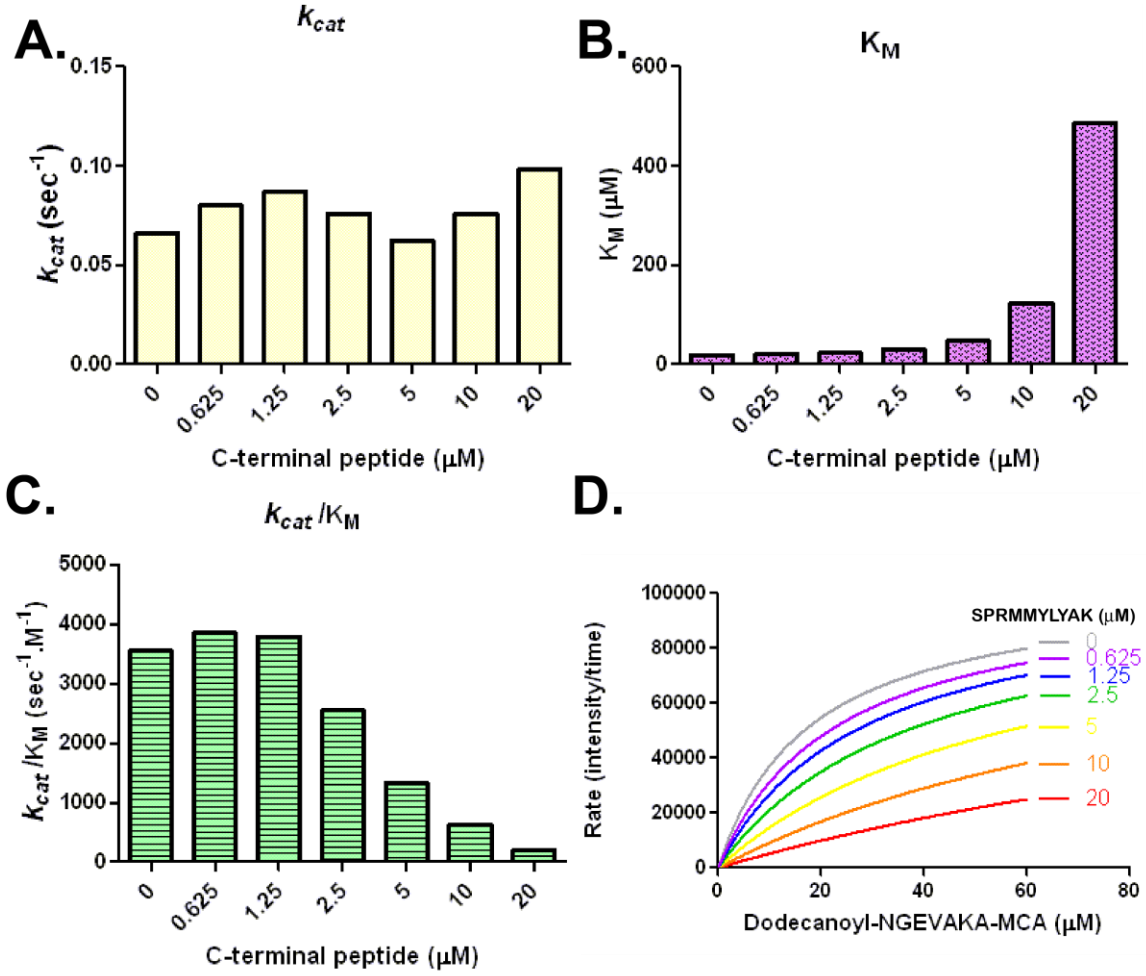


Figure 3.6. Measured $SppA_{BS}$ k_{cat}, K_M and k_{cat}/K_M values (using the substrate dodecanoyl-NGEVAKA-MCA), in presence of difference concentrations of $SppA_{BS}$ C-terminal peptide (^{326}SPRMMYLYAK^{335}).

The bar graphs shows the **A.** k_{cat}, (yellow) **B.** K_M (pink) and **C.** K_{cat}/K_M (green) values at different concentration of $SppA_{BS}$ C-terminal peptide (^{326}SPRMMYLYAK^{335}). **D.** The Michaelis-Menten curve of substrate (dodecanoyl-NGEVAKA-MCA) concentration (μM) versus the rate (intensity/time) using competitive inhibition equation. Different colored graph lines represent the resulting fluorescence measured at different concentration of the C-terminal peptide added.

3.3.8. $SppA_{BS}$ C-terminal truncation studies to investigate its potential role in the oligomerization

Various C-terminus truncations were made to test if the C-terminus plays a role in the oligomerization of $SppA_{BS}$. Three different truncations at the C-terminus were made; the Δ329-335 truncation ends before the conserved residues ^{329}MMYLYK^{335}, the Δ307-335 truncation was based on the self-cleaved wild type $SppA_{BS}^{\Delta 2\text{-}54}$'s molecular

mass estimation from SDS-PAGE gels plus the positive ESI mass spectrometry identified peptide, and the Δ295-335 truncation was based on the last residue (F294) where the electron density is no longer seen in the structure. The three C-terminus truncated constructs were purified from the soluble cytosolic fraction of the *E. coli* expression system and were subjected to limited proteolysis following the protocol which resulted in the solution of $SppA_{BS}^{\Delta 1\text{-}25}$K199A's octameric structure. All three C-terminal truncations showed the same gel-filtration elution profile as the thermolysin treated octameric $SppA_{BS}^{\Delta 1\text{-}25}$K199A, which suggests that the C-terminus truncations do not affect the oligomerization of $SppA_{BS}$ purified from the cytosolic fraction (Figure 3.7) and therefore the C-terminus is not absolutely essential for the correct oligomerization.

Although a fraction of the protein from each expressed construct was soluble in solution, each expressed construct produced a significant fraction of the protein as insoluble inclusion bodies. Interestingly, the faction of protein that formed inclusion bodies from the C-terminal truncated $SppA_{BS}$ ($SppA_{BS}^{\Delta 2\text{-}54,\Delta 329\text{-}335}$K199A) increased significantly compared to the untruncated $SppA_{BS}$ ($SppA_{BS}^{\Delta 2\text{-}54}$K199A) construct. The truncated and untruncated $SppA_{BS}$ inclusion bodies were denatured, refolded and applied to gel-filtration. The refolded untruncated $SppA_{BS}$ was observed to have the same elution profile as the cytosolic fraction purified $SppA_{BS}$, however, the C-terminal truncated $SppA_{BS}$ eluted later in the chromatogram, suggesting that it did not form octamers. The above results are preliminary but suggest that the C-terminus is required for full oligomeric (octameric) assembly of $SppA_{BS}$ *in vitro*.

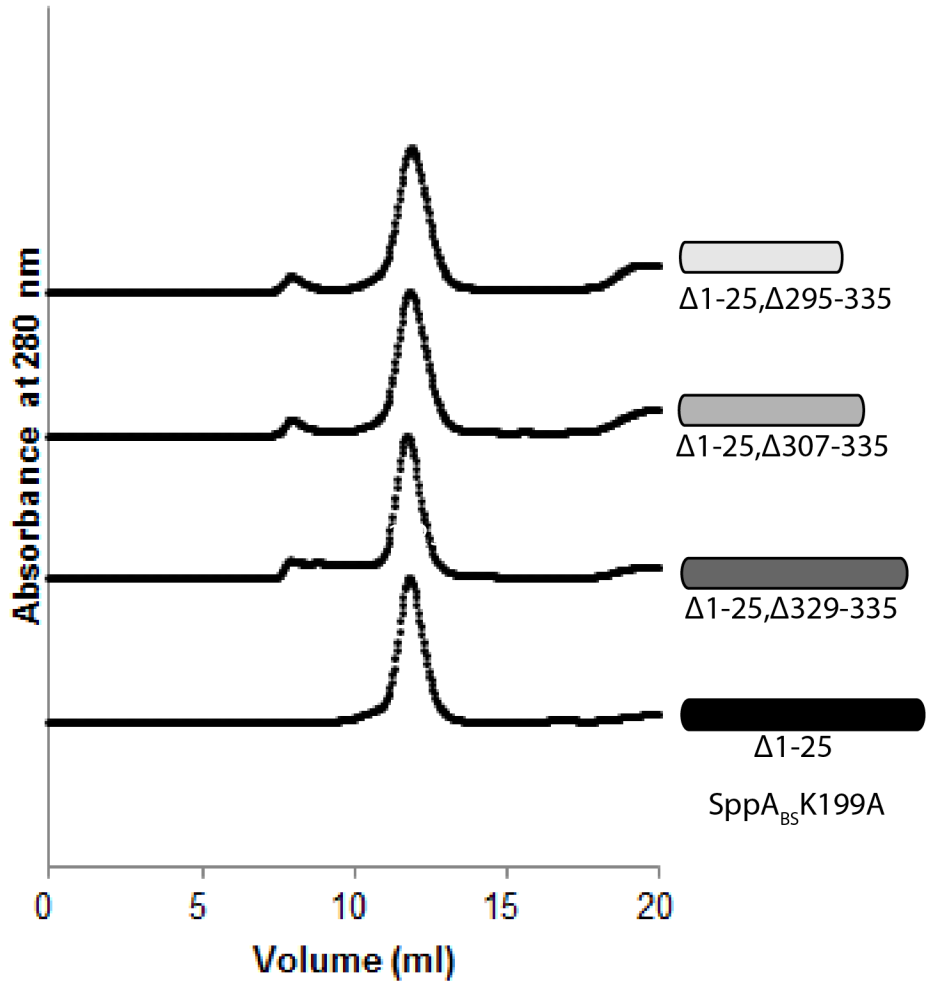


Figure 3.7. Truncating the C-terminus of SppA$_{BS}$ does not prevent a soluble fraction from forming octameric assembly when expressed in the cytoplasm of E. coli.

Gel-filtration chromatography of limited proteolysis treated C-terminus truncated SppA$_{BS}$ constructs along with untruncated SppA$_{BS}$ are shown as elution volume (ml) versus absorbance at 280 nm. A schematic of each construct is shown beside the chromatogram. From bottom to top, SppA$_{BS}^{\Delta 1\text{-}25}$, SppA$_{BS}^{\Delta 1\text{-}25,\Delta 329\text{-}335}$, SppA$_{BS}^{\Delta 1\text{-}25,\ \Delta 307\text{-}335}$, and SppA$_{BS}^{\Delta 1\text{-}25,\ \Delta 295\text{-}335}$. Each construct has the catalytic lysine mutated to alanine (K199A).

3.3.9. Structural changes within SppA$_{BS}$ upon binding a peptide.

The superimposition of the apo-structure and the SppA$_{BS}$ structure with a peptide from its own C-terminus bound in the substrate binding groove has an r.m.s.d. value of 0.60 Å (all atoms). Most differences were observed in β-strand 5 and α-helix 4 of the extension region. Ile1173 and Met174 are part of β-strand 5 which hydrogen bonds to

the C-terminal peptide, and α-helix 4 follows β-strand 5, making the roof of the octameric dome. There are also significant differences found in the active site. First, in the $SppA_{BS}$ structure with the bound C-terminus, the hydroxyl group of Ser147 is rotated approximately 70^{o} away from the peptide bound in the active site (Figure 3.8). Second, in the apo-structure, different rotamer conformations were found in Glu164 and Arg254 which are part of the S3 substrate specificity pocket, but only one conformation was observed for both Glu164 and Arg254 in the $SppA_{BS}$ structure with the bound C-terminus. Third, the residues of the C-terminus peptide bound to $SppA_{BS}$, Ala146 to Gly148, Pro112 to Val116 and Ile173 to Met174 have moved closer to the C-terminal peptide compared to the apo-structure of $SppA_{BS}$ (Figure 3.8). These residues are involved in hydrogen bonding interactions (Figure 3.4).

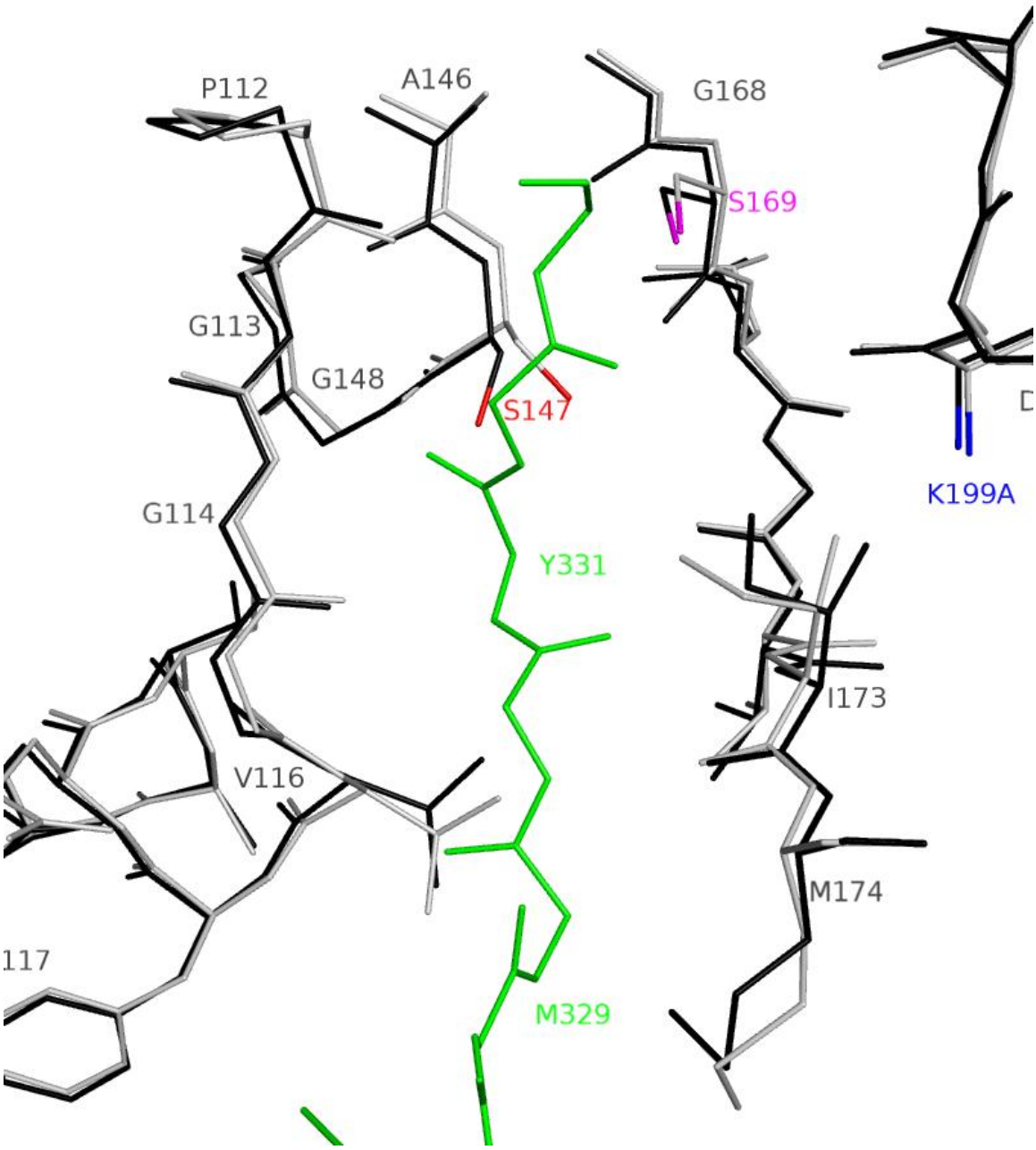


Figure 3.8. Active site superposition of apo- and C-terminus bound $SppA_{BS}$.

Thin stick representation. Grey is the C-terminus bound $SppA_{BS}$, black is the apo-structure of $SppA_{BS}$ and green is the C-terminal peptide (only main chain shown). The nucleophile $O\gamma$ atom of S147 in both structure is in red, the general base K199A $C\beta$ atom is in blue and the general base orienting residue S167's $O\gamma$ atom is in magenta.

3.4. Discussion

In this thesis, I have shown that the overall structure of active sites occupied in $SppA_{BS}$ shares similarity with the previously solved apo-structure. Both $SppA_{BS}$ structures have a homo-octameric complex that resembles a dome in shape. There is a subtle difference between the apo-structure and the C-terminus peptide bound structure of SppA in the extension regions. The hydroxyl group of the nucleophile Ser147 in the C-terminal peptide bound structure has rotated and is now within hydrogen bonding distance of the carbonyl carbon of the P1 residue in order to accommodate the peptide and allows attack on the carbonyl carbon from the *si*-face. In addition, the main chain of the C-terminal peptide bound structure has moved slightly closer to the bound peptide compared to the apo-structure.

There was one unexpected difference, however; a positive difference density map was observed in the $SppA_{BS}$ substrate binding groove, forming a continuous circle inside the octameric complex. This discovery led to a series of experiments, all of which support the hypothesis that the $SppA_{BS}$ substrate binding groove binds to its own C-terminal region. The peptides built in are shorter in length than the peptide analyzed by the mass spectrometry. This could be because the remainder of the peptide's N-terminus may bent out from the active site toward the empty cavity and exposed to the solvent, making it flexible and disordered and therefore not seen in the crystal structure, even though it is evident in the mass spectrometry measurement.

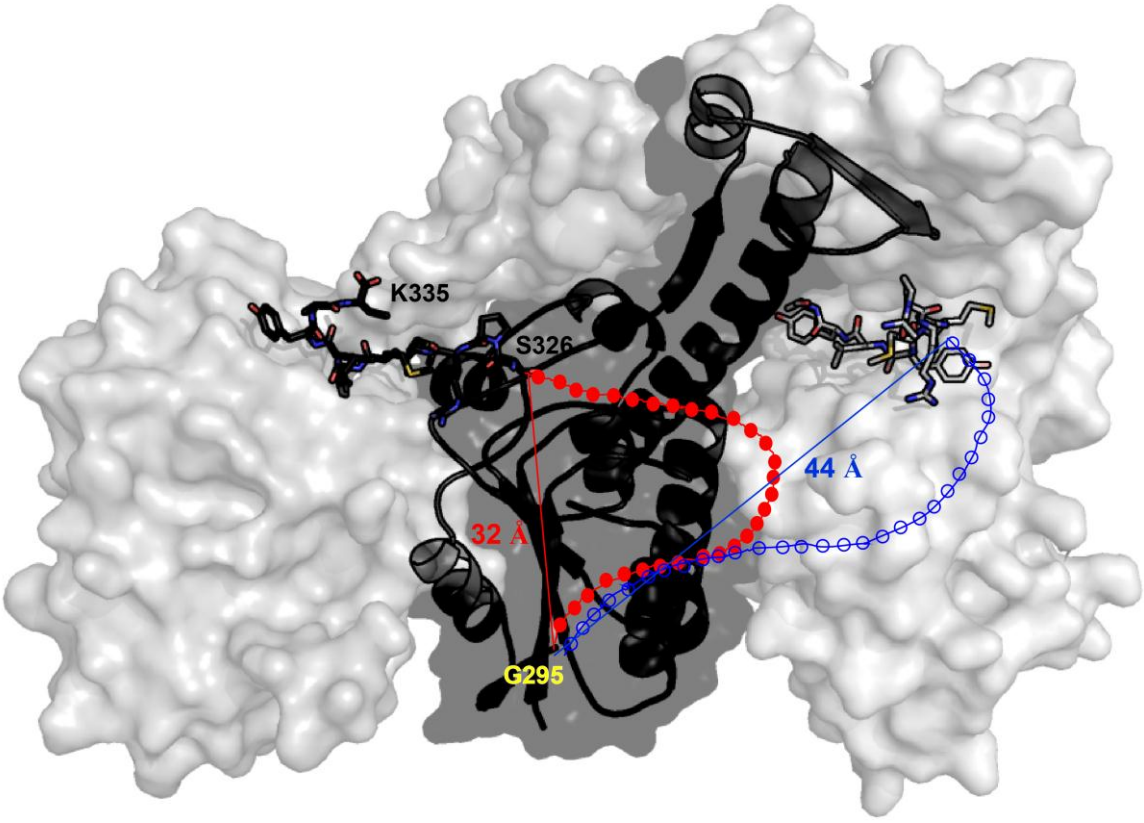


Figure 3.9. A proposed model of how the C-terminus of $SppA_{BS}$ occupies the active site of the neighboring $SppA_{BS}$ molecules.

Presentation of the C-terminus of the black $SppA_{BS}$ bound into the substrate binding groove that sits at the interface of each adjacent protomer, right or left $SppA_{BS}$ molecules (grey). Long loop in red and blue with spheres are drawn to connect black molecule. Each sphere represents one residue. S326 and K335 is the N- and C-terminus end of the C-terminal peptide, respectively. G295 (yellow) is last residue with the clear electron density. The distance between G295 to S326 is shown.

3.4.1. What is the function of the C-terminal fragment of $SppA_{BS}$ bound in active site of $SppA_{BS}$?

The distance measured from the last residue, Gly295 (black molecule in Figure 3.9), of the protomer for which there is clear electron density, to the first residue Ser326 of the C-terminal peptide in either the adjacent left or right neighboring protomer (grey) is approximately 32 Å or 44 Å long, respectively (Figure 3.9). Our new $SppA_{BS}$ structure in complex with a peptide corresponding to its own C-terminus posed a new question. How and why does the C-terminal fragment of $SppA_{BS}$ bind into its own active site? There are 40 residues between the clear density observed last residue, Gly295 and $SppA_{BS}$' C-

terminal end residue Lys335. When the last residue seen in the structure, Gly295, was extended as a poly-alanine chain, the C-terminus of the protomer was long enough to fit into the binding site of either side of the neighboring protomers (Figure 3.9). Based on our modeling studies, the minimum number of residues that are required to reach the S1 substrate specificity pocket from Gly295 is 13 residues, assuming that there is no helical structure in the C-terminus after Gly295. Thirteen residues after Gly295 is Phe308 which is where the wild type $SppA_{BS}$ is thought to self-cleave based on the molecular mass estimation from SDS-PAGE gel. Interestingly, F308, is also an aromatic amino acid, similar to Tyr331 of the C-terminus peptide which is bound in the S1 pocket. Surprisingly, wild type *E. coli* SppA is purified as smaller fragments, which is similar to what has been observed in wild type $SppA_{BS}^{\Delta 2\text{-}54}$ purification (data not shown). Furthermore, like *Bacillus* SppAs, there is a significant amount of conservation at the C-terminus among Gram-negative SppAs (Figure 2.10) which suggests that the C-terminus of SppA in both Gram-negative bacteria and Gram-positive may be important for some unknown function.

There are several proteases that process themselves at either the C- or N-terminus. For example, secreted proteases in the *Bacillus* genus, such as subtilisin, have propeptides at the N-termini, which are not present in the final mature forms. The propeptide usually works as an intra-molecular chaperone to help the protein fold into a tertiary or quaternary structure, followed by cleavage of the propeptide region. Type II intra-molecular chaperones are usually found in the C-terminus region of the protein and are involved in the formation of the quaternary structure of the protein assembly, which is necessary for them to be functional (Chen and Inouye, 2008). The C-terminus region is then self-cleaved off if it is part of a protease or by other proteases. The Type I intra-molecular chaperones which are found at the N-terminus are sometimes not essential for the general folding of the protein but rather the correct conformation of the active site.

$SppA_{BS}$, which is an octameric species and binds to and self-cleaves its C-terminus, is a good candidate for having a type II intramolecular chaperone. My C-terminal truncation studies, show that all the constructs formed oligomers, despite the different C-terminus truncations, when they are purified from the cytosolic fraction. However, when these same constructs are denatured and refolded the C-terminus was required for the oligomeric assembly *in vitro*. $SppA_{BS}$, a Gram-positive integral membrane protease, was over-expressed and purified from the Gram-negative *E. coli*

cytosolic fraction, but this is a completely different environment from where $SppA_{BS}$ exists physiologically. Further study is needed to investigate the function of the C-terminus as a Type II intra-molecular chaperone as discussed in Chapter 4, section 4.3.1.

Another potential function of the propeptide is to act as an inhibitor for the mature protease, a type of regulatory function on top of the self-compartmentallization. Carboxypeptidase Y, from the yeast *Saccharomyces cerevisiae*, is a Ser/His/Asp utilizing protease that is synthesized as a precursor form with an N-terminal propeptide extension (Nagayama et al., 2012). The activity of Carboxypeptidase Y was inhibited in a competitive inhibition manner when a nine residue long peptide, corresponding to the N-terminal propeptide region, was added. Similarly, I showed that a synthetic peptide, corresponding to the $SppA_{BS}$ C-terminal peptide, competes for the substrate binding groove in a competitive manner with a fluorogenic substrate.

Other types of proteases that are able to undergo self-cleavage are LexA, cIλ, UmuD and DegP (Burckhardt et al., 1988; Jomaa et al., 2009; Little, 1984). These proteases self-process to either inactivate or activate themselves. The LexA repressor is inactivated, and in contrast UmuD is activated after self-cleavage, whereas DegP becomes inactive after cleaving off its N-terminus, possibly to remove excess DegP from the system (Jomaa et al., 2009). Similar to these proteases, the C-terminus of $SppA_{BS}$ may be involved in activation and inactivation, although further investigation on this matter is needed. Finding the answers to why and how $SppA_{BS}$ recognizes and processes its own C-terminus could offer significant insight about the function and molecular mechanism of $SppA_{BS}$.

4. Conclusion and future directions

4.1. Conclusion

Signal peptide peptidase A (SppA), found in eubacteria, archaea and plant chloroplasts, belongs to the S49 protease family which utilizes the Ser/Lys catalytic dyad. Bacterial SppA is a membrane bound protease with a large, soluble, self-compartmentalized catalytic domain that sits on the extracellular side of the cytoplasmic membrane, and its function is to cleave remnant signal peptides (Hussain et al., 1982b; Kim et al., 2008; Wang et al., 2008). The Gram-negative bacterial SppA structure has been solved earlier in the Paetzel lab. In this thesis, the novel Gram-positive bacterial SppA apo-structure and SppA in complex with its own C-terminus peptide are presented.

The apo-structure of the catalytic domain of *B. subtilis* SppA has been solved using X-ray crystallography and refined to 2.4 Å resolution. There are eight $SppA_{BS}$ molecules in the asymmetric unit, forming a dome shaped, octameric complex. The octameric state of $SppA_{BS}$ is also observed in solution by size exclusion chromatography and multi-angle light scattering analysis. The octameric complex has a small, positively charged opening at the top and a wider opening at the bottom of the dome. Inside the dome is a concave groove where the substrate binding sites are located. The overall electrostatic analysis showed that the exterior surface of the dome is evenly distributed with patches of positive and negative charges while the interior of the dome has mostly hydrophobic surface with negative charges around the concave groove.

Each protomer of $SppA_{BS}$ contains a serine nucleophile (Ser147) and a lysine general base (K199A); however, they are located in different regions of the protein, more than 29 Å apart. Only upon its assembly into the octameric state do the serine and lysine come into close proximity, with neighboring protomers each providing one half of the catalytic dyad, thus producing eight separate active sites at the interfaces between the protomers. One complete active site is created when three $SppA_{BS}$ protomers come

together. Lys199 is contributed by the first protomer and Ser147, the S1 substrate specificity pocket, and the oxyanion hole are from second protomer. Lastly, the S3 substrate specificity pocket is created between the second and third protomer.

The $SppA_{BS}$ S1 substrate specificity pocket is deep, narrow and hydrophobic but with a polar base. The S3 pocket, which is constructed from two neighboring proteins, is shallower, wider and more polar than the S1 pocket. A comparison of the substrate binding groove in $SppA_{BS}$ to that in $SppA_{EC}$ reveals a significant difference in the size and shape of the S1 pocket which is reflected in the type of peptides the enzymes are capable of cleaving. Substrate preference was screened using different ranges of peptide-MCA substrates. The result showed that $SppA_{BS}$ prefers substrates with arginine, leucine and tyrosine at the P1 position and leucine at the P3 position. This agrees with the observation that the S1 substrate specificity pocket is deep and hydrophobic with a polar base. $SppA_{EC}$ also had a preference for leucine but a significant difference was observed when, unlike $SppA_{BS}$, $SppA_{EC}$ was not able to cleave a substrate with Tyr at the P1 position, as the S1 pocket of $SppA_{EC}$ is not deep enough to accommodate tyrosine.

The same $SppA_{BS}$ construct, but purified in a different way, resulted in an $SppA_{BS}$ having its own C-terminal peptide bound into each of the eight active sites, creating a perfect circle of peptides. The complex shows the C-terminus peptide is bound to the substrate binding groove, in an anti-parallel β-sheet fashion, with a buried surface area of 667 Å. The C-terminus residues, Tyr331, Met329 and Tyr333, are bound in substrate specificity pockets S1, S3 and S2', respectively. Ten hydrogen bonds are observed between the C-terminus peptide atoms and the substrate binding groove atoms, as well as a salt bridge between Arg328 of the peptide and Glu118 of the protomer. The interaction is further stabilized by conserved waters that form hydrogen bonds between the C-terminal peptide and the protomer.

Wild type $SppA_{BS}$ showed self-cleavage at its C-terminus but self-cleavage was dramatically decreased when the general base Lys199 was mutated to Ala. Mutating Tyr331 to an alanine prevented self-cleavage completely in $SppA_{BS}$K199A. Modeling studies show that it is possible for the C-terminus of each of the SppA protomers to bind into the active sites that are formed at the interfaces between protomers. A series of truncations at the C-terminus indicate that the octameric assembly is not affected by C-

terminus deletions. However, kinetic analysis suggests that the C-terminal peptide displays a competitive inhibitor characteristic and therefore may be involved in regulating activity. Further research is

Previous SppA research focused primarily on *E. coli* SppA. Researchers reported that *E. coli* SppA forms a tetrameric, dome-shaped structure that utilizes the Ser/Lys catalytic dyad to preferably cleave the hydrophobic region of signal peptide (Kim et al., 2008; Novak and Dev, 1988). However, the only known information on *B. subtilis* SppA was that it processes signal peptide from pre-protein and is not an essential gene for viability (Bolhuis et al., 1999). From my research on *B. subtilis* SppA, we now have new information that has not been reported previously; 1. $SppA_{BS}$ forms an octameric complex and its active site is created only upon assembly of octamer because Ser and Lys are donated from separate protomers to make one active site. 2. $SppA_{BS}$ not only has a preference for hydrophobic residues at the P1 position but also prefers the charged residue, arginine and the aromatic residue, tyrosine, 3. $SppA_{BS}$ has a C-terminus propeptide region which may function as an inhibitor for the mature protease, 4. *E. coli* SppA and *B. subtilis* SppA are able to digest folded proteins.

4.2. Evaluation and limitation of the experimental techniques used to analyze *B. subtilis* SppA structure and substrate preference

X-ray crystallography has provided us with in depth knowledge into SppA's structure and how it is able to recognize and interact with its own C-terminus. We are confident that $SppA_{BS}$ exists in an octameric state based on both the solved crystal structure and supporting data from multi angle light scattering analysis.

For both apo- and C-terminus bound structures, the diffraction data's quality and reliability were supported by statistical values such as completeness, redundancy and R_{merge}. The overall completeness of the reflection measurement to the maximum resolution was 99 % for both crystals, and the redundancy of the measurement were 7.4 and 6.8 for apo- and C-terminus bound protein crystal, respectively. The overall R_{merge} values, a measure of disagreement among multiple measurements, were less than 9 %

for both. These statistics indicated that the diffraction datasets were complete and the reflection intensity measurements were of good quality for structure elucidation and refinement.

The R_{crys} value is used to measure and validate the agreement between the experimental data and the refined model. As a rule of thumb, the refined atomic positions of a structure are considered reliable fit to the electron density when the R_{crys} values is less than its highest resolution used to refine the model. For example, when $SppA_{BS}$ model was refined to 2.4 Å resolution, the R_{crys} value was expected to be lower than 24 %. On the other hand, the R_{free} value is calculated from 5 % of randomly selected reflections which were not used in the R_{crys} calculation, and R_{free} is used to cross-validate the R_{crys}. The R_{free} value is usually higher than R_{crys} but should remain similar in the absence of model bias during refinement (Blow, 2004; 2002). During repeated refinement cycles of apo- and C-terminus bound $SppA_{BS}$ structures, the R_{crys} and R_{free} values decreased steadily, and for the final refined structures the values were 20.6/24.1 and 20.6/24.0 (R_{crys}/R_{free}), respectively. These values provide strong validation regarding the fit and accuracy of the refined structures.

For both apo- and C-terminus bound structures, the clear electron density that fits accurately to specific residues, especially around the aromatic residues tyrosines (residue numbers: 81, 117, 138, 150, 151, 179, 224, 251 and 268, for C-terminus bound structure 331 and 333 as well) and phenylalanines (residue numbers: 89, 160, 190, 227 and 267) indicated that the overall protein sequence was in correct register. Additional structure validation was provided by various structural parameter indicators. Examination of the geometry of the peptide linkage and side chains demonstrated that no outlier residues were present in the unfavorable region of the Ramachandran plot. As well, the r.m.s.d. values of the bond angle and bond length were also within the theoretical values.

A significant limitation to the $SppA_{BS}$ analyzed in this thesis is that it does not represent the full length, native form of *B. subtilis* under physiological conditions. The crystal structure was obtained and the electron density map shows only the soluble catalytic domain of $SppA_{BS}$ (residues 57-295), lacking a significant portion of the C-terminus (residues 296-335), and the crystal was grown in a highly organic condition, consisting of tert-butanol, MPD and paraffin oil. In addition, we were not able to visualize

the flexible loop region (residues 72-82) which is missing in both the *E. coli* and *B. subtilis* solved structures. To overcome these limitations, additional experiments are proposed in the Future Direction section of this thesis.

The substrate preference of *B. subtilis* SppA was determined using short peptides linked to a fluorogenic moiety, MCA. All the conditions including buffer, temperature, DMSO concentration, preparation for enzyme and substrate were kept consistent throughout the experiments. Therefore, the relative activity of *B. subtilis* SppA throughout the range of peptide-MCA substrates yielded significant results. It was fortunate that *B. subtilis* is able to accommodate such a large moiety (MCA) in the substrate specificity pocket S1', but this was also a drawback because we could not derive information on the substrate preference for the P' positions of the substrate. Another drawback is that the experimental peptides were shorter than SppA's known endogenous substrate, signal peptide. Analyzing signal peptide or membrane protein digestion using HPLC or FRET substrates with longer sequences would be helpful for further study of the enzymatic and biological function of *B. subtilis* SppA.

4.3. Future Directions

During this thesis project, the molecular structure of *B. subtilis* SppA was solved and its substrate preference determined. These findings have also raised many new and interesting questions. Here, I propose some future experiments to investigate some intriguing queries that may result in a greater understanding of SppA's function and structure.

4.3.1. What is the function of the C-terminus?

The C-terminal bound structure solved by X-ray crystallography, together with the results from other experiments, showed that $SppA_{BS}$ recognizes, binds and cleaves its own C-terminus. The C-terminus truncation experiment suggests that truncating the C-terminus does not affect the oligomerization of $SppA_{BS}$ which has been expressed in the cytosolic fraction of the *E. coli* expression system, however it does affect oligomerization when the proteins need to be folded *in vitro*. Further investigation is required to determine the function of the C-terminus. To investigate if the C-terminus is involved in

oligomer formation, one could denature the truncated $SppA_{BS}$K199A (full-length with transmembrane segment intact) and refold again, alongside untruncated $SppA_{BS}$, and observe if the truncated $SppA_{BS}$ is still able to form an octamer after refolding using gel-filtration or MALS. In refolding experiments one could add exogenous peptides corresponding to the C-terminal sequence to see if it were able to rescue the octameric oligomerization. Another important experiment would be to truncate the wildtype SppA and observe if there are significant changes in activity between the truncated wildtype SppA and untruncated wildtype SppA. This may show if the C-terminus plays a role in SppA's functioning as an enzyme, i.e. forming functional active sites.

4.3.2. How does SppA associate with the membrane? Does SppA extract the membrane embedded signal peptide?

As mentioned previously, SppA's structure was solved from a crystal grown in an artificial environment. SppA is a membrane protease which is thought to sit on the membrane (Hussain et al., 1982b; Wang et al., 2008). It will be interesting to discover if there are conformational changes when SppA is exposed to physiological conditions and to find out how strongly SppA associates with the membrane. To determine this, an SPR (surface plasmon resonance) experiment could be carried out to examine the binding affinity of the soluble domain of SppA towards a liposome created from the membrane extract. Meanwhile, co-crystallization of SppA with the liposome could reveal structural changes due to association with the lipid, which would mimic physiological conditions given that SppA is a membrane bound protease.

There have been no direct studies to show that *B. subtilis* SppA, a compartmentalized protease that sits on the membrane with its active site above the membrane, extracts and degrades signal peptides. One experiment approach that could show if SppA is able to digest signal peptides from the membrane would be to purify pro-OmpA nuclease A (PONA), a protein which contains the signal peptide sequence of OmpA linked to nuclease A, and insert the signal peptide portion of PONA, after SPase I digestion, into the liposome. Then, digestion could be observed by using SDS-PAGE gel or by identifying the digested fragments using mass spectrometry. Another experiment would be to monitor digestion using a fluorogenic substrate which has a lipid tail at the

N-terminus, allowing for insertion into the liposome. This may reveal if SppA alone is able to digest the membrane embedded signal peptide.

4.3.3. Does SppA have an interacting partner?

YteJ (UniProt accession number: O34424) is an uncharacterized protein that is thought to be transcribed in the same operon as SppA in the *B. subtilis* genome. YteJ is a transmembrane protein, 164 amino acids in length, whose function and structure are unknown. However, YteJ may be involved in interaction with SppA. Since YteJ is predicted to be a membrane spanning protein, it is possible YteJ could interact with the signal peptide embedded in the membrane and deliver the signal peptide into the dome of the SppA complex. Some self-compartmentalized proteases similar to SppA, such as ClpP and Lon proteases, have hetero-multi-subunit complexes which have a separate subunit that works to unfold and deliver the substrate to the catalytic subunit of the complex (Sauer and Baker, 2011). If experiments using SPR and multi-angle light scattering confirm the interaction between SppA and YteJ, co-crystallizing *B. subtilis* SppA with YteJ and solving the structure would be the next step. In addition, co-crystallization of the catalytic domain or full length SppA in a liposome, either in the presence or absence of YteJ, could provide new information on SppA's structure under physiological conditions.

4.3.4. How is SppA able to degrade folded protein? And what could be its endogenous substrate?

As mentioned previously, SppA is able to digest folded lipoprotein from *E. coli in vitro*. How is SppA able to recognize and degrade folded protein when it does not have an ATPase or PDZ domain and when its active site is outside the cell where there is no access to ATP? One experimental approach to investigate the unfoldase activity of SppA would be to monitor changes in the intensity of green fluorescent protein (GFP) or in the intrinsic fluorescence of the substrate protein as the catalytically inactive SppA mutant unfolds the substrates. GFP has a β-barrel structure with a molecular mass 27 of kDa. Since SppA is able to digest proteins with an all β-strand secondary structure, GFP may be an ideal protein for studying unfoldase activity by measuring decrease in fluorescence intensity as $SppA_{BS}$ unfolds GFP (Rothballer et al., 2007). If SppA is found

to have unfoldase activity, what would be its substrate *in vivo*? One method could be performing two-dimensional electrophoresis analysis of the cell lysate from *B. subtilis* with intact SppA gene compared to a knockout strain (Major et al., 2006). By comparing the relative spot intensities, it may be possible to identify the endogenous substrate for SppA because the lysate from the knockout strain could contain the accumulated endogenous substrates.

4.4. Where could studying *B. subtilis* SppA lead us?

Over the years, *Bacillus* strains have been used as protein 'factories' producing pharmaceutical proteins, membrane proteins, and protein complexes because of their ability to secrete proteins into the surrounding medium (Schallmey et al., 2004). *Bacillus* strains are preferred over *E. coli* expression systems because inclusion body (aggregation of misfolded proteins) formations can be prevented. Moreover, expressed pharmaceutical proteins do not become contaminated with lipopolysaccharide (LPS) which is located on the outer membrane of *E. coli* and is harmful to humans, causing septic shock. There have been attempts to improve the level of secreted protein in the medium using *Bacillus* by optimizing the signal peptide sequence, as well as work on propeptides to encourage proper folding of the protein once it is secreted into the medium (Degering et al., 2010; Takagi and Takahashi, 2003; Zweers et al., 2008). In order to meet the artificial demand for bulk production of proteins, promotion of faster transportation and translocation and more efficient hydrolysis of the signal peptides embedded in the membrane are required, since the membrane has limited space. However, faster transport and translocation of the pre-proteins may result in heavy traffic in the cytoplasmic membrane which in turn increase the number of damaged or aggregated proteins. Further research on SppA's function as a putative membrane protein quality control protease may lead to novel strategies to improve the clearing the cytoplasmic membrane of these misfolded or damaged proteins. Furthermore, this research provides a greater understanding of signal peptide processing in the *Bacillus* model for Gram-positive bacteria, and could lead to the engineering of a signal peptide substrate more susceptible to SppA allowing faster clearance from the membrane. This would ultimately improve the *Bacillus*' function as a protein expression host organism for the pharmaceutical industry. In addition, structural and activity investigations into SppA

could lead to the development of SppA inhibitors that could function as antibiotic compounds that interfere with bacterial survival by interrupting signal peptide hydrolysis. Therefore, an inhibitor for SppA could make the bacterial cell more susceptible to antibiotics.

References

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W.*, et al.* (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D *66,* 213-221.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Research *38,* W529-W533.

Baird, L., Lipinska, B., Raina, S., and Georgopoulos, C. (1991). Identification of the Escherichia coli sohB gene, a multicopy suppressor of the HtrA (DegP) null phenotype. J. Bacteriol. *173,* 5763-5770.

Barnett, J.P., Eijlander, R.T., Kuipers, O.P., and Robinson, C. (2008). A minimal Tat system from a gram-positive organism: a bifunctional TatA subunit participates in discrete TatAC and TatA complexes. J. Biol. Chem. *283,* 2534-2542.

Bechtluft, P., Nouwen, N., Tans, S.J., and Driessen, A.J. (2010). SecB--a chaperone dedicated to protein translocation. Mol. Biosyst *6,* 620-627.

Beveridge, T.J., and Davies, J.A. (1983). Cellular responses of Bacillus subtilis and Escherichia coli to the Gram stain. J. Bacteriol. *156,* 846-858.

Blow, D.M. (2004; 2002). Outline of crystallography for biologists (Oxford ; New York: Oxford University Press).

Bolhuis, A., Matzen, A., Hyyrylainen, H.L., Kontinen, V.P., Meima, R., Chapuis, J., Venema, G., Bron, S., Freudl, R., and van Dijl, J.M. (1999). Signal peptide peptidase- and ClpP-like proteins of Bacillus subtilis required for efficient translocation and processing of secretory proteins. J. Biol. Chem. *274,* 24585-24592.

Bos, M.P., Robert, V., and Tommassen, J. (2007). Biogenesis of the gram-negative bacterial outer membrane. Annu. Rev. Microbiol. *61,* 191-214.

Botos, I., Melnikov, E.E., Cherry, S., Tropea, J.E., Khalatova, A.G., Rasulova, F., Dauter, Z., Maurizi, M.R., Rotanova, T.V., Wlodawer, A., and Gustchina, A. (2004). The catalytic domain of Escherichia coli Lon protease has a unique fold and a Ser-Lys dyad in the active site. J. Biol. Chem. *279,* 8140-8148.

Braun, V. (1975). Covalent lipoprotein from the outer membrane of Escherichia coli. Biochim. Biophys. Acta *415,* 335-377.

Briggs, M.S., Cornell, D.G., Dluhy, R.A., and Gierasch, L.M. (1986). Conformations of signal peptides induced by lipids suggest initial steps in protein export. Science *233,* 206-208.

Burckhardt, S.E., Woodgate, R., Scheuermann, R.H., and Echols, H. (1988). UmuD mutagenesis protein of Escherichia coli: overproduction, purification, and cleavage by RecA. Proc. Natl. Acad. Sci. U. S. A. *85,* 1811-1815.

Butler, S.M., Festa, R.A., Pearce, M.J., and Darwin, K.H. (2006). Self-compartmentalized bacterial proteases and pathogenesis. Mol. Microbiol. *60,* 553-562.

Chayen, N.E. (1997). A novel technique to control the rate of vapour diffusion, giving larger protein crystals. J. Appl. Cryst. *30,* 198-202.

Chen, L., Tai, P.C., Briggs, M.S., and Gierasch, L.M. (1987). Protein translocation into Escherichia coli membrane vesicles is inhibited by functional synthetic signal peptides. J. Biol. Chem. *262,* 1427-1429.

Chen, Y.J., and Inouye, M. (2008). The intramolecular chaperone-mediated protein folding. Current Opinion in Structural Biology *18,* 765-770.

Choo, K.H., Tan, T.W., and Ranganathan, S. (2005). SPdb--a signal peptide database. BMC Bioinformatics *6,* 249.

Copeland, R. (2000). Enzymes.*2nd edition,*

Cristobal, S., de Gier, J.W., Nielsen, H., and von Heijne, G. (1999). Competition between Sec- and TAT-dependent protein translocation in Escherichia coli. EMBO J. *18,* 2982-2990.

Dalbey, R.E., Wang, P., and van Dijl, J.M. (2012). Membrane proteases in the bacterial protein secretion and quality control pathway. Microbiol. Mol. Biol. Rev. *76,* 311-330.

Date, T., and Wickner, W. (1981). Isolation of the Escherichia coli leader peptidase gene and effects of leader peptidase overproduction in vivo. Proc. Natl. Acad. Sci. U. S. A. *78,* 6106-6110.

Degering, C., Eggert, T., Puls, M., Bongaerts, J., Evers, S., Maurer, K.H., and Jaeger, K.E. (2010). Optimization of protease secretion in Bacillus subtilis and Bacillus licheniformis by screening of homologous and heterologous signal peptides. Appl. Environ. Microbiol. *76,* 6370-6376.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA. USA

Driessen, A.J., and Nouwen, N. (2008). Protein translocation across the bacterial cytoplasmic membrane. Annu. Rev. Biochem. *77,* 643-667.

Ekici, O.D., Paetzel, M., and Dalbey, R.E. (2008). Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration. Protein Sci. *17,* 2023-2037.

Emr, S.D., and Silhavy, T.J. (1983). Importance of secondary structure in the signal sequence for protein secretion. Proc. Natl. Acad. Sci. U. S. A. *80,* 4599-4603.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallographica *60,* 2126-2132.

Fairman, J.W., Noinaj, N., and Buchanan, S.K. (2011). The structural biology of beta-barrel membrane proteins: a summary of recent reports. Curr. Opin. Struct. Biol. *21,* 523-531.

Feldman, A.R., Lee, J., Delmas, B., and Paetzel, M. (2006). Crystal structure of a novel viral protease with a serine/lysine catalytic dyad mechanism. J. Mol. Biol. *358,* 1378-1389.

Fikes, J.D., Barkocy-Gallagher, G.A., Klapper, D.G., and Bassford, P.J.,Jr. (1990). Maturation of Escherichia coli maltose-binding protein by signal peptidase I in vivo. Sequence requirements for efficient processing and demonstration of an alternate cleavage site. J. Biol. Chem. *265,* 3417-3423.

Gan, L., Chen, S., and Jensen, G.J. (2008). Molecular organization of Gram-negative peptidoglycan. Proc. Natl. Acad. Sci. U. S. A. *105,* 18953-18957.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). Protein identification and Analysis Tools on the ExPASy Server. John M. Walker (Ed): The Proteomics Protocols Handbook, Humana Press 571-607.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics (Oxford, England) *19,* 163-164.

Gohlke, U., Pullan, L., McDevitt, C.A., Porcelli, I., de Leeuw, E., Palmer, T., Saibil, H.R., and Berks, B.C. (2005). The TatA component of the twin-arginine protein transport system forms channel complexes of variable diameter. Proc. Natl. Acad. Sci. U. S. A. *102,* 10482-10486.

Gouet, P., Courcelle, E., Stuart, D.I., and Metoz, F. (1999). ESPript: multiple sequence alignments in PostScript. Bioinformatics, *14,* 305-308.

Gur, E., Vishkautzan, M., and Sauer, R.T. (2012). Protein unfolding and degradation by the AAA+ Lon protease. Protein Sci. *21,* 268-278.

Hussain, M., Ichihara, S., and Mizushima, S. (1982a). Mechanism of signal peptide cleavage in the biosynthesis of the major lipoprotein of the Escherichia coli outer membrane. J. Biol. Chem. *257,* 5177-5182.

Hussain, M., Ozawa, Y., Ichihara, S., and Mizushima, S. (1982b). Signal peptide digestion in Escherichia coli. Effect of protease inhibitors on hydrolysis of the cleaved signal peptide of the major outer-membrane lipoprotein. European Journal of Biochemistry / FEBS *129,* 233-239.

Hutchinson, E.G., and Thornton, J.M. (1996). PROMOTIF--a program to identify and analyze structural motifs in proteins. Protein Sci *5,* 212-220.

Ichihara, S., Suzuki, T., Suzuki, M., and Mizushima, S. (1986). Molecular cloning and sequencing of the sppA gene and characterization of the encoded protease IV, a signal peptide peptidase, of Escherichia coli. J. Biol. Chem *261,* 9405-9411.

Iwanczyk, J., Damjanovic, D., Kooistra, J., Leong, V., Jomaa, A., Ghirlando, R., and Ortega, J. (2007). Role of the PDZ domains in Escherichia coli DegP protein. J. Bacteriol. *189,* 3176-3186.

James, G.T. (1978). Inactivation of the protease inhibitor phenylmethylsulfonyl fluoride in buffers. Anal. Biochem. *86,* 574-579.

Jansen, E.F., and Balls, A.K. (1952). The inhibition of beta- and gamma-chymotrypsin and trypsin by diisopropyl fluorophosphate. J. Biol. Chem. *194,* 721-727.

Jomaa, A., Iwanczyk, J., Tran, J., and Ortega, J. (2009). Characterization of the autocleavage process of the Escherichia coli HtrA protein: implications for its physiological role. J. Bacteriol. *191,* 1924-1932.

Kantardjieff, K.A., and Rupp, B. (2003). Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. Protein Sci *12,* 1865-1871.

Khan, A.R., and James, M.N. (1998). Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes. Protein Sci. *7,* 815-836.

Kim, A.C., Oliver, D.C., and Paetzel, M. (2008). Crystal structure of a bacterial signal Peptide peptidase. J. Mol. Biol. *376,* 352-366.

Kim, K.H., Aulakh, S., and Paetzel, M. (2011). Crystal structure of beta-barrel assembly machinery BamCD protein complex. J. Biol. Chem. *286,* 39116-39121.

Kim, K.H., and Paetzel, M. (2011). Crystal structure of Escherichia coli BamB, a lipoprotein component of the beta-barrel assembly machinery complex. J. Mol. Biol. *406,* 667-678.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. *372,* 774-797.

Krojer, T., Garrido-Franco, M., Huber, R., Ehrmann, M., and Clausen, T. (2002). Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. Nature *416,* 455-459.

Kumamoto, C.A. (1989). Escherichia coli SecB protein associates with exported protein precursors in vivo. Proc. Natl. Acad. Sci. U. S. A. *86,* 5320-5324.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Research *33,* W299-302.

Langklotz, S., Baumann, U., and Narberhaus, F. (2012). Structure and function of the bacterial AAA protease FtsH. Biochim. Biophys. Acta *1823,* 40-48.

Laskowski, R.A. (2001). PDBsum: summaries and analyses of PDB structures. Nucleic Acids Research *29,* 221-222.

Liang, J., Edelsbrunner, H., and Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci *7,* 1884-1897.

Little, J.W. (1984). Autodigestion of lexA and phage lambda repressors. Proc. Natl. Acad. Sci. U. S. A. *81,* 1375-1379.

Liu, J., and Mushegian, A. (2004). Displacements of prohead protease genes in the late operons of double-stranded-DNA bacteriophages. J. Bacteriol. *186,* 4369-4375.

Luo, Y., Pfuetzner, R.A., Mosimann, S., Paetzel, M., Frey, E.A., Cherney, M., Kim, B., Little, J.W., and Strynadka, N.C. (2001). Crystal structure of LexA: a conformational switch for regulation of self-cleavage. Cell *106,* 585-594.

Major, T., von Janowsky, B., Ruppert, T., Mogk, A., and Voos, W. (2006). Proteomic analysis of mitochondrial protein turnover: identification of novel substrate proteins of the matrix protease pim1. Mol. Cell. Biol. *26,* 762-776.

Matsumi, R., Atomi, H., and Imanaka, T. (2005). Biochemical properties of a putative signal peptide peptidase from the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1. Journal of Bacteriology *187,* 7072-7080.

McCoy, A.J., Grosse-Kunstleve, R.W., Storoni, L.C., and Read, R.J. (2005). Likelihood-enhanced fast translation functions. Acta Crystallographica *61,* 458-464.

Monteferrante, C.G., Baglieri, J., Robinson, C., and van Dijl, J.M. (2012). TatAc, the Third TatA Subunit of Bacillus subtilis, Can Form Active Twin-Arginine Translocases with the TatCd and TatCy Subunits. Appl. Environ. Microbiol. *78,* 4999-5001.

Muller, J., Walter, F., van Dijl, J.M., and Behnke, D. (1992). Suppression of the growth and export defects of an Escherichia coli secA(Ts) mutant by a gene cloned from Bacillus subtilis. Mol. Gen. Genet. *235,* 89-96.

Murshudov, G.N., Skubak, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F., and Vagin, A.A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallographica Section D *67,* 355-367.

Nagayama, M., Kuroda, K., and Ueda, M. (2012). Identification of interaction site of propeptide toward mature carboxypeptidase Y (mCPY) based on the similarity between propeptide and CPY inhibitor (IC). Biosci. Biotechnol. Biochem. *76,* 153-156.

Nikaido, H. (2003). Molecular basis of bacterial outer membrane permeability revisited. Microbiol. Mol. Biol. Rev. *67,* 593-656.

Novak, P., and Dev, I.K. (1988). Degradation of a signal peptide by protease IV and oligopeptidase A. Journal of Bacteriology *170,* 5067-5075.

Otwinowski, Z., Denzo. In: Sawyer, N., Isaacs, N., Baily, S., and Editors. (1993). Denzo, Daresbury Laboratory, Warrington, UK.56-62.

Pacaud, M. (1982). Purification and characterization of two novel proteolytic enzymes in membranes of Escherichia coli. Protease IV and protease V. The Journal of Biological Chemistry *257,* 4333-4339.

Paetzel, M., and Dalbey, R.E. (1997). Catalytic hydroxyl/amine dyads within serine proteases. Trends Biochem. Sci. *22,* 28-31.

Paetzel, M., Dalbey, R.E., and Strynadka, N.C. (2002a). Crystal structure of a bacterial signal peptidase apoenzyme: implications for signal peptide binding and the Ser-Lys dyad mechanism. J. Biol. Chem. *277,* 9512-9519.

Paetzel, M., Karla, A., Strynadka, N.C., and Dalbey, R.E. (2002b). Signal peptidases. Chem Rev. *102,* 4549-4580.

Paetzel, M., and Strynadka, N.C. (1999). Common protein architecture and binding sites in proteases utilizing a Ser/Lys dyad mechanism. Protein Sci. *8,* 2533-2536.

Paetzel, M., Strynadka, N.C., Tschantz, W.R., Casareno, R., Bullinger, P.R., and Dalbey, R.E. (1997). Use of site-directed chemical modification to study an essential lysine in Escherichia coli leader peptidase. J. Biol. Chem. *272,* 9994-10003.

Painter, J., and Merritt, E.A. (2006). Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta Crystallographica *62,* 439-450.

Raetz, C.R., and Dowhan, W. (1990). Biosynthesis and function of phospholipids in Escherichia coli. J. Biol. Chem. *265,* 1235-1238.

Raetz, C.R., and Whitfield, C. (2002). Lipopolysaccharide endotoxins. Annu. Rev. Biochem. *71,* 635-700.

Raju, R.M., Goldberg, A.L., and Rubin, E.J. (2012). Bacterial proteolytic complexes as therapeutic targets. Nat. Rev. Drug Discov. *11,* 777-789.

Rawlings, N.D., and Barrett, A.J. (2000). MEROPS: the peptidase database. Nucleic Acids Research *28,* 323-325.

Rawlings, N.D., and Barrett, A.J. (1999). MEROPS: the peptidase database. Nucleic Acids Research *27,* 325-331.

Régnier, P. (1981). The purification of protease IV of E.coli and the demonstration that it is an endoproteolytic enzyme. Biochemical and Biophysical Research Communications *99,* 1369-1376.

Remaut, H., Bompard-Gilles, C., Goffin, C., Frere, J.M., and Van Beeumen, J. (2001). Structure of the Bacillus subtilis D-aminopeptidase DppA reveals a novel self-compartmentalizing protease. Nat. Struct. Biol. *8,* 674-678.

Reynolds, C., Damerell, D., and Jones, S. (2009). ProtorP: a protein-protein interaction analysis server. Bioinformatics (Oxford, England) *25,* 413-414.

Rothballer, A., Tzvetkov, N., and Zwickl, P. (2007). Mutations in p97/VCP induce unfolding activity. FEBS Lett. *581,* 1197-1201.

Sankaran, K., and Wu, H.C. (1994). Lipid modification of bacterial prolipoprotein. Transfer of diacylglyceryl moiety from phosphatidylglycerol. J. Biol. Chem. *269,* 19701-19706.

Sauer, R.T., and Baker, T.A. (2011). AAA+ proteases: ATP-fueled machines of protein destruction. Annu. Rev. Biochem. *80,* 587-612.

Schallmey, M., Singh, A., and Ward, O.P. (2004). Developments in the use of Bacillus species for industrial production. Can. J. Microbiol. *50,* 1-17.

Silhavy, T.J., Kahne, D., and Walker, S. (2010). The bacterial cell envelope. Cold Spring Harb Perspect. Biol. *2:a000414,* 1-16.

Stein, N. (2008). CHAINSAW;: a program for mutating pdb files used as templates in molecular replacement. J. Appl. Cryst. *41,* 641-643.

Stewart, G.C. (2005). Taking shape: control of bacterial cell wall biosynthesis. Mol. Microbiol. *57,* 1177-1181.

Suno, R., Niwa, H., Tsuchiya, D., Zhang, X., Yoshida, M., and Morikawa, K. (2006). Structure of the whole cytosolic region of ATP-dependent protease FtsH. Mol. Cell *22,* 575-585.

Takagi, H., and Takahashi, M. (2003). A new approach for alteration of protease functions: pro-sequence engineering. Appl. Microbiol. Biotechnol. *63,* 1-9.

The_CCP4_Suite. (1994). Programs for Protein Crystallography. Acta Cryst. *D50,* 760-763.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research *22,* 4673-4680.

Tjalsma, H., Antelmann, H., Jongbloed, J.D., Braun, P.G., Darmon, E., Dorenbos, R., Dubois, J.Y., Westers, H., Zanen, G., Quax, W.J.*, et al.* (2004). Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome. Microbiol Mol Biol Rev *68,* 207-233.

Tjalsma, H., Bolhuis, A., Jongbloed, J.D., Bron, S., and van Dijl, J.M. (2000). Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. Microbiol Mol Biol Rev *64,* 515-547.

Tjalsma, H., Zanen, G., Venema, G., Bron, S., and van Dijl, J.M. (1999). The potential active site of the lipoprotein-specific (type II) signal peptidase of Bacillus subtilis. J. Biol. Chem. *274,* 28191-28197.

Tokunaga, M., Tokunaga, H., and Wu, H.C. (1982). Post-translational modification and processing of Escherichia coli prolipoprotein in vitro. Proc. Natl. Acad. Sci. U. S. A. *79,* 2255-2259.

Tomoyasu, T., Yuki, T., Morimura, S., Mori, H., Yamanaka, K., Niki, H., Hiraga, S., and Ogura, T. (1993). The Escherichia coli FtsH protein is a prokaryotic member of a protein family of putative ATPases involved in membrane functions, cell cycle control, and gene expression. J. Bacteriol. *175,* 1344-1351.

Tschantz, W.R., and Dalbey, R.E. (1994). Bacterial leader peptidase 1. Methods Enzymol. *244,* 285-301.

van der Wal, F.J., Oudega, B., Kater, M.M., ten Hagen-Jongman, C.M., de Graaf, F.K., and Luirink, J. (1992). The stable BRP signal peptide causes lethality but is unable to provoke the translocation of cloacin DF13 across the cytoplasmic membrane of Escherichia coli. Mol. Microbiol. *6,* 2309-2318.

van Roosmalen, M.L., Geukens, N., Jongbloed, J.D., Tjalsma, H., Dubois, J.Y., Bron, S., van Dijl, J.M., and Anne, J. (2004). Type I signal peptidases of Gram-positive bacteria. Biochimica Et Biophysica Acta *1694,* 279-297.

Vidal-Ingigliardi, D., Lewenza, S., and Buddelmeijer, N. (2007). Identification of essential residues in apolipoprotein N-acyl transferase, a member of the CN hydrolase family. J. Bacteriol. *189,* 4456-4464.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J. Mol. Biol. *225,* 487-494.

von Heijne, G. (1990). The signal peptide. J. Membr. Biol. *115,* 195-201.

Wang, P., Shim, E., Cravatt, B., Jacobsen, R., Schoeniger, J., Kim, A.C., Paetzel, M., and Dalbey, R.E. (2008). Escherichia coli signal peptide peptidase A is a serine-lysine protease with a lysine recruited to the nonconserved amino-terminal domain in the S49 protease family. Biochemistry *47,* 6361-6369.

Watson, M.E. (1984). Compilation of published signal sequences. Nucleic Acids Res. *12,* 5145-5164.

Wetzel, C.M., Harmacek, L.D., Yuan, L.H., Wopereis, J.L., Chubb, R., and Turini, P. (2009). Loss of chloroplast protease SPPA function alters high light acclimation processes in Arabidopsis thaliana L. (Heynh.). J. Exp. Bot. *60,* 1715-1727.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G., McCoy, A.*, et al.* (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. *67,* 235-242.

Winn, M.D., Isupov, M.N., and Murshudov, G.N. (2001). Use of TLS parameters to model anisotropic displacements in macromolecular refinement. Acta Crystallographica *57,* 122-133.

Winn, M.D., Murshudov, G.N., and Papiz, M.Z. (2003). Macromolecular TLS refinement in REFMAC at moderate resolutions. Methods in Enzymology *374,* 300-321.

Yokoyama, H., Matsui, E., Akiba, T., Harata, K., and Matsui, I. (2006). Molecular structure of a novel membrane protease specific for a stomatin homolog from the hyperthermophilic archaeon Pyrococcus horikoshii. J. Mol. Biol. *358,* 1152-1164.

Yuan, J., Zweers, J.C., van Dijl, J.M., and Dalbey, R.E. (2010). Protein transport across and into cell membranes in bacteria and archaea. Cell Mol. Life Sci. *67,* 179-199.

Zimm, B.H. (1948). The Scattering of Light and the Radial Distribution Function. J. Chem. Phys. *16,* 1093-1099.

Zweers, J.C., Barak, I., Becher, D., Driessen, A.J., Hecker, M., Kontinen, V.P., Saller, M.J., Vavrova, L., and van Dijl, J.M. (2008). Towards the development of Bacillus subtilis as a cell factory for membrane proteins and protein complexes. Microb. Cell. Fact. *7:10,* 1-20.

Appendices

Appendix A. $SppA_{BS}$ C-terminal peptide refining procedure

The large continuous positive difference electron density (F_o - F_c) observed in the substrate binding groove of $SppA_{BS}$ resembled the shape of a peptide (Figure A.1.A). The program Coot (Emsley and Cowtan, 2004) was used to build a poly-alanine peptide, 10 residues long, that fit nicely into the positive difference electron density. The peptide was built within the binding groove in an extended (anti-parallel) β–sheet fashion. The structure was then refined using restrained refinement with the program Refmac5 (Murshudov et al., 2011; Winn et al., 2003). The resulting polypeptide model fit such that the mainchain carbonyl oxygen of the 6th residue pointed directly into the oxyanion hole. After refinement, $2F_o$ - F_c electron density was observed surrounding the polypeptide and there was positive difference electron densities (F_o - F_c) for the side chains of each residue of the peptide (Figure A.1.B). The poly-alanine peptide was then mutated to match different regions of the C-terminal region of $SppA_{BS}$ that was the most abundant species in the mass spectrometry analysis, ^{307}FKSEIDFLNMREILSQSGSPRMMYLYAK^{335}. The positive electron density located at the residues 3rd from the end and 5th from the end of the poly-alanine peptide resembled an aromatic side chain amino acid. The C-terminus sequence ^{331}YLY^{333} was found to be a good candidate for these residues and therefore would correspond to the P1 and P2' position, since Tyr331 is closest to the serine nucleophile. After multiple cycles of restrain refinement with different C-terminal sequences, the last ten residues of $SppA_{BS}$, ^{326}SPRMMYLYAK^{335} was found to best fit the electron density (Figure A.1.D). An example of a different side chain that was attempted to fit the difference electron density map but didn't fit well was when residue 331 was mutated to lysine and refined. The resulting F_o - F_c map showed that the length of the density matches lysine however, the side of the density was fatter, plus there was extra positive electron density, which resembles the electron density for an aromatic side chain such as tyrosine (Figure A.1.C).

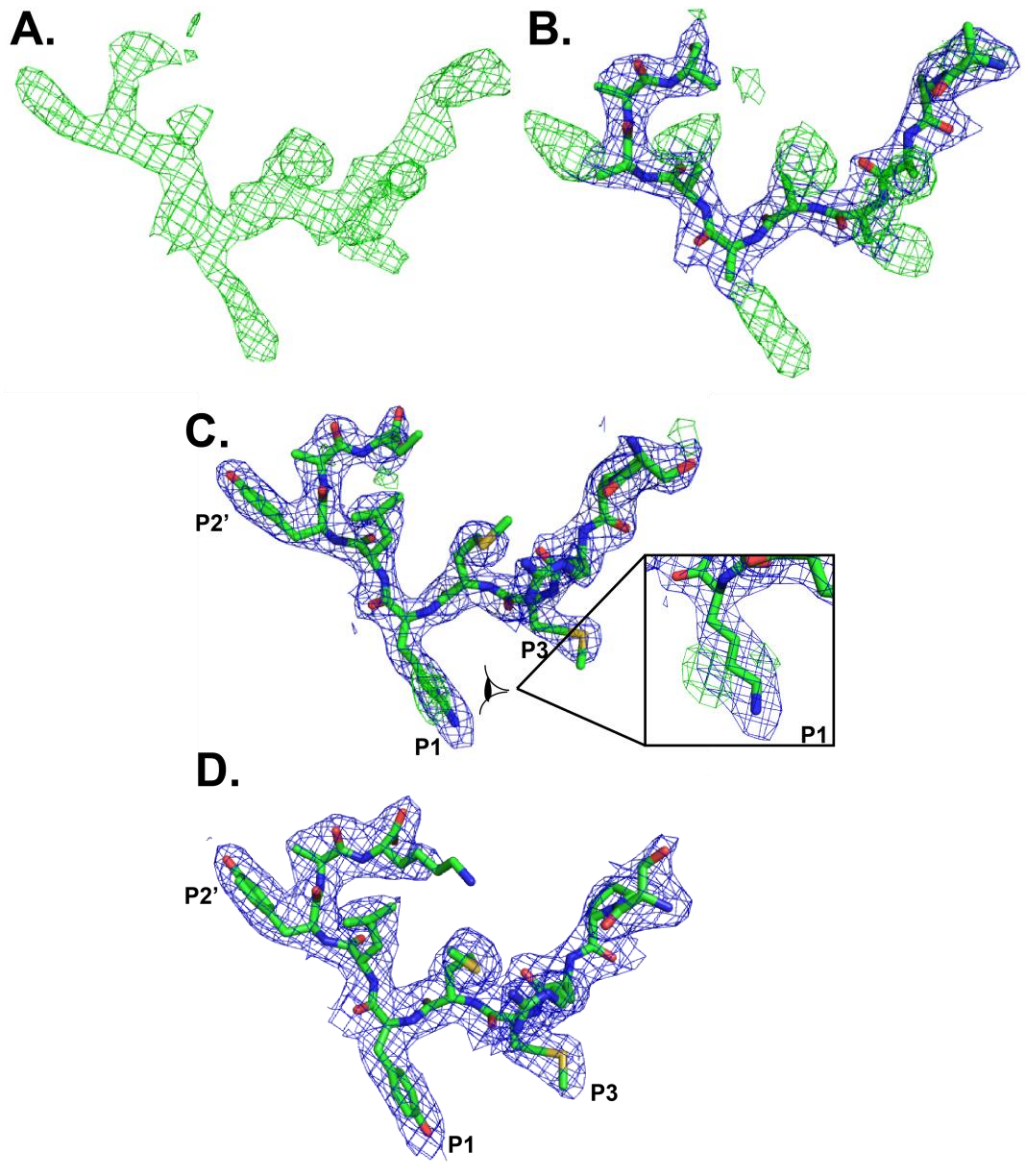


Figure A. 1. The electron density map and the C-terminal peptide building process.

A. F_o - F_c positive difference electron density map shown in green mesh. The positive difference electron density is observed in each of the eight substrate binding groove of $SppA_{BS}$. This results in a perfect continuous ring of electron density for the complete $SppA_{BS}$ octamer. **B.** A poly-alanine peptide (shown in stick model with oxygen in red, nitrogen in blue, and carbon in green) is fit into the F_o - F_c positive difference electron density map and refined, submitted. The resulting $2F_o$ - F_c electron density map is shown in blue mesh, the resulting F_o - F_c positive difference density is shown in green mesh. **C.** A peptide from the C-terminus of $SppA_{BS}$, ^{326}SPRMMYLYAK^{335}, with Tyr331 mutated to lysine, is fit into the $2F_o$ - F_c electron density map (blue mesh). The zoomed in view in the direction the eye looks shows the electron density surrounding the modeled lysine side chain. Clear positive F_o - F_c difference elctron density can be seen on both sides of the lysine side chain. **D.** $SppA_{BS}$ C-terminal peptide, ^{326}SPRMMYLYAK^{335}, is shown with surrounding $2F_o$ - F_c map. For all panels, the F_o-F_c electron density is contoured at 3σ level and the $2F_o$ - F_c electron density map is contoured at 1σ level.