Sequence and Structural Investigation of the Nonribosomal Peptide Synthetases of *Bacillus atrophaeus* UCMB 5137(63Z)

A mini-thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE OF RHODES UNIVERSITY

by

Coursework / Thesis

in

Bioinformatics and Computational Molecular Biology

In the Department Of Biochemistry, Microbiology & Biotechnology

Faculty of Science

by

Candice Ryan

February 2013

ABSTRACT

Due to increased plant resistance to the existing antibiotics produced, there is a need to develop alternatives. Nonribosomal peptides (NRPs) are important plant phytopathogens synthesized by nonribosomal peptide synthetases (NRPSs). In this study, a newly sequenced *Bacillus* strain *Bacillus atrophaeus* UCMB 5137 (63Z), found to have increased phytopathogenic activity, was investigated to gain insights to the possible reason behind this activity.

NRPS modules were identified using a novel script that can act on unannotated, raw DNA sequences. The Structure Based Sequence Analysis Webserver was used to identify the amino acids incorporated into the final NRP, which were compared to the NRP database. Five NRPSs were found within the strain; fengycin/ plipstatin, mycosubtilin, surfactin, bacillibactin and bacitracin. Some of the modules usually present for these NRPSs were not present in the test strain and only a few modules were found. A phylogenetic study was carried out and the topologies of the trees showed that genes were not transferred horizontally. It did, however, lead to the hypothesis that different NRPS genes are under different adaptive evolutionary pressures.

Only slight conformational changes between L and D-conformation of amino acids were seen between the test and neighboring strains. All of the linker and terminal regions of synthetases were found to exhibit a large amount of conservation overall.

Homology modeling was performed on the test strain on selected modules, TE and A-domains of fengycin and mycosubtilin synthetases. TE-domains between the different synthetases are different and specific for the NRP they facilitate release for. The NRPS from which the A-domain originates also influences substrate specificity as well as the module in which the A-domain occurs within the NRPS. Binding pockets of A-domains of differing substrate specificity were compared. Future work will include; refinement of the models and docking studies within the A-domain binding pocket.

DECLARATION

I declare that the dissertation, which I hereby submit for the degree MSc - Bioinformatics at Rhodes University, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

Date: February 2013

ACKNOWLEDGEMENTS

Firstly I would like to express my sincere gratitude to my advisor Dr Özlem Tastan Bishop for the continuous support of my MSc study and research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis as well as her pain-staking effort in proof reading the drafts. I could not have imagined having a better advisor and mentor for my MSc study. I would also like to thank my co-supervisor, Prof Oleg Reva, for his insightful comments and hard questions.

I would like to thank my colleagues in RUBi for their advice and help. Especially to Crystal Clitheroe and Rowan Hatherley who were always available to give advice and listen to my rants.

I would like to thank my family for their love and support during all my studies over all the years and leading up to this point. For teaching me to work hard for the goals I wanted to achieve and allowing me to function in my little bubble.

Especially to my fiancé, Chris Rafael, for his love, support and belief in me and for dealing with all the mood swings associated with my work and writing process. For being my support system and helping me to survive and stay sane. For without you I simply would not have finished this thesis. I dedicate this thesis to Chris with all my love.

I would like to thank Rhodes University and the Sandisa Imbewu Scholarship for funding this work. I would like to thank the South African Genetics Society (SAGS) and the South African Society for Bioinformatics (SASBi) for granting me a travel fellowship to attend and present this work at the joint meeting of SAGS and SASBi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Asn	Asparagine
B.	Bacillus
Bcb	Bacillibactin synthetase
Bct	Bacitracin synthetase
Feng	Fengycin synthetase
Glu	Glutamine
Myc	Mycosubtilin synthetase
NRP	Non-ribosomal peptide
NRPDB	NRP database
NRPS	Non-ribosomal peptide synthetase
Phe	Phenylalanine
PKS	Polyketide synthetase
Plp	Plipstatin synthetase
Pro	Proline
SBSPKS	Structure Based Sequence Analysis of Polyketide synthetases
Srf	Surfactin synthetase
Tyr	Tyrosine

FIRST CHAPTER: LITERATURE REVIEW

1.1 Overview

This chapter introduces the main concepts necessary for this research and gives an overview of the literature available on the subject. The chapter will first outline the background literature and then offers more details on the various sections. Once the available knowledge is discussed a problem statement is posed to show the motivation for the project. The aims and objectives of the research are then presented.

1.2 Introduction

The promotion of plant growth and the control of plant disease are a pressing need for the 21st century all over the world. This need is driven by an increasing human population, and there is a demand, particularly in the developed countries, for high quality food that is free from unacceptable levels of chemicals such as herbicides and pesticides (Nagórska et al., 2007; Ongena & Jacques, 2008). There is clearly a need to develop new agricultural practices that supplement and, ultimately, even replace existing disease control strategies. One such strategy that has not yet been fully exploited, but which has the potential to supplement and possibly even replace the use of synthetic chemical pesticides, is the use of biocontrol agents (Ongena & Jacques, 2008). Biocontrol agents is a term often used to describe microbes which are able to produce biologically active compounds that poses the ability to inhibit phytopathogen growth while promoting the growth of their host organism or plant (Ongena & Jacques, 2008).

Microorganisms have become a promising alternative to pesticides as biological control agents. Microorganisms are a preferred choice to pesticides as they are less expensive and have shown to be less accumulative in plants and soil and

thereby having less adverse effects on humans than pesticides do (Nagórska et al., 2007; Ongena & Jacques, 2008). The emergence of pesticide and fungicide resistance has also lead to the urge to develop alternatives. Non-pathogenic bacteria, such as rhizobacteria have been found to coexist with plants as well as enhance the plants adaptive potential and growth (Nagórska *et al.,* 2007). Rhizobacteria have been found to have many favorable biological control agent characteristics including; suppression of phytopathogen disease, activation of the hosts defenses and increased uptake of nutrients such as nitrogen, from the soil (Nagórska *et al.,* 2007). *Bacillus subtilis* is a rhizobacterium that has been well studied and used as a biocontrol agent. *Bacillus* species as a model organism is discussed in more detail in Section 1.3. One of the reasons for this is that they are able to produce lipopeptides, which have antimicrobial properties (Ongena & Jacques, 2008). The method in which the mutualistic relationship occurs between *Bacillus* strains and the plant rhizospheres is still not fully understood but is speculated to be linked to the production of lipopeptides and sideophores or antibiotics, known as non-ribosomal peptides (NRPs) produced by non-ribosomal peptide synthetases (NRPSs) (Finking and Marahiel, 2004; Koumoutsi *et al.,* 2004).

NRPSs are large enzyme complexes that are made up of several modules, which together regulate the production of the NRP end product. The structure of the NRP is determined by the organization and number of modules within the NRPS. Each module is made up of several domains (Figure 1.1) that together facilitate the production of the NRP each with their own responsibilities within the complex discussed further in Section 1.4. NRPs are short peptides, which have many clinical applications due to their lipopeptide or sideophore possessing properties. NRPs are discussed below in Section 1.5 in more detail.

Due to the fact that NRPS are assembled using modules, and the order and number of modules and domains determines the end NRP product, research is focused at the discovery of novel NRPS that will lead to the production of "unnatural natural products" through shuffling the enzymatic domains (Nguyen *et al.,* 2006; Weissman and Leadlay, 2005). The construction of chimeric enzymes

has been an area of interest with NRPSs due to their organization. This construction would allow for the exchange of intact modules or single domains and thereby lead to the synthesis of novel non-ribosomal peptides. It has been found that the intact C domain – A domain interface could be integral to the catalytic process of this step (Tanovic e*t al.,* 2008). Tanovic (2008) proposed that this would be possible through the exchange of C domain – A domain pairs.

The strain studied in this research is a member of the *Bacillus subtilis* group and has only recently been sequenced. Due to this, there is no literature published on this particular strain as of yet. Our collaborators at the University of Pretoria have, however, in molecular studies have shown that this particular strain, *B. atrophaeus* UCMB 5137 (63Z), is much stronger than other tested similar strains and is able to inhibit the growth of *Leptosphaeria maculans*, a potent phytopathogen, when the plant is grown in conjunction with this strain (Personal correspondence). This discovery has made the prospect of using this strain for its antibiotic properties very favorable and more research would be needed on this strain to take this next step. The test strain has not previously been annotated or investigated specifically for its NRPS contents on a sequence level.

There are crystal and NMR structures available within the Protein Data Bank (PDB) for some of the NRPS domains (refer to Chapter 3 Section 2, Table 3.1). There are, however, still many, which are absent, and there is no structural information available for this particular strain's NRPS domains as well as for the neighboring, more well studied strains used in this study.

Linker regions between the modules and the domains have been found to play a role in the production and correct functioning of the end NRP produced and their disruption has been shown to disrupt the NRP produced (Lai *et al.,* 2006). The exact mechanism as to how the linker regions affect and regulate the modules and facilitate the biosynthetic transfer of the correct amino acid to the specific module is still not known. Linker regions are discussed in Section 1.4 and Section 2.5.6.

1.3 Model organism:

Model organisms are those within the *B. subtilis* group, which includes strains from the following species; *B subtilis, B. amyloliquifaciens and B. atrophaeus.* Members from this group have already been widely used as a biological control agents (Nagórska *et al.,* 2007). *B. subtilis* has a broad host range and is able to maintain a stable relationship with its host and promote its growth. The endospores produced by *B. subtilis* are also able to produce a wide range of broad spectrum antibiotics (Moszer, 1998) that are resistant to wide ranges of temperatures and pHs (Souto *et al.,* 2004).

The genomic organization of the *B. subtilis* group is very flexible and natural rearrangements occur frequently therein, thereby allowing for the natural selection of favorable compounds such as those able to add selective advantages to the host (Stein, 2005). This therefore leads to access of genetic manipulations by NRPSs, which could provide a means of development of novel antibiotics and therefore a possible path for the development of novel plant biocontrol agents (Sieber & Marahiel, 2003). Isihihara *et al.* (2002) speculated that the production of NRP antibiotics is strain-specific rather than species specific since different strains of the same species had shown to produce different antibiotics.

1.3.1. B. atropheaus UMCB 5137 (63Z):

The main strain which will be investigated in this study will be *B. atrophaeus* UMBC 5137 (63Z). *B. atrophaeus* is a new group made up of organisms which produced a brownish-black pigment on one medium and a brown pigment on the other, showed low levels of DNA hybridization with groups 2 and 3 (both representing *B. subtilis* species), this includes two strains previously called *B. subtilis* DSM 675 and DSM 2277 (Fritze & Pukall, 2001). Other strains which are used in the study for comparison and control porposes are; *B. atropheaus* 1942, *B. amyloliquifaciens* FZB42, *B. subtilis subspecies subtilis* 168 and *B. subtilis* BS5n.

1.3.2. B. subtilis subsp. subtilis 168:

When sequenced the *B. subtilis subsp. subtilis* 168 was found to contain the full sequences of the NRPSs encoding for surfactin (srf) and bacillibactin (bcb). On inspection by May *et al.,* (2001) it was found that the *sfp* gene contained a defect, which would result in the carrier protein domains staying constantly in their inactive/ apo form and therefore explaining why the strain does not produce either of the NRPs. This is a laboratory strain which is frequently used due to its ability to take up foreign/introduced DNA and be genetically manipulated (Barbe *et al*., 2009).

1.3.3. B. amyloliquiefaciens:

B. amyloliquefaciens is a Gram-positive bacterium known to promote plant growth as well as suppress plant pathogenic bacteria and fungi. Siezen and Khayatt (2008) found four more clusters in this bacteria, apart from the five known clusters of *B. subtilis* that are involved in mediating the biosynthesis of secondary metabolites such as; antibiotics and siderophores. The known gene clusters from *B. subtilis* are surfactin synthetase, plipstatin/ fengycin (plp/ feng) synthetase, bcb synthetase, bacilysin and bacillaene. The four additional clusters identified from *B. amyloliquefaciens* are bacillomycin D, macrolactin, difficidin and a putative siderophore (Siezen & Khayatt, 2008).

Koumoutsi *et al.,* (2004) found that *B. amyliolifaciens* FZB42 has 7.5% of its entire genome devoted to the production of PKs and NRPs that enable it to survive in the presence of competing organisms which is twice the amount of *B. subtilis* lab model organism 168 (Wipat & Harwood, 1999). Bacillomycin (*Bmy*) was identified by Koumoutsi *et al.,* (2004) has the main cluster involved in the synthesis of NRPs. They proved that the clusters were imperative to the biosynthesis of the lipopeptides by disrupting the gene and noting that the disruption resulted in absence of the lipopeptides production (Koumoutsi *et al.,* 2004).

1.4. Non-ribosomal peptide synthetases:

Non-ribosomal peptide synthetases (NRPSs) were described by Challis and Naismith (2004) as "complex molecular machines". NRPSs are composed of many modules that each contain several separately folded catalytic domains (Challis and Naismith, 2004). NRPSs are multidomain enzymes that catalyse the formation of secondary metabolites such as medicinally important antibiotics and promote plant growth (Finking and Marahiel, 2004, Strieker *et al.,* 2010). NRPS were discovered to be large proteins comprised of several fused repeating enzymes (Challis and Naismith, 2004).

1.4.1. Structure of NRPS:

Each NRPS is made up of several subunits each of which contains their own N and C-terminal regions. Each subunit is made up of a number of modules connected by linker regions. Each module is further divided into domains (e.g. catalytic, adenylation (A), thiolation (T), esterification (E) and thioesterase(TE)). Each domain is connected by another linker region. The A-domain is responsible for the specification of the amino acid transferred to the growing chain of the NRP being formed resulting in the end antibiotic product (Figure 1.1).

Modules:

Modules of NRPSs are made up of about 1000 residues each of which control a single reaction cycle adding one amino acid to the NRP chain (Ongena & Jacques, 2008a). Hillson and Walsh (2003) found evidence to suggest that the modules regulate activity by forming functional dimers which thereby allow synthesis to move in a chain from one protein to the next as well as within a chain (Hillson and Walsh, 2003).

Starcevic *et al.,* (2008) found that catalytic domains could be grouped into modules that control each successive round of chain elongation which implies that synthesis is due to a collinear principle in which the final chemical arrangement is reliant on the individual modules gene sequence. Domains related

to each module perform specific reactions that allow for the prediction of module specificity through domain specificity (Starcevic *et al.,* 2008).

Domain regions:

NRPSs contain three catalytic domains: the condensation (C) domain, responsible for amino acid condensation, the adenylation (A) domain, responsible for amino acid activation, and the peptidyl carrier (PCP) domain, responsible for proliferation. These domains are required for elongation of the peptide and are therefore ubiquitous in NRP synthesis. The thiolation (TE) domain is a catalytic domain involved with product release and is found in the termination module. Catalysis in this domain is catalyzed through either hydrolysis or macrocyclization (Strieker *et al*., 2010). A two-step reaction is carried out to activate each amino acid by its respective module by the appropriate coupled domain. The A (adenylation) domain selects the amino acid from the available substrate and ATP is then used to generate a corresponding aminoacyl adenylate. This is then covalently attached to the phosphopantheinyl (ppant) of the PCP domain. The ppant group is then added to the TE domains by ppant transferase and thereby post-translationally modifying them. The C (condensation) domain controls the addition of the peptidyl chain (Stachelhaus *et al.,* 1999).

Machinery responsible for the acquiring of an activated, covalently linked amino acid, ready to undergo peptide bond formation, is located within the A and PCP domains. The A domain activates the amino acid, generating ATP and a highly reactive aminoacyl adenylate, which then reacts with the thiol (located at the terminus of the phosphopantetheinyl arm). This is then attached, post-translationally, to a conserved serine residue within the PCP domain and forms an activated thioester derivative. The C domain is responsible for the catalysis of the peptide bond formation between the aminoacyl thioester attached to the PCP domain. The first initiation module does not have any C domain. The last module has an extra TE domain that cleaves the peptide from the PCP domain once the peptide is fully assembled. The C domain is sometimes substituted by a Cy domain, which catalyzes the condensation and intramolecular heterocyclisation of

serine, cysteine or threonine (Stachelhaus *et al.,* 1999; Challis and Naismith, 2004).

A large amount of conformational freedom has been found necessary within the domains of NRPSs to facilitate the synthesis of non-ribosomal peptides especially in the case of the PCP domain as it acts as an intermediate between the A domain and the C domain (Samel, *et al.,* 2007).

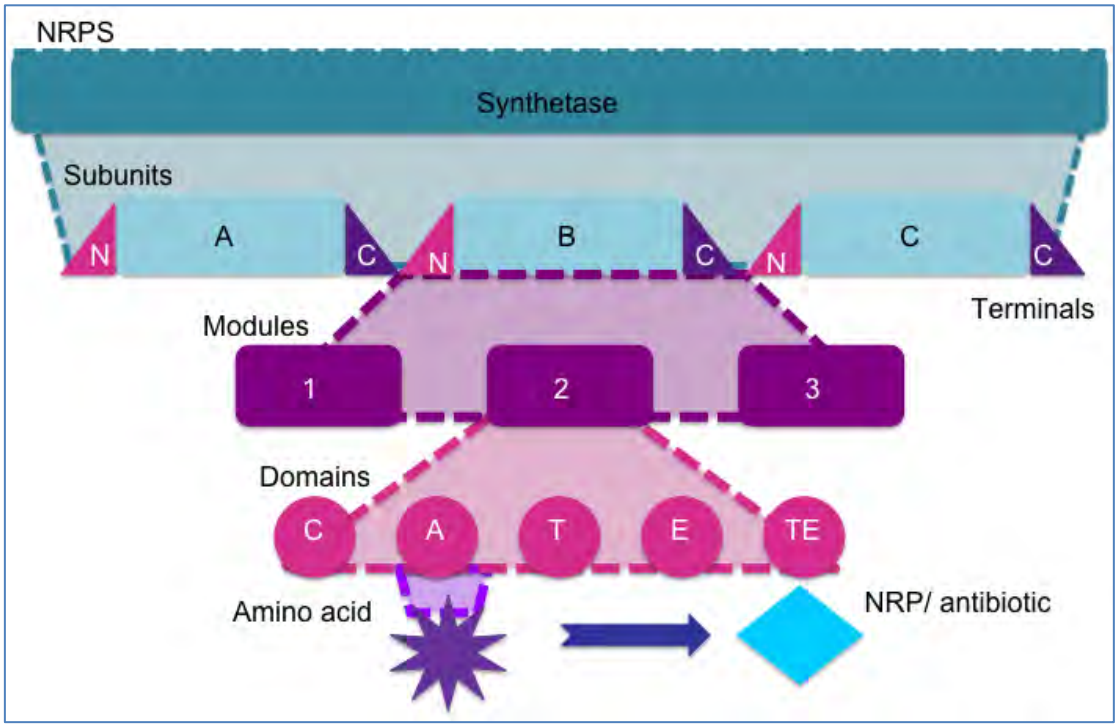


Figure 1.1: Schematic representation of the structure of NRPS units to NRP formation. Each NRPS is made up of several subunits each of which contains their own N and C-terminal regions. Each subunit is made up of a number of modules connected by linker regions. Each module is further divided into domains (e.g. catalytic, adenylation, thiolation, esterification and thioesterase). Each domain is connected by another linker region. The A-domain is responsible for the specification of the amino acid transferred to the growing chain of the NRP being formed resulting in the end antibiotic product

Peptidyl Carrier Protein Domain:

The Peptidyl Carrier Protein (PCP) domain can occur in 3 different conformations; A (apo), A/H, and H (holo) state. The most common state is the A/H state, however, during the priming reaction the ppant-primed PCP domain is generated through the exclusive interaction between srf synthetase-activating enzyme (Sfp) and the A-state (Figure 1.2). Mis-priming events interrupting NRP

assembly occur since approximately 80% of the ppant cofactor precursor (CoA) is acetylated in bacteria (Vallari *et al.*, 1987). The TE domain is involved in correcting the mis-primed PCP species by interacting with the H state of the PCP domain (Schwarzer *et al.,* 2002).

Adenylation (A)-Domain:

The A domain in its open conformation is able to bind to the amino acid and ATP forming an adenylate intermediate (speculated from the crystal structure of DhbE). The structure is closed due to the cleavage of the phosphodiester bond by the aminoacyl adenylate formation and release of a pyrophospahate. This is followed by the thioester-forming step wherein the activated amino acid is then transferred to the PCP. This product is then released from the PCP domain by condensation resulting in the A domain reverting to the open confirmation and the cycle is restarted (Strieker *et al.,* 2010).

Structurally similar amino acids have been found in equivalent positions of NRPs possibly indicating the NRPSs A domain have a relaxed substrate specificity thereby suggesting that novel natural products could be produced (Hoffmann et al., 2003). Christiansen *et al*., (2011) found that a substitution of an amino acid within the cyclic backbone does not effect the bioactivity of the peptide as opposed to those not contained within the cyclic backbone in which case a substitution results in a large effect on the peptides functionality/ activity (Christiansen et al., 2011). Novel lipopeptides have successfully been synthesized using srf synthetase through the rearrangement of modules (Stachelhaus et al., 1995).

Termination (TE-) Domain:

The TE-domain has a molecular weight of about 28kDa and is found on the C-terminal end of the NRPS module. A catalytic triad of Ser-His-Asp were found to be conserved in all TE domains and is responsible for the release of the final peptide (Grunewald and Marahiel, 2006).

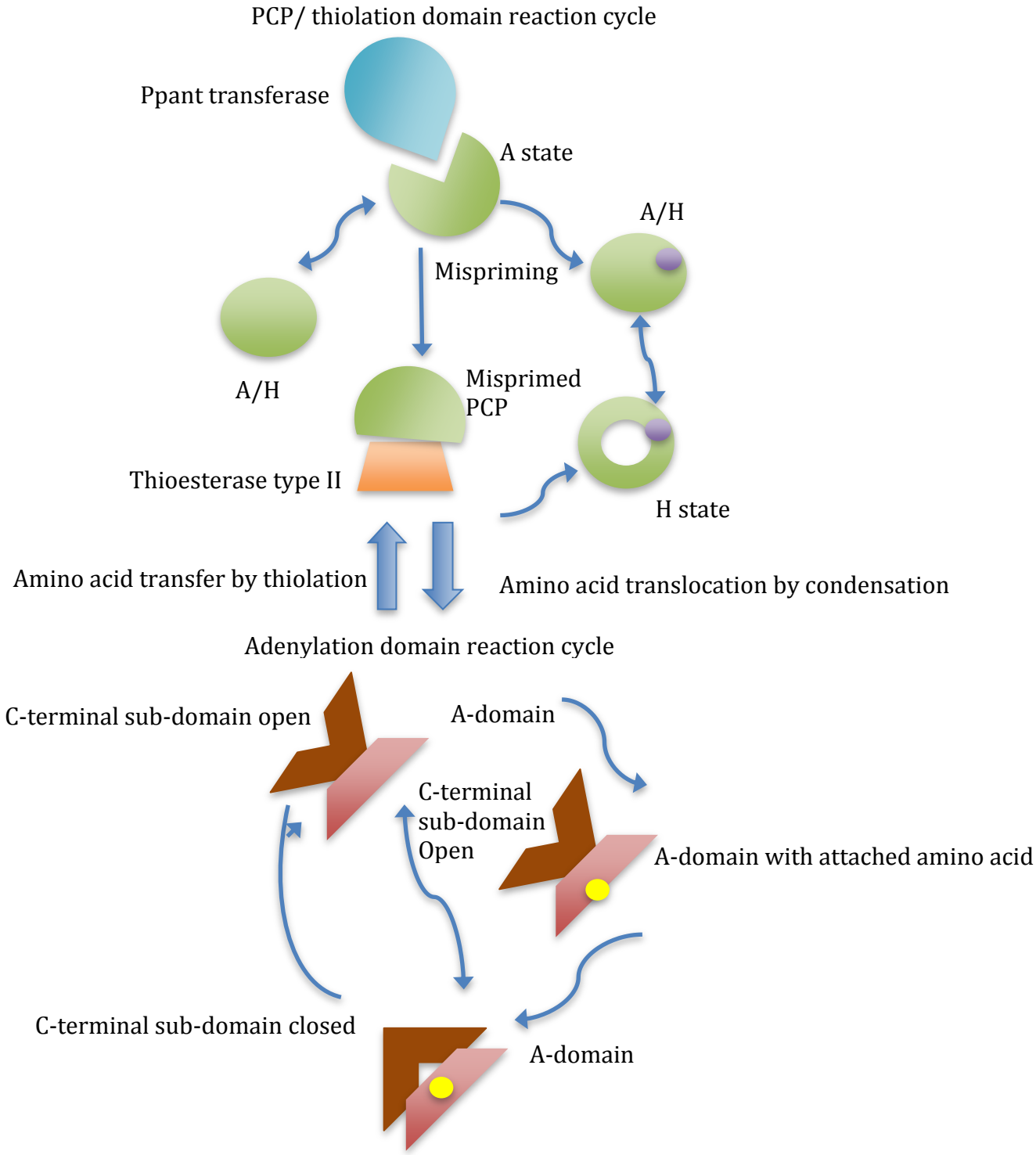


Figure 1.2: Schematic representation of the adenylation and PCP domain cycles.
The top half (green) shows the PCP-domain cycle wherein ppant-transferase converts the apo-form of the domain to the holo-form when attached to the active site serine (shown in purple). The bottom half (orange) shows the adenylation cycle which occurs during mispriming wherein the A-domain interacts with the holo-PCP facilitating the new amino acid to be loaded. Once the amino acid has been successfully transferred the holo-PCP can be loaded with another amino acid. The second cycle shows the c-terminal subdomain (brown) of the A-domain (red). The A-domain is able to bind the incoming amino acid and ATP when it is in the open conformation. When in the closed conformation the aminoacyl-AMP, formed during the previous step and release of pyrophosphate, is protected from additional solvent thereby allowing the amino acid to be transferred to the PCP-domain (Adapted from Strieker *et al*., 2010).

Linkers:

Linker regions are found between the modules as well as the domains of the NRPS (Figure 1.1). Linkers have been found to be involved in the communication between modules as well as to possess a regulatory role (Gokhale & Khosla, 2000). Gokhale and Khosla (2000) found linkers to be crucial in the structural and functional assembly of multi-domain proteins such as NRPSs through biosynthesis and signal transduction and linkers were found to direct the movement of the domains.

1.5. Non-ribosomal peptides:

Proteins are usually synthesized by ribosomes catalyzing peptide bond formation, but what is often overlooked, is that some peptides may be catalyzed by non-ribosomal peptide synthetases. These non-ribosomal synthesized products, NRPs, are usually cyclic and contain a high density of proteinogenic amino acids (Challis *et al.* 2004). These peptide products are 2-48 residues in length and there are over 300 amino acids, which can be incorporated due to the presence of D-configured, N-methylated, and other non-proteinogenic residues. NRPs can form linear, cyclic, or branched cyclic structures and can undergo many different forms of modification such as acylation, glycosylation, and heterocyclizaton (Du & Lou, 2010).

NRPs form a group of biologically important compounds with broad clinical applications such as last resort antibiotics, antitumor or antifungal drugs and immunosuppressant's as well as plant growth promotion (Walsh, 2008). NRPS and its associated enzymes are responsible for the formation of vancomycin, a last resort antibiotic, and its associated enzymes as well as most peptide-based antibiotics. NRPSs are also responsible for the production of bleomycin, an anti-tumor compound, and cyclosporine, an immune suppressor integral to organ transplantation. (Weber *et al.,* 2004). Several distinct modules, which each aid in the incorporation of a single monomer to the peptide, together constitute NRPSs (Strieker *et al.,* 2010). The NRPSs can be classified as either lipopeptides or sideophores.

1.5.1. Lipopeptides:

Bacillus species often synthesize lipopeptides, which are amphipathic, cyclic antibiotics. Lipopeptides have been divided into three groups (srf, iturin and feng) according to length, branching fatty acid chains and amino acid substitutions. Lipopeptides with longer hydrocarbon side chains have been found to be more bioactive (Toure *et al.*, 2004). Iturin has been found to form small vesicles of segregation membrane-spanning particles and release electrolytes and high molecular mass products which disrupt the plasma membrane of yeast and degrade phospholipids respectively (Toure *et al.,* 2004). Iturin and feng lipopetides have been found to display potential antifungal activity and to suppress many plant pathogens growth. Iturins have displayed antifungal and limited antibacterial activity (Maget-Dana and Peypoux, 1994; Souto *et al*., 2004). Several studies have highlighted the synergistic interactions between lipopeptides in different combinations of the three increasing their effects (Maget-Dana et al., 1992; Ongena et al., 2007; Romero et al., 2007). Ongena and Jacques (2007) speculated that *Bacillus* strains that are able to co-produce lipopeptides from each of the three groups would contain more powerful biocontrol characteristics since each of the different lipopeptides groups exhibit different strengths and levels of efficiency (Ongena & Jacques, 2008b).

Fengycin:

Feng is one of the lipopeptides reported to be produced by *B. subtilis*. Feng is a cyclic lipopeptide and has been found to inhibit phospholipase A2 of filamentous fungi (Nielsen & Sorensen, 2003). The structure of feng is comprised of a β-hydroxy fatty acid joined to a 10 amino acid peptide chain, 8 of which are in a cyclic conformation. There are two different isoforms of feng that were determined by Deleu *et al.,* (2005) in which isoform A has a D-Ala whereas isoform B has a D-Val in the same position. It was determined that the mechanism of action of feng is through interference with the phospholipid molecules of the cell membrane that are necessary for tight packaging thereby exhibiting a similar effect to that caused by cholesterol (Deleu, Paquot, & Nylander, 2005). Fengs

inhibit phytopathogen growth as well as acting as immune-stimulants and stimulate root colonization (Ongena & Jacques, 2008a).

Surfactin:

Srf is classified as a lipoheptapeptide which is made up of four different enzyme constituents (Menkhaus *et al.,* 1993). The structure of surfactin has also been determined to be cyclic in nature joined to a β–hydroxy fatty acid component. Srf has been found to effect membrane integrity and stability and thereby displays antimicrobial and antiviral properties (Peypoux *et al.*, 1999). Fengs and srfs increase and stimulate the immune response of the plant thereby increasing the resistance of the plant (Ongena & Jacques, 2008b).

Iturins:

Iturns have limited antibacterial activity as well as strong antifungal and hemolytic activities (Phae et al., 1990, Maget-Dana & Peypoux, 1994). Iturin lipopeptides are PKS-NRPS hybrids. This action is attributed to their membrane permeability properties. Mycosubtilin (myc) is an example of an iturin lipopeptide and is made up of four open reading frames/ subunits, three of which are responsible for the coding of the NRPS; *mycA. mycB* and *mycC* (Ongena & Jacques, 2008b). Duitman *et al.,* (1999) recorded the myc heptapeptide to be made up of the following amino acids; Asn-(D)Tyr-(D)Asn-Gln-Pro-(D)Ser-Asn and is reported to span 38 kb. The A subunit of myc (*mycA*) has functional domains from peptide synthetases, fatty-acid synthetases and amino transferases and has therefore been termed the "first example of a natural hybrid between these enzyme families" (Duitman *et al*., 1999). Leclere *et al.,* (2005) found that overproduction of myc could confer phytopathogen characteristics to a strain which otherwise does not usually produce biocontrol abilities. They also determined that an altered myc wherein threonine is substituted for serine exists indicating that myc synthetase adenylation domain is able to activate both residues (Leclère et al., 2005).

1.5.2. Siderophores:

Siderophores are low molecular weight iron chelating compounds synthesized by bacteria and they are essential for growth by most organisms (Ratledge & Dover, 2000). Disruption of siderophore production is a target for eliminating the growth of pathogeneic organisms since siderophores compete with other iron chelating agents, such as haem, for iron (Quadri, 2000).

Bacillibactin:

B. subtilis produces a cyclic catecholic siderophore, bcb, through the NRPS made up of the following enzymes; DhbB, DbhE and DbhF, encoded by the dhb operon (Rowland *et al.*, 1996) in response to iron deprivation (May et al., 2001).

1.6. NRP biosynthesis:

The overall biosynthesis of NRP by NRPSs involves 3 main steps (initiation, elongation and termination), each of which are carried out by separate NRPS modules (Figure 1.3) (Siezen & Khayatt, 2008). The differing specificity of the A domain, as well as occasional additional domains such as a methyltransferase domain, dictate the structural and compositional differences in the end peptide products (Walsh *et al.,* 2001).

Challis *et al*. (2000) and Stachelhaus *et al*. (1999) found that nine residues in the adenylation domain in the substrate-binding pocket are responsible for determining which amino acid is incorporated into the peptide by NRPSs. This knowledge made it possible to relate the amino acid to a corresponding specificity residue and thereby enabling the prediction of the NRPS adenylation domain specificities (Challis *et al*., 2000; Stachelhaus *et al.,* 1999). *B. amyloliquefaciens* FZB42 was found to have 8.5% of its chromosome devoted to NRPSs through the identification of these sites (Chen *et al.*, 2007).

Interactions between the domains are necessary to the functioning of NRPSs (Tsuji, *et al.,* 2001). During the biosynthesis of NRPs the intermediates are tethered to the carrier proteins and the determinants of the carrier-protein recognition by the other synthetase domains is an important area of study (Lai *et*

al., 2006). Fuma *et al.* (1993) found that the number of modules is directly related to the number of chain elongation steps that take place during NRPS biosynthesis. They also discovered that the domains within the module are responsible for what the module transfers to the growing peptide.

NRPSs use unnatural amino acids as well as the naturally occurring 20 amino acids thereby leading to an increased amount of diversity in the peptide products they are able to produce (Sieber and Marahiel, 2005). Methylation or reductase can be performed on the carrier-protein tethered peptidyl chain during the elongation step and thereby introduce further variation and functionality (Meier & Burkart, 2009). There can be over a billion possible structures produced through the exploitation of the building blocks in the elongation steps, amount of reduction following the condensation reaction and the processing of products post-synthetically (Minowa *et al.,* 2007).

Intermediates that contain biosynthetic errors are corrected/ removed by the trans-acting TE-domains (Heathcote *et al.,* 2001; Schwarzer *et al.,* 2002; Yeh *et al.*, 2004). Type II TE has been found to have an active site that is easily accessible as well as relatively unspecific. Its main function however is its ability to repair. It recognizes and repairs misloaded carrier domains. It has been found in many different conformations and controls substrates access to the active site through the use of a helical lid (Koglin *et al.,* 2008).

1.7. NRPS Prediction Programs:

Protein-protein interactions of NRPSs' predictive accuracy was limited until the crystal structure of the first NRPS module was published which has since led to more positive research (Tanovic, Samel, Essen, & Marahiel, 2008). A group has since dedicated an entire platform to NRPS and PKS in which all knowledge of the chemical structures, characterization, experimental data, domain organization and functionality (Yadav *et al.,* 2003). This database allows for the identification of domains as well as location of the nearest match within the database to an inputted sequence (Starcevic *et al*., 2008).

Domain organization of NRPS has been limited due to the high sequence divergence between members of the same domain family. The NRPS A domain makes use of approximately 50 natural and unnatural amino acids thereby introducing even more variability. Ansari *et al.,* (2004) found that Conserved Domain Database (Marchler-Bauer *et al.,* 2013) and InterPro (Hunter *et al.,* 2012) were unable to accurately identify the organization of the domains with the modules due to the multi-functional nature of the proteins. SEARCHPKS was used in the developing of the NRPS-PKS webserver by integrating it with SEARCHNRPS since NRPS domains and Polyketide synthetase domains have been found in microbial genomes (Yadav et al., 2003).

Ansari *et. al.,* (2004) developed the NRPS-PKS system that is an automated computational platform that can be used to locate polyketide synthetases (PKS) and NRPS domains and predict the substrate specificity (Ansari *et al.*, 2004) (Table 1.1). This system was developed making use of SEARCHNRPS and

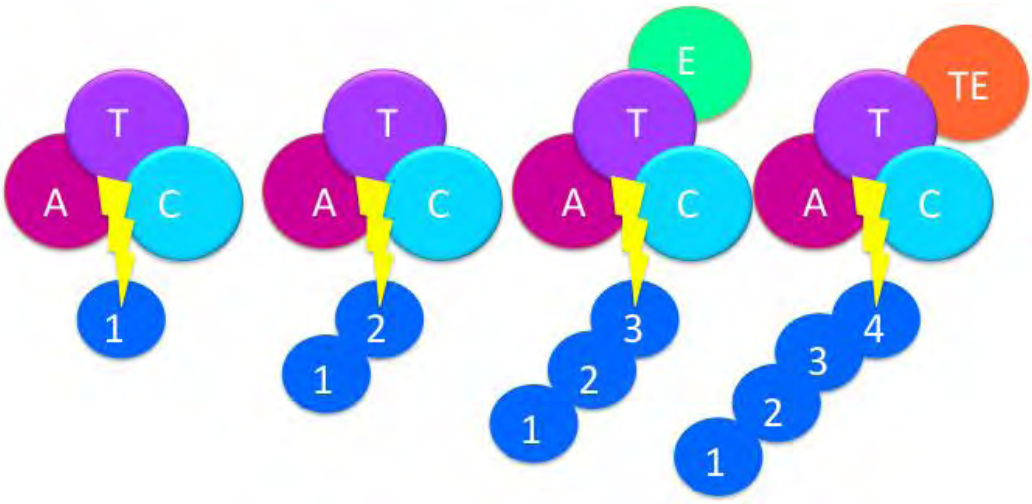


Figure 1.3: Schematic representation of the stages of NRP biosynthesis. NRP biosynthesis is very similar to that of peptides synthesized via ribosomes in that they follow the same processes of initiation, elongation, modification and termination. The first amino acid enters and is activated by the A-domain (initiation). The second amino acid is then activated by the second module of the NRPS to which the first amino acid is transferred (elongation). Modification occurs at NRPS modules which contain epimerization or methylation domains. The final peptide is released at the TE domain where termination takes place.

integrating it with previous prediction programs. Linker regions (defined as the C-terminal and N-terminal sequence regions) were used by Ansari *et al.* (2004) to

predict protein-protein interactions of NRPSs/PKSs. This was done based on the fact that the amino acid sequences of proteins that interact are functionally restricted. This was done using an algorithm, which searched for interacting sequence clusters within phylogenetic trees. This was thereby used to show that, unlike previously thought, the pathways of NRPSs/PKSs do not occur collinearly (Ansari *et al.,* 2004).

The Structure Based Sequence Analysis of Polyketides Synthetase software was developed by Anand *et al.,* (2010) (Table 1.1). This software has many uses specific to polyketide synthetase prediction as well as NPRSs. In this study the specific tool used was the NRPS-PKS prediction tool, which was developed to enable the identification of the catalytic domains within a sequence. The query sequence is compared with a database of NRPS and PKS clusters that have previously been experimentally characterized. This database contains 167 experimentally characterized NRPS and PKS gene clusters totaling approximately 4 400 catalytic domains. This tool performs a sequenced based analysis of the submitted sequence and analyses it for putative type I and II PKS proteins, NRPSs and PKS/NRPS hybrids (Anand *et al*., 2010).

Table 1.1: NRPS Prediction Programs. Names, abbreviations, web addresses and references of the NRPS prediction programs used.

Name	URL	Reference:
PKS/NRPS Analysis Web-site	nrps.igs.umaryland.edu/nrps/	(Challis et al., 2000)
SBSPKS (Structure Based Sequence Analysis of Polyketide Synthases)	www.nii.ac.in/~pksdb/sbspks/master.html	(Ansari *et. al.,* 2004)

1.8. Problem Statement:

Biocontrol agents and microbes are an emerging preferred alternative to pesticides within the agricultural sector; however, antibiotic resistance is a growing problem leading to the need to develop alternatives to the antibiotics currently in use. NRPs produced by microbes such as those within the *B. subtilis*

group are important source of antibiotics. Since their production is determined by the enzymes which produce them, NRPSs, as well as the order and number in which the NRPS modules occur, NRPSs are an area of interest for the development of novel antibiotics through their rearrangement and manipulation.

Previous studies have been able to determine important information behind understanding the mechanism of NRPS synthesis. However there are still gaps in the overall mechanism and more research is necessary to fully understand the entire mechanism of NRPS functionality, regulation and an extraordinary rate of genetic modification. Without a complete understanding of the mechanism through which NRPs are synthesized by NRPSs the enzymes cannot be safely and efficiently manipulated to produce useful novel antibiotics.

There are still gaps in the understanding of the evolution of the NRPSs in terms of the module and domain arrangement and shuffling. The full extent to which the linker regions affect this process and the end product formed is also not yet fully understood as well as the amount of conservation within the linker and terminal regions of the NRPSs.

The test strain has shown promising anti-phytogenic properties in laboratory testing however no published data is available on the structure of NRPSs or sequence of this strain. The amino acid composition as well as the order of the modules has therefore not been compared to that of neighboring, better studied/ published strains.

1.9. Aims:

This project was a pilot study on the test strain, *B. atrophaeus* UCMB 5137 (63Z), that aims to develop a better understanding of the strain. Methods tested in this study would in future work be applied to several more strains, which have since been sequenced by our collaborators at the University of Pretoria.

The study aimed to investigate the amino acid composition of the NRPs within the test strain in comparison to other neighboring strains of the *B. subtilis* group (*B. atropheaus* 1942, *B. subtilis* BS5n*, B. subtilis subsp. Subtilis* 168 and *B. amyloliquifaciens* FZB42). A further aim was to determine the level of conservation between the linker and terminal regions of the NRPSs of the test strain. Preliminary 3D structures of the domains of NRPSs of the test strain were calculated with the aim of further investigation of the active site residues within the A-domain as well as in future studies of interactions and reshuffling.

1.10. Objectives:

1. To identify the NRPS modules contained within the sequence of the test strain.
2. To identify which of these NRPS modules found in the test strain genome are true positives and which are false positives.
3. To identify duplications within the predicted NRPS modules.
4. To determine whether predicted NRPS modules compose NRPS genes or gene clusters.
5. To identify close neighboring strains to the test strain through a phylogenetic analysis.
6. To determine the NRPS modules found within the neighboring strains.
7. To determine which amino acid the NRPS module encodes attachment for in both the test and neighboring strains.
8. To compare the NRPS modules found within the test strain to those found within the neighboring strains.
9. To investigate the conservation between the linker and terminal regions of the NRPSs of the test strain.
10. To construct a homology model of the NRPS modules and/or domains of the test strain.

SECOND CHAPTER: MODULE IDENTIFICATION, CLASSIFICATION AND CHARACTERIZATION

2.1 Overview

In this Chapter the test strain, *B. atrophaeus* UCMB 5137 (63Z), was investigated to determine the NRPSs it contains and the amino acids encoded by these NRPSs to produce NRPs. The strain was also investigated in a phylogenetic study to determine its closest neighboring strains. Once the neighboring strains were determined this information was used to compare the test strain to that of the neighboring strains in terms of NRPSs contained within the strains as well as the NRP amino acids produced. A study was also performed on the terminal and linker regions of the NRPSs to gain insight into their possible involvement in the synthesis of NRPs.

2.2 Introduction

Non-ribosomal peptide synthetases are made up of several modules, which act together to add amino acids to a growing peptide chain producing non-ribosomal peptides/ antibiotics (Nagórska et al., 2007; Ongena & Jacques, 2008). The order in which modules are arranged within the genes dictate the final product produced by the NRPS. Modules can be rearranged within the organisms genome leading to different NRPs being produced. This is an important area of interest in the hopes of discovering novel natural products and antibiotics (Nguyen *et al.,* 2006; Weissman and Leadlay, 2005). The modules of *B. atrophaeus* UMBC 5137 (63Z) were investigated, in relation to neighboring strains, to determine if there was any rearrangement or changes between the NRP products produced by the NRPSs.

Sequencing has become a lot more affordable recently leading to an increase in the amount of data sequenced and requiring analysis. Generally the main drawback in analysis of a genome is the fact that assembly and manual annotation are time consuming. Annotation is usually done automatically through programs such as RAST however there are drawbacks to automated annotation since they are not manually curated. Manually curating or checking data is a long and specialized process usually delaying the analysis of the data. The script used in this project (nrps.py) is therefore beneficial as an initial analysis tool since it can be used on a raw DNA sequence.

2.3. Phylogenetic analysis:

Studies have been conducted to ascertain the clustering and phylogenetic routing of the A-domains from different organisms. This was done by aligning sequences of interest, trimming the sequences down to the presumed binding pocket constituents, realigning the sequences and carrying out phylogenetic studies on the selected sites. These selections were assumed to represent the codons of substrate specificity and that the phylogenetically similar domains would recognize the same substrate and would therefore be found to cluster closely in a phylogenetic tree. Bacterial and fungal sequences were not separated but rather dispersed according to their structural homology and the representation of the selected residues. This method could therefore be useful in the future in the determination and prediction of phylogenetic origin of the substrates of newly discovered domains (Stevens *et al.,* 2006).

Evolutionary and phylogenetic studies can provide insightful information into the intricate processes and mechanisms involved in the design of NRPSs (Jenke-Kodama & Dittmann, 2009). There are two main areas of molecular evolutionary analysis; one geared to determining the principles of macromolecular evolution, through attempting to determine the mechanisms involved in the observed changes and thereby forming evolutionary patterns, and the second is focused on the DNA or protein sequence history in relation to the species evolutionary past

(Harrison & Langdale, 2006). Phylogenetics is also a useful approach in the classification of proteins according to functional specificity thereby indicating a selection process for enzymatic functions as well as for the detection of functionally important genes (Jenke-Kodama & Dittmann, 2009).

2.4. Methodology:

This work consisted of several different steps (Figure 2.1). The first steps involving sequencing of the test strain *B. atrophaeus* UCMB 5137 (63Z), assembly of the genome after sequencing and the annotation were carried out by the project collaborators at the University of Pretoria. The subsequent steps were carried out during this project at Rhodes University.

2.4.1. Background:

The elements in this section were carried out or organized by collaborators at the University of Pretoria (Figure 2.1). The genome of the *B. atrophaeus* UMBC 5137 (63Z) strain was sequenced by Macrogen in South Korea and initially 52 contigs were constructed by using Velvet and CLC Genomics Workbench. Fifty-two contig sequences were of the total length of above 6 Mbps. The total length of the *Bacillus* genome is approximately 4 Mbp indicating there were many overlapping regions and duplications especially in the NRPS genes due to the fact that they are built of repeated elements therefore making assembly generally very difficult.

2.4.2. Module identification, classification

This section onward was carried out at RUBi (Figure 2.1). The *B. atrophaeus* UMBC 5137 (63Z) strains genome was annotated using RAST (Aziz et al., 2008). The annotated genome was then passed through a python script developed by O. Reva (2011) (Appendix 1.2) to identify NRPS genes/ modules. The script searches through the submitted genome of the organism and searches for NRPS modules by comparing them to a library of known NRPS modules compiled by O. Reva (Appendix 1.2). The module search by nrps.py could be carried out before RAST annotation as the script is able to operate on a raw, unannotated DNA sequence.

The genome annotation step was carried out to make future dealings with the genome easier and more manageable, especially visualization steps.

The script (nrps.py) passes the sequences of the modules through a BLAST search to determine the best prediction of what the module codes for. The sequences were then examined manually for false positives and duplications. Once duplications and false positives were removed the remaining sequences were treated as potential NRPS modules.

These gene locations were observed using Artemis (Rutherford et al., 2000). Gene sequences identified by nrps.py to contain a NRPS module, as well as neighboring genes, were extracted from the genome. These sequences were used to confirm the NRPS modules by two webservers containing known NRPS module information; PKS/NRPS Analysis Web-site (PKSAW) (Bachmann & Ravel, 2009) and Structure Based Sequence Analysis of Polyketide Synthases (SBSPKS) (Anand et al., 2010) (Table 1.1). The modules recognized by these webservers were compared to those identified by the python scripted program (nrps.py) to determine which were true positives and which were false positives. The NRPSs identified were shown on the SVG map (Figure 2.2).

2.4.3. Module composition

Module locations were viewed in the genome using Artemis (Rutherford et al., 2000) to determine whether they were from a single gene or a cluster of genes. Due to the fact that the genome was assembled using contigs (See Chapter 2.4.1) and automatically annotated, using RAST (See Chapter 2.4.2), gene cluster areas could be representative of a single gene which had not been correctly annotated as a separate gene, giving the appearance of gene clusters as apposed to a single gene if it had been incorrectly separated or joined within the gene or between multiple genes.

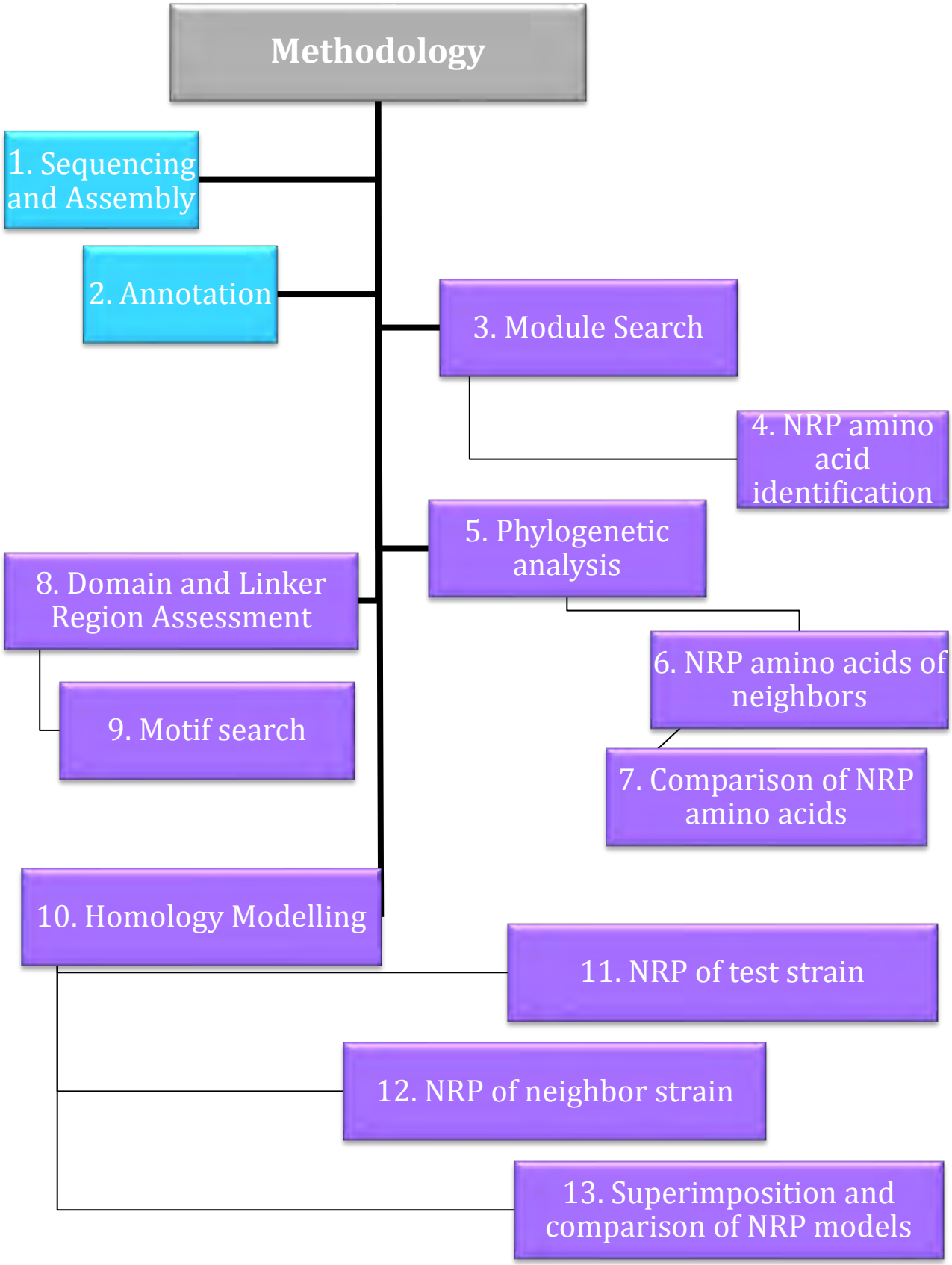


Figure 2. 1: Flow Diagram of the Methodology for this work. Collaborators at University of Pretoria carried out blue colored steps. Steps shown in purple were carried out personally during this work.

2.4.4. Comparison of polypeptides encoded by NRPS genes

The amino acid for which the A-domains within the modules encode attachment for was also predicted by the two webservers (SBSPKS and PKSAW) for the test strain, *B. atrophaeus* UMBC 5137 (63Z) (Table 1.1). These amino acid incorporation predictions were compared to that of the same non-ribosomal peptide synthetase module in the SBSPKS database of known NRPS and in some cases NRPS-PKS hybrids (Anand *et al.*, 2010).

Modules identified within the *B. atrophaeus* UMBC 5137 (63Z) strain were compared to neighboring strains identified in Chapter 2.2.5.2 (Figure 2.3). Four reference strains; *B. atrophaeus* 1942*, B. subtilis* BS5n, *B. subtilis subsp. subtilis* 168 and *B. amyloliquifaciens* FZB42 were also run through the python script (nrps.py) (Appendix 1.1) in the same manner as with the test strain (Chapter 2.2.1 and 2.2.3). The resulting identified modules and domains were compared to those from the *B. atrophaeus* UMBC 5137 (63Z) strain (Table 2.3).

The SBSPKS database was also tested by submitting sequences which were listed on the site under a specific NRPS to confirm the results. This was done using the genes recorded for feng synthetase (AF087452, AF023465, AF023464, L42523 and AJ011849) (Lin et al., 1999) and myc synthetase (AF184956.1) for the *Bacillus subtilis* F29-3 and ATCC 6633 strains respectively (Table 2.4) (Duitman *et al.*, 1999).

2.4.5. Phylogenetic Analysis:

Species Level

A species tree was generated with the other *B atrophaeus* strains, which have been sequenced by the University of Pretoria lab but not yet fully characterized as well as the other *B. subtilis* and *B. amyloliquifaciens* strains (Figure 2.3). The species tree was created using MEGA 5.05 (Tamura *et al.,* 2011) by constructing a multi-locus tree by aligning all orthologous proteins found in all the reference genomes using MUSCLE (Edgar, 2004). The alignment was then edited using

Gblocks to remove any areas of the alignment, which were ambiguous. A Neighbour-Joining tree as well as a Minimum evolution tree, each using 500 bootstrap replicates, were then constructed (Castresana, 2000).

Gene Level

Gene trees of the genes containing NRPS modules were also constructed to determine the amount of deviation/ evolution between the different neighboring strains (Figures 2.4-2.8). MEGA was used to generate the Maximum Likelihood gene trees using 1000 bootstrap replicates (Tamura *et al.,* 2011).

2.4.6. Linker and Terminal Region Investigation:

The positions of domains were identified using the SBSPKS webserver in Chapter 1.2.4. The absence of domains was used to indicate areas of linkers or terminal regions within the gene sequence known to contain NRPS modules. The range extractor of The Sequence Manipulation Suite (Stothard, 2000) was used to extract the sequences of the linker and terminal regions from the gene sequence identified to contain the NRPS. Linker regions occur between the modules, referred further to as terminals, as well as between the domains within the modules, referred to further as linkers (Figure 1.1).

2.5. Results and Discussion:

2.5.1. NRPS identification in newly sequenced genome

B. atrophaeus UCMB 5137 (63Z) strain was investigated to determine which NRPS modules it contains. This was done using the script nrps.py (developed by O. Reva (2011)) and genes identified to contain NRPSs were run through two online servers (Table 2.1). The whole genome was not submitted to the online servers as they are only able to take short peptide sequences and not full genomes. The script developed is, however, able to perform this search on a whole genome even in a raw, unannotated DNA format.

The nrps.py program originally identified 48 DNA fragments, which appeared to represent NRPS modules, however, on closer inspection of the sequences it was found that some of the sequences were translatable to proteins whereas some were not. The non-translatable sequences were either false predictions or pseudo-genes, which had been knocked out due to mutations. Once the duplications had been removed the remainder were investigated for NRPS modules.

These genes were confirmed to contain NRPS modules by both of the online programs (Table 1.1). The locations within the genes were slightly differently predicted by all three methods; nrps.py, SBSPKS and PKSAW. The end products predicted by the programs were the same for all three programs with the exception of one NRPS which was recognized by nrps.py as an NRPS and identified as plp synthetase in its closest neighbor strain (*B. atrophaeus* 1942) by means of a BLAST search whereas the other two online programs found the NRPS to code for feng synthetase, however, feng and plp synthetases are synonymous (Table 2.1).

The nrps.py identified a total of 29 modules, 21 modules of which code for 6 different known NRPSs. SBSPKS identified 27 modules coding for 5-7 NRPSs. PKS/NRPS Analysis Website identified 23 modules to code for 8 NRPSs, one of which was identified as a novel A-domain nature. The NRPS modules identified by nrps.py were found to have the closest matches to the following NRPS modules within the database; Myc synthetase, plp/ feng synthetase, bct synthetase, bcb synthetase, polyketide synthetase and srf synthetase (Table 2.1). These were mapped to the genome of the test strain (Figure 2.2) by an SVG mapper by O. Reva (2012) (Appendix 1.3). The modules within the test genome in most cases contained less modules in some of the NRPSs in comparison to those recorded within the database. The most complete are of myc synthetase and plp/ feng synthetase.

For one of the NRPS genes all three of the programs identified a different product; polyketide synthetase, lichenycin synthetase, microcystin synthetase and

saframycin Mx1 synthetase and could possibly indicate a false positive (Table 2.1). However, since the two online programs both identified a NRPS product within the gene identified by nrps.py it could possibly be attributed to a novel sequence signature. Since this cannot be further tested in this instance without physical testing and the focus of the study was NRPSs and not PKS's this was not further studied in this case. Another possible explanation could be attributed to the fact that the script searches for NRPS sequences to which it was similar as is usually the case where long-chain-fatty-acids are incorrectly annotated and are in fact NRPSs.

A few other genes were identified by nrps.py to contain NRPS modules, however, the program either identified their products as unknown or as products, which are not produced by non-ribosomal peptide synthetases. Only one of the genes identified by nrps.py to contain a NRPS module, but classified as an unknown product during the BLAST search, was identified by both SBSPKS and PKS/NRPS as feng synthetase. Due to the location of the gene within the genome and the surrounding genes it is reasonable to assume this identification is correct as it is between two other genes identified to contain NRPS modules for feng/ plp synthetase (Table 2.1). Locations of predicted PKS/NRPS genes in ordered contigs of UCMB-5137 (63Z) are shown in Figure 2.2.

None of the modules were found to be stand-alone. All were found to either be a part of a gene or a gene cluster. Myc synthetase is made up of 3 genes, which constitute the three subunits of myc synthetase. The A and C subunits are shorter genes flanking the B subunit which was found to be very long in comparison to the other two genes. Bcb synthetase, bct synthetase and srf synthetase were all found to be attributed to a single gene. According to the SBSPKS database srf synthetase is generally made up of a gene cluster of 3 genes with 7 modules. Bct synthetase was identified to be made up of 3 genes and 12 modules by the SBSPKS database and bcb synthetase to have 3 modules within 2 genes. On inspection of these areas in the genome it was found that the test strain only contained fragments of the NRPSs bct synthetase and bcb synthetase within the contigs. Srf synthetase, however, was truncated by the contig border and it is

therefore not possible to determine whether the remainder of the synthetase is contained within the contig border/ overlap or whether it is, like with the bct and bcb synthetases, not complete. This is an assembly problem, which is being addressed by our collaborators.

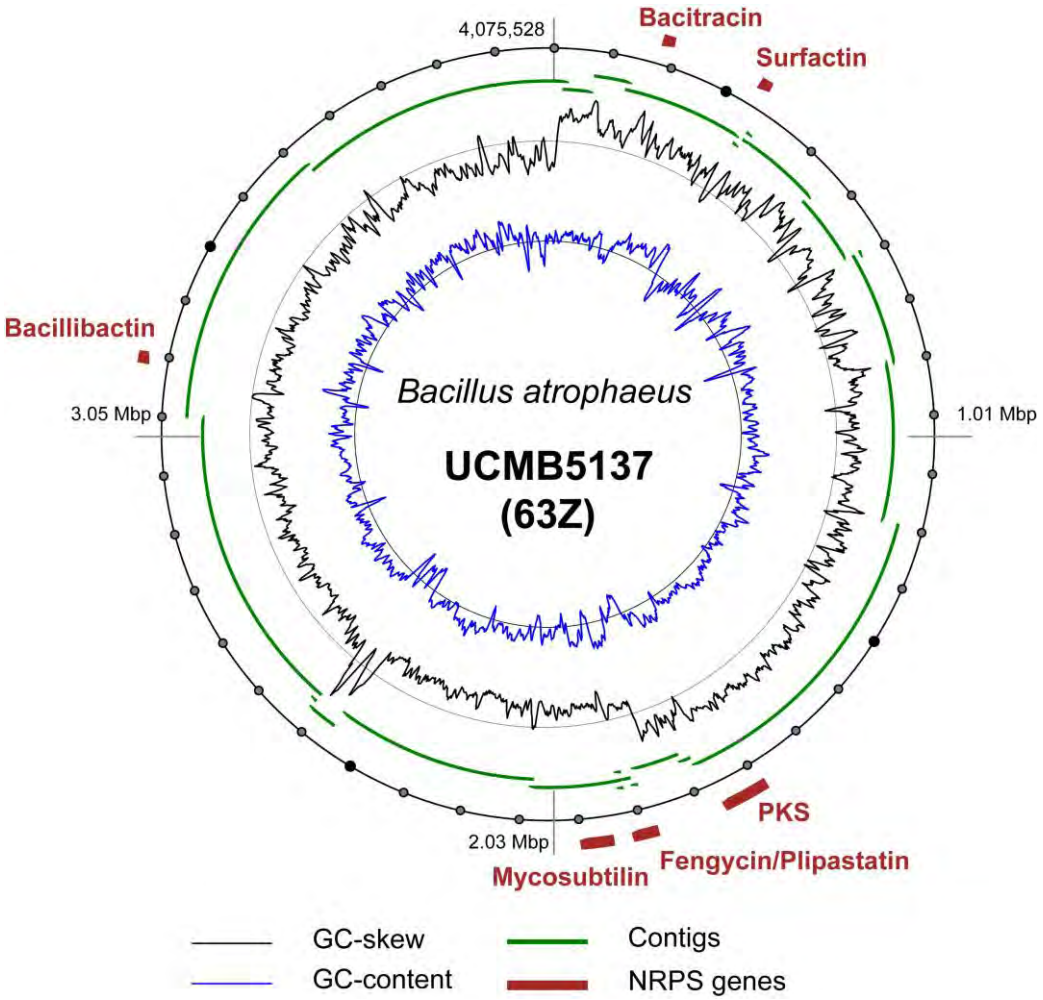


Figure 2.2: *Bacillus atropaheus* UMBC-5137 (63Z) genome map showing identified non-ribosomal peptide synthetase modules. Constructed with SVG mapper. NRPSs identified are shown in red; Bcb synthetase, Bct synthetase, Srf synthetase, Myc synthetase, plp/ feng synthetase and Polyketide synthetase. Contigs constructed are shown in green with GC skew and GC content shown in black and blue respectively.

Polyketide synthetase encoded by the gene pks in *Bacillus* is generally made up of 5 long genes however in this case only 2 of the 5 genes, the 1st and 4th genes in the group of 5, were identified to contain A-domains.

Table 2.1: NRPS modules identified in *Bacillus atropheaus* UCMB 5137 (63Z) strain. Genes identified by nrps.py to contain NRPS modules were run through two online programs (SBSPKS and PKS/NRPS Analysis webserver) and results compared.

UMBC 5137 (63Z)	Identified by nrps.py		SBSPKS		PKS/NRPS Analysis Webserver	
Gene	**# of Modules**	**NRPS**	**# of Modules**	**NRPS**	**# of Modules**	**NRPS**
28740	2	bacillibactin synthetase	2	bacillibactin synthetase	2	Nostopeptolide/**bacillibactin** / CDA peptideII /Coelichelin/Pyoverdin/ fengycin
40250	2	bacitracin synthetase 1	2	tyroc/mycos/ituri/grami/**bacitracin synthetase**	2	Myxothiazol/Epothilone /**Bacitracin/NOVEL A-DOMAIN SIGNATURE**
41820	4	surfactin synthetase	2	surfactin synthetase	2	surfactin synthetase
13580	1	PKS	2	lichenycin / microcystin synthetase	1	saframycin Mx1 synthetase
15550	1	mycosubtilin subunit C	2	mycosubutilin synthetase	2	mycosubutilin(A/B/C)/tyrocidine/bacitracin/ Phosphitricin// MicrocistinA
15560	3	mycosubtilin subunit B	4	mycosubutilin synthetase	4	mycosubbutilin/nostopeptolode/fengycin
15570	1	mycosubtilin subunit A	2	iturin/**mycos**	1	mycosubutilin(A/B/C)/tyrocidine/bacitracin
15030	1	plipastatin synthetase	1	fengycin synthetase	1	fengycin synthetase
15050-60	1		3	fengycin/ tyrocidin	2	fengycin synthetase
15060	1		3	fengycin/ gramicidin/iturin	2	fengycin synthetase
15070	1		0	None found	0	None found
15080-90	1		1	fengycin synthetase	1	fengycin synthetase
15100	2		2	fengycin synthetase	2	Coelichelin /Pyoverdin/fengycin
15000	1	acyl-CoA dehydrogenase	0	None found	0	None found
15040	1	unknown	1	fengycin synthetase	1	fengycin synthetase
15580	1	Malonyl CoA-acyl carrier protein transacylase	0	None found	0	None found
21990	1	putative transcriptional regulator	0	None found	0	None found
25940-50	1	YtmB; , putative hydrolase	0	None found	0	None found
32530	1	putative membrane bound transcriptional regulator	0	None found	0	None found
41840-60	1	unknown	0	None found	0	None found
41890-900	1	unknown	0	None found	0	None found

Feng synthetase in *B. subtilis* according to the SBSPKS is made up of 10 modules in 5 genes. Plipastatin synthetase operon identified in the *B. atrophaeus* UMBC 5137 (63Z) is spread over 5 contigs in which 8 ORFs were identified to contain modules encoding for plp synthetase. Many of these ORFs were truncated by contig borders and mapping them against the plp synthetase of the reference strain *B. atrophaeus* 1942 showed that they code for the same 5 genes and an overall of 10 modules were identified from the SBSPKS Database.

Originally on investigation of the modules within the genes of the *B. atrophaeus* UMBC 5137 (63Z) strain there were many novel modules found. On closer inspection the modules appeared to be rearranged in a different order to that of the database record, however, this is due to the direction of the strand and there were in fact no rearrangements found in the test strain in comparison to that of the neighboring strands and database.

2.5.2. Comparison of polypeptides encoded by NRPS genes

NRPS encode peptides with important antibiotic properties whose sequence can be determined through the specificity and order of the A-domains within the NRPS genes (see Chapter 1). The NRP sequence was therefore investigated.

The amino acids encoded by the modules of the genes recognized by nrps.py to contain NRPSs were also confirmed by the two online webservers. However, only the SBSPKS website contains identified known amino acids, domains and modules of NRPS antibiotics and is therefore the only one which can be used for comparative purposes (Table 2.1).

The modules within the genes appear to be in the same order coding for the same amino acids as those in the SBSPKS database (Table 2.2). Some of the genes were in a different order though due to the direction of the strand, however, when viewed in the same direction as that recorded in the database, no rearrangements are found.

Myc synthetase was the most completely assembled in 63Z and corresponded to that recorded in the SBSPKS database (Table 2.2). Feng/ plp synthetase was quite complete in comparison to the database record (Table 2.2). The test strains feng synthetase contained 10 modules, the same as in the database. On first inspection of the amino acids encoded for in the test strain in comparison to the database seem to be exactly the same.
When looking at the modules for the different NRPSs and the amino acids they encode attachment for it was noted that the same NRPS were mostly found in the other strains and no other NRPS (other than the 5 identified in the test strain) were found in the other strains. The only identified NRPSs in all the strains examined were; feng/ plp synthetase, myc synthetase, srf synthetase, bcb synthetase and bct synthetase. No additional NRPSs, not found in the test strain, were found in the other strains.

Bacitracin (bct) synthetase was also found in the *B. atrophaeus* 1942 and *B. amyloliquifaciens* FZB42 strains however they only contained a small subset of the amino acids known to make up the first described bct synthetase from *B. licheniformis* ATCC 10716 (Konz D et al., 2005) (Table 2.2).

Bcb synthetase was found in all the strains with the exception of the first gene encoding the attachment of the first amino acid, Dhb, that was reported in the homolog bcb synthetase in a *B. subtilis* producer (May *et al*., 2001) (Table 2.2).

Srf synthetase was found in all the strains however in all reference genomes the genes of srf lipopeptide encoded smaller number of amino acid residues in the final product than was reported by Bruner *et al.,* (2002). *B. atrophaeus* 1942, *B. subtilis* BS5n and *B. amyloliquifaciens* FZB42 were found to contain the full srf synthetase modules with the exception of the last Leucine (Table 2.2).

Myc synthetase was not found in the *B. subtilis subsp. subtilis* 168 and *B. subtilis* BS5n strains. The remainder of the strains, *B. atrophaeus* 1942 and *B. amyloliquifaciens* FZB42, showed myc synthetase to be in the same order as that of the test strain (Table 2.2). The only exception was in the case of *B.*

amyloliquifaciens FZB42 where the last amino acid attached was found to be Tyr whereas it is Asn in the other strains and database.

Table 2.2: Detected antibiotic NRP product amino acid sequence in *Bacillus atropheaus* UCMB 5137 (63Z) strain in comparison to the SBSPKS database. Amino acids appearing in both the strain as well as the database are color-coordinated.

NRPS	Genes	Modules	Amino	Database NRPS	Genes	Modules	Amino
Bacitracin synthetase 1	1	2	Hpg	Bacitracin synthetase	3	10	Ile
			Cys				Cys
							Leu
							Glu-D
							Ile
							Lys
							Orn-D
							Ile
							Phe-D
							His
							Asp-D
							Asn
Bacillibactin synthetase	1	2		Bacillibactin synthetase	2	3	Dhb
			Gly				Gly
			Thr				Thr
Surfactin synthetase	1	2	Glu	Surfactin synthetase	3	7	Glu
			Leu				Leu
							Leu-D
							Val
							Asp
							Leu-D
							Leu
Mycosubtilin synthase	3	8	ACoL	Mycosubtilin synthetase	3	8	AcoL
			Asn				Asn
			Tyr-D				Tyr-D
			Asn-D				Asn-D
			Gln				Gln
			Pro				Pro
			Ser-D				Ser-D
			Asn				Asn
Plipstatin/ fengycin Synthetase	9	10	Glu	Fengycin Synthetase	5	10	Glu
			Orn				Orn-D
			Tyr				Tyr
			Thr				Thr-D
			Glu				Glu
			Ala/Val				Ala/Val
			Pro				Pro
			Glu				Glu
			Tyr				Tyr-D
			Ile				Ile

Feng synthetase, also referred to as plp synthetase, was found in all of the strains and appeared in the same order to the database in each case (Table 2.2). The only differences found were in the cases of the two *B. atrophaeus* strains (UCMB 5137 and 1942) and *B.* amyloliquifaciens FZB42 strain where the Orn and Tyr amino acids were in the L conformation whereas these amino acids appear in the D-conformation in the other *B. subtilis* and database strains.

Bacillus amyloliquifaciens ssp. plantarum FZB42 strain has been well annotated was therefore used as a control. It was also a reference strain in this research. The nrps.py script did not appear to miss any NRPSs present in the genome that had been annotated as such. It did manage to pick up some of the NRPSs that had been annotated as Long-chain-fatty-acids-CoA-ligases, peptide synthetases and hypothetical proteins.

2.5.3. Database inconsistency

A comparison was done using the sequences from which the database was compiled. *B. subtilis* ATCC 6633 used for myc synthetase and *B. subtilis* F 29-3 used for plp/ feng synthetase (Table 2.4). The accession numbers for the sequences were obtained from the PUBMED article link attached to the NRPSs in the SBSPK database.

For the plp/ feng synthetase reference the only notable differences were that some of the amino acids were not in the D-confirmation even though in the database they are listed as being in the D-conformation. An additional Ile was detected within the same FASTA file not shown in the original database entry. For the myc synthetase reference only the B subunit was found within the sequence when tested through the SBSPKS database (Table 2.4).

Table 2.3: Module comparison between test strain and neighboring strains. Modules detected in test strain were compared to those in neighboring strains by comparing the amino acid incorporated determined by SBSPKS Webserver. Amino acids incorporated into the NRP which correlated between strains were colored-coordinated.

NRPS	*B.atrophaeus* UMBC 5137 (63Z)	*B. atrophaeus* 1942	*B. subtilis* 168	*B. subtilis* BS5n	*B. amyloliquifaciens* FZB42	SBSPKS	Reference strain
Bacitracin synthetase 1	Hpg	Hpg			Cys	Ile	*B. licheniformis* ATCC 10716
	Cys	Cys			Cys	Cys	
						Leu	
						Glu	
						Ile	
						Lys	
						Orn	
						Ile	
						Phe	
						His	
						Asp	
						Asn	
Bacillibactin synthetase						Dhb	*B. subtilis*
	Gly	Gly	Gly	Gly	Gly	Gly	
	Thr	Thr	Thr	Thr	Thr	Thr	
Surfactin synthetase	Glu	Glu	Glu	Glu	Glu	Glu	?
	Leu	Leu	Leu	Leu	Leu	Leu	
		Leu	Leu	Leu	Leu	Leu	
		Val		Val	Val	Val	
		Asp		Asp	Asp	Asp	
		Leu		Leu	Leu	Leu	
		Leu		Leu	Leu	Leu	

Table 2.3 (Continued): Module comparison between test strain and neighboring strains. Modules detected in test strain were compared to those in neighboring strains by comparing the amino acid incorporated determined by SBSPKS Webserver. Amino acids incorporated into the NRP which correlated between strains were colored-coordinated.

NRPS	*B.atrophaeus* UMBC 5137 (63Z)	*B. atrophaeus* 1942	*B. subtilis* 168	*B. subtilis* BS5n	*B. amyloliquifaciens* FZB42	SBSPKS	Reference strain
Mycosubtilin synthase	ACoL	ACoL			ACoL	ACoL	*B. subtilis* ATCC 6633
	Asn	Asn			Asn	Asn	
	Tyr-D	Tyr-D			Tyr-D	Tyr-D	
	Asn	Asn			Asn	Asn-D	
	Gln	Gln			Pro	Gln	
	Pro	Pro			Glu	Pro	
	Ser	Ser			Ser	Ser-D	
	Asn	Asn			Thr	Asn	
Plipstatin/ fengycin Synthetase	Glu	Glu	Glu	Glu	Glu	Glu	*B. subtilis* F 29-3
	Orn	Orn	Orn-D	Orn-D	Orn	Orn-D	
	Tyr	Tyr	Tyr	Tyr	Tyr	Tyr	
	Thr	Thr	Thr	Thr	Thr	Thr-D	
	Glu	Glu	Glu	Glu	Glu	Glu	
	Ala/Val	Ala/Val	Ala/Val	Ala/Val	Ala/Val	Ala/Val	
	Pro	Pro	Pro	Pro	Pro	Pro	
	Glu	Glu	Glu	Glu	Glu	Glu	
	Tyr	Tyr	Tyr-D	Tyr-D	Tyr	Tyr-D	
	Ile	Ile	Ile	Ile	Ile	Ile	

Table 2.5: Control test of database against sequences used to generate the database. *Bacillus subtilis* ATCC 6633 used for myc synthetase and *Bacillus subtilis* F 29-3 used for feng synthetase as per the Databases listed strains for the particular NRP.

NRPS	NRPSDB	Database Test
Mycosubtilin synthase	ACoL	
	Asn	
	Tyr-D	Tyr-D
	Asn-D	Asn
	Gln	Gln
	Pro	Pro
	Ser-D	
	Asn	
Plipstatin/ fengycin Synthetase	Glu	Glu
	Orn-D	Orn
	Tyr	Tyr
	Thr-D	Thr
	Glu	Glu
	Ala/Val	Ala/Val
	Pro	Pro
	Glu	Glu
	Tyr-D	Tyr-D
	Ile	Ile
		Ile

2.5.4. Phylogenetic comparison of NRPS genes

All orthologous genes in all reference genomes were located through a reciprocal BLAST search and then these sequences were translated, aligned and edited using Gblocks. These sequences were then concatenated into an alignment and used to construct a multi-locus species tree using MEGA 5.05 (Figure 2.3).

Species trees were constructed using the Neighboring joining and Minimum evolution methods using 500 bootstrap replicates in each case, both producing identical trees (Figure 2.3). The species tree showed that the closest neighbors to *B. atrophaeus* UMBC 5137 (63Z) strain are those of the other genomes sequenced from *B. atrophaeus* followed by those from *B. subtilis* and *B. amyloliquifaciens.*

The species tree demonstrates that the three species are well separated and the test strain unambiguously belongs to the *B. atrophaeus* group.

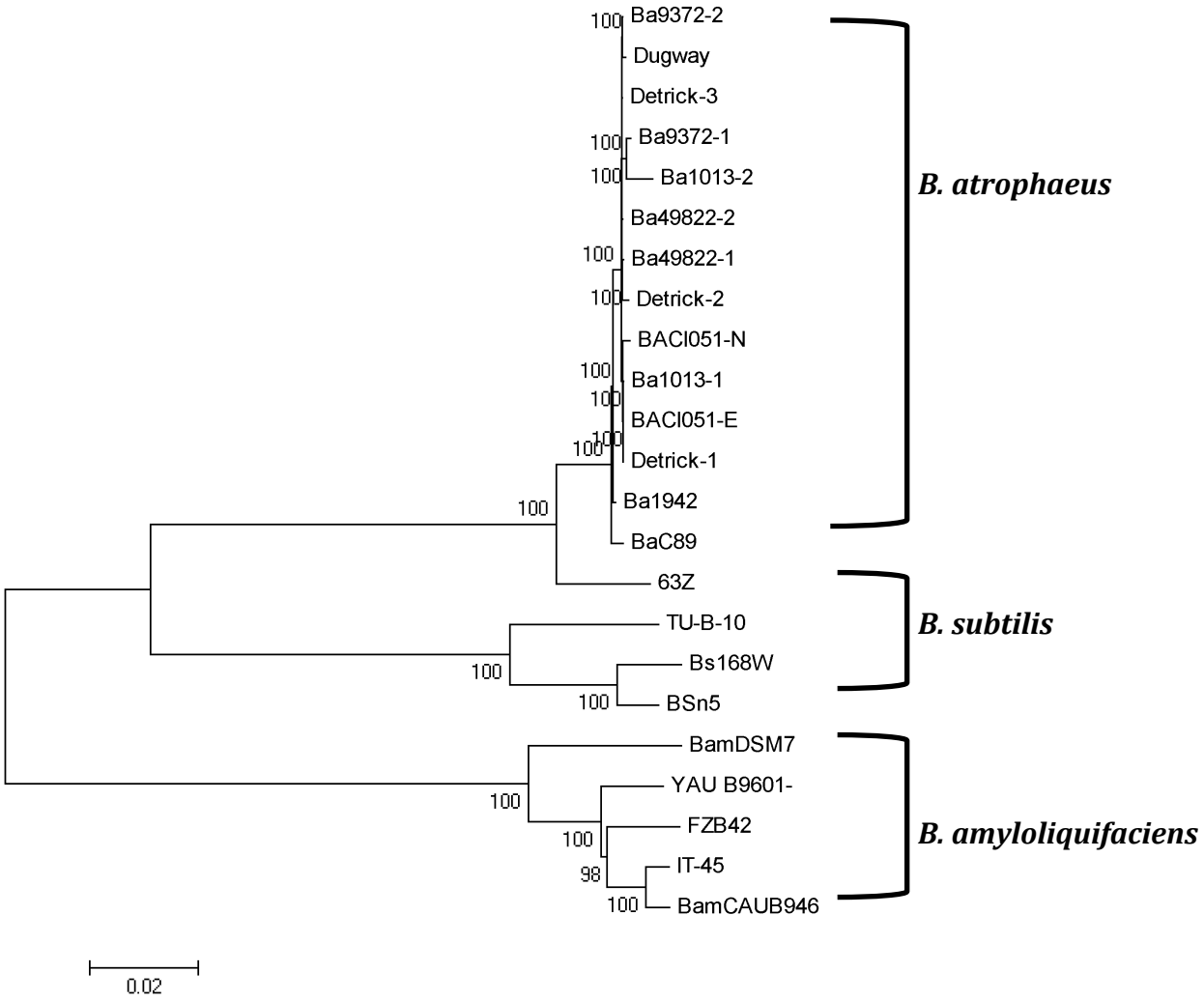


Figure 2.3: Species tree of *Bacillus atrophaeus* UMBC 5137 (63Z). Constructed using the neighbor joining and Minimum evolution methods using 500 bootstrap replicates in each case. 2,517 orthologous proteins of the total length of 549,615 amc. were concatenated into alignment which was then used to construct this tree. Groupings of species within the *B. subtilis* group are shown in bold on the right of the tree.

Gene trees were constructed for bcb synthetase (Figure 2.4), srf synthetase (Figure 2.5), feng synthetase (Figure 2.6) and myc synthetase (Figure 2.7). When reviewing the gene trees it appears in the same order as that of the species tree; *B. atrophaeus* followed by *B. subtilis* and then *B. amyloliquifaciens* (Figure 2.3). Maximum-Likelihood gene trees with 1000 bootstrap replicates were created of each of the gene trees within the test strain containing NRPS modules. A gene tree was constructed for srf synthetase with only the common gene in the test and neighboring strains. The tree revealed the closest neighbor to be from *B.*

atrophaeus 1942 followed by the *B. subtilis* and then *B. amyloliquifaciens* strains (Figure 2.5.).

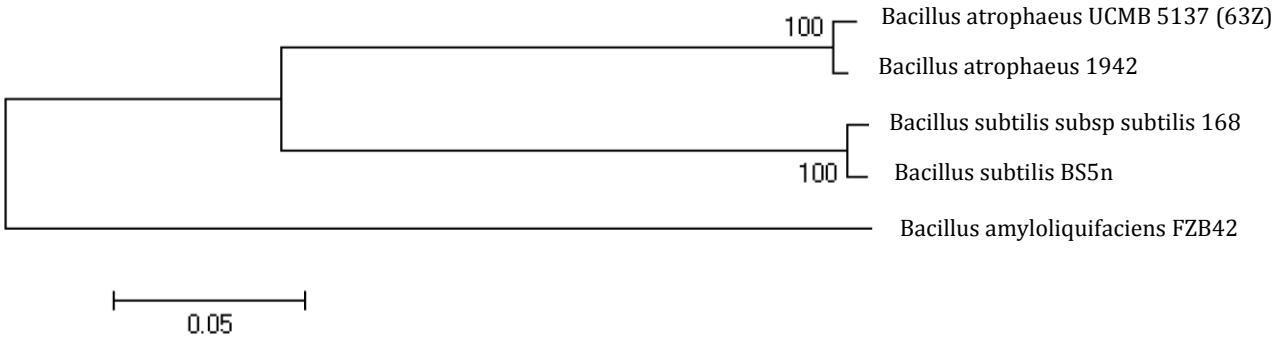


Figure 2.4: Gene tree of Bcb synthetase of *B. atrophaeus* UCMB 5137 (63Z) strain. Constructed using the Maximum-Likelihood method with 1000 bootstrap replicates.

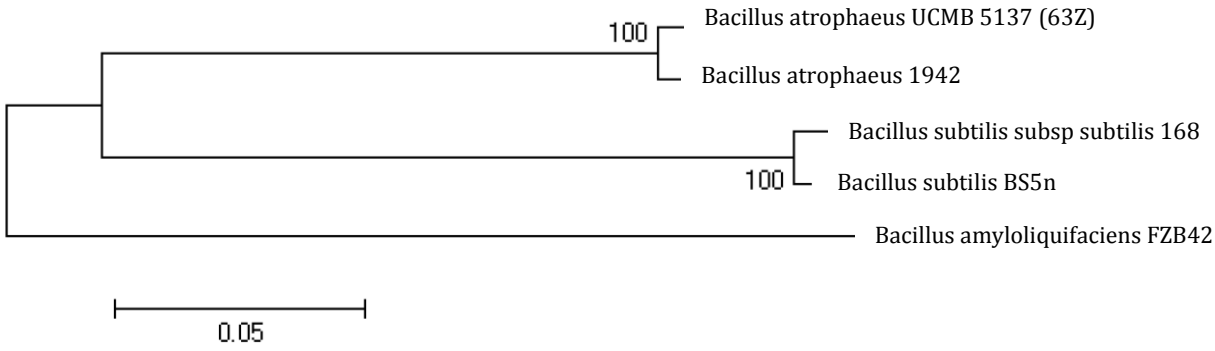


Figure 2.5: Gene tree of Surfactin synthetase of *B. atrophaeus* UCMB 5137 (63Z) strain. Constructed using the Maximum-Likelihood method with 1000 bootstrap replicates.

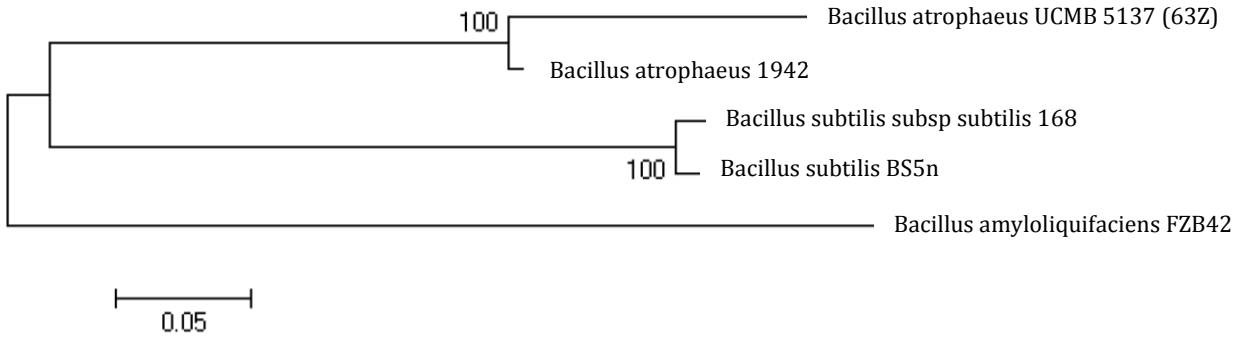


Figure 2.6:Gene tree of feng synthetase of *B. atrophaeus* UCMB 5137 (63Z) strain. Constructed using the Maximum-Likelihood method with 1000 bootstrap replicates.

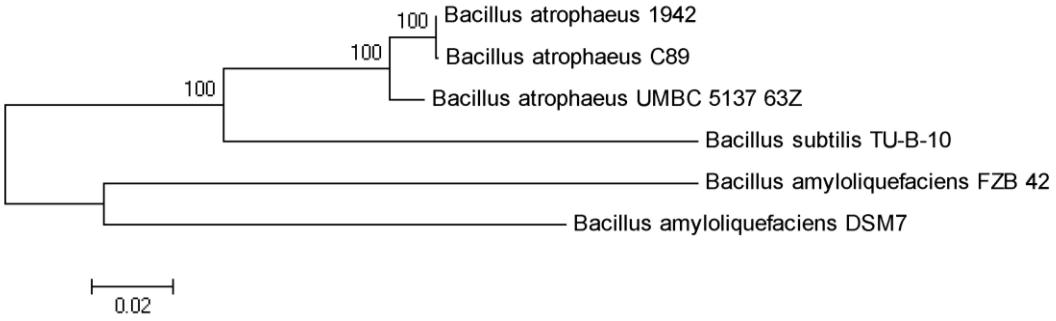


Figure 2.7: Gene tree of myc synthetase of *B. atrophaeus* UCMB 5137 (63Z) strain.
Constructed using the Maximum-Likelihood method with 1000 bootstrap replicates..

The NRPS trees constructed indicate that the genes were seldom-exchanged horizontally. There were however branch length differences seen which shows the variation in the rate at which evolution is occurring. It was therefore evident that myc synthetase is more conserved within *B. atrophaeus* however the plp/ feng synthetase within the test strain is less conserved within the *B. atrophaeus* group which could be an indication of an adaptive evolution of this gene within the test strain.

2.5.5. Linker and Terminals Region Investigation:

Overall:

Linker regions occur between the modules, referred further to as terminals, as well as between the domains within the modules, referred to further as linkers (Figure 1.1). Linker regions have been implicated in the correct functioning of the end NRP product, with disruption or removal of the linker regions resulting in disruption or abolishing of function of the NRP. Linker regions were thought to be important in the correct positioning of the active sites of NRPS modules and influence the synthetic performance of the NRPS (Lai *et al*., 2006; Hahn and Stachelhaus, 2004; O'Connor *et al.,* 2003; Gokhale *et al.,* 1999).

Linker regions were extracted from the NRPS genes and the level of conservation within the homologous genes and between the different NRPS genes investigated. Some of the domain regions were found to overlap creating a slight

confusion for where the one domain region ended and where the next actually began. Linker regions were considered to be areas, which did not fall in the areas of sequence identified to contain a NRPS domain. In general the C-terminals were found to be shorter than the N-terminals and the linkers were found to be shorter on a whole than the terminal regions.

Linker Regions:

In most cases there were no linker regions found between the A and C-domains and in most cases the A and C-domains were found to have overlapping coordinates. There were a few occasions were linker regions were found between the C and A-domains however the majority of the linker regions occurred between the other domains (e.g. between A and T or E etc.).

Within Test Strain:

Analysis of all the linker regions within the test strain collected between the domains of each of the NRPS modules returned several identified motifs from MEME (Bailey *et al.*, 2009). When considering all the different linkers between all the different domains of the NRPS modules identified in the test strain with MEME motifs were found in each of the NRPSs as a whole but not necessarily each module or between each domain (Appendix 1.4).

In comparison to neighboring strains:

Linker regions and terminal between the following NRPS domains; bcb synthetase, plp/ feng synthetase and myc synthetase were investigated (Appendix 1.5-7 and Appendix 1.9-1.11).

Terminal Regions:

The terminal regions presented a slight problem in that; the terminal regions on the ends of each module before the next module began could be biased for modules at the beginning and end of a complete NRPS group. This is due to the fact the gene identified to contain a NRPS module was run through the SBSPKS webserver to determine the location of the domains and by inverse the terminals and linkers, however, areas before the gene and after the gene would not be

considered as only the gene was parsed and any terminal region overlapping a previous or following gene would be ignored. This would possibly be affected if the boundaries of genes were questionable and had not been correctly annotated (Appendix 1.8).

Motif searches conducted on the linker and terminal regions of the NRPS modules showed the amount of conservation within these regions. This could possibly be instrumental in the functioning of the assembly of the NRPs. It is not yet evident to what extent the linker regions are involved however it is clear that these regions are very conserved between species as well as within the strain. This amount of conservation could be indicative of the importance of the linker regions and its possible role in the assembly of NRPs.

2.6. Conclusions:

The test strain *Bacillus* atrophaeus UCMB-5137 (63Z) was found to contain 5 NRPS; myc synthetase, srf synthetase, bcb synthetase, bct synthetase and plp/ feng synthetase. Not all the NRPS were assembled and some modules were missed, that limited our possibility for the whole genome comparison. In a concurrent HPLS analysis of secondary metabolites synthesized by *B. atrophaeus* 63Z it was found that this strain synthesizes all above mentioned polypeptides, however, some minor variations in structures of these metabolites compared to known homolog polypeptides are quite possible (personal communication of yet unpublished data).

NRPS genes are known to be fast evolving and it was therefore of interest to determine whether the genes were horizontally transferred and thereby contributing to the evolution and the rate at which substitutions occur between different NRPS genes. This was investigated through the construction of a species phylogenetic tree using a concatenated alignment of multiple homologous protein sequences. This species tree was then compared with gene trees constructed from individual NRPS genes. The neighboring organisms to the strain were from the following groups; *B. atrophaeus, B. subtilis* and *B.*

amyloliquifaciens. From these neighboring organism the following strains were found to be; *B. atrophaeus* 1942*, B. subtilis subsp subtilis* 168, *B. subtilis* BS5N and *B. amyloliquifaciens* FZB42.

Topologies of the trees were found to be similar and the hypothesis that the genes were transferred horizontally was therefore rejected. The branch lengths where however found to be significantly different which led to the hypothesis that different NRPS genes are under different adaptive evolutionary pressures.

The NRPS modules contained within the test strain were compared to neighboring strains identified during the phylogenetic study. Some of the NRPS modules found within the test strain in comparison to benchmark strains in the database were shorter and not all the modules were found within the strain. Of the NRPSs found within the test strain there do not appear to be any differences or rearrangements to that of the sequences recorded in the database and in comparison to the neighboring strain. The only slight difference noted was that in the *B. atrophaeus* and *B. amyloliquifaciens* strains tested contained Orn and Tyr in the L-conformation for feng in comparison to other strains tested and the database where they appeared in the D-conformation. The NRPS from different strains were found to encode the same products with only minor changes seen in the number of amino acids or their conformational form (L or D conformation).

Motif searches conducted on the linker and terminal regions of the NRPS modules showed the amount of conservation within these regions. This could possibly be instrumental in the functioning of the assembly of the NRPs. It is not yet evident to what extent the linker regions are involved however it is clear that these regions are very conserved between species as well as within the strain. This amount of conservation could be indicative of the importance of the linker regions and its possible role in the assembly of NRPs.

The structure of the synthetase was investigated further in Chapter 3. Homology modeling was carried out and assessment of the A-domains in relation to

conservation and substrate interacting residues within the A-domain binding pocket were investigated.

THIRD CHAPTER: STRUCTURAL ANALYSIS

3.1. Overview

In this Chapter, the structure of the NRP domains of the test strain, *B. atrophaeus* UCMB 5137 (63Z), was investigated. Homology modeling was carried out to construct models of some of the full modules of the NRPSs plp/feng synthetase and myc synthetase as well as the TE-domains and A-domains of the individual modules. The conservation of the A-domain of the individual modules was investigated in relation to the substrate for which they code attachment in the final NRP product. Binding pockets of A-domains were analyzed to determine the role of individual binding pocket residues for substrate specificity.

3.2. Introduction

Multi-domain enzymes are responsible for repetitive catalysis of chemical reactions producing secondary metabolites (Stein, 2005). The order of the NRPS modules is an indication of the final peptide's amino acid sequence (Sieber & Marahiel, 2003). Four interactions are essential in the functioning of NRPSs; 1) structural arrangement of the different domains within modules, 2) role of linker regions connecting the domain, 3) structural control of interaction order, and 4) associations between subunits within the complex (Weissman and Müller, 2008). Domains related to each module perform specific reactions that allow for the prediction of module specificity through domain specificity (Starcevic *et al.*, 2008). The A-domain is of particular interest as it is the domain, which specifies the amino acid that binds and is added to the growing NRP product (refer to Chapter 1 Section 4.1). Therefore when looking for a possible target for modification in the development of novel natural products/ antibiotics in the form of NRPs, the A-domain could be an area of interest. The TE-domain is

involved in the release of the NRP from the NRPS. It is therefore a domain of interest in the investigation of the specificity of the TE-domain to the NRPS.

3.2.1. A-domain Binding Pocket

Stachelhaus et al., (1999) determined the signature sequences within the A-domain, which interacted with the phenylalanine in the substrate-binding pocket. This was used to determine the core interacting amino acids by matching up core motifs within the domain. The interacting residues within the binding pocket were found to dictate the shape of the pocket as well as to be involved in the recognition of the side chain of the incoming substrate (Stachelhaus *et al*., 1999). From the interacting residues within the binding pocket they speculated that the basic amino acid (His322) was involved in an interaction with the acid side chain and that the Leu236 was responsible for the closing of the pocket in the presence of larger substrates. In cases where larger basic side chains are being interacted with they speculated that acidic residues such as Glu278 and Asp331 would be more involved in the recognition process. For a hydrophobic substrate such as Val it is thought that the activation involves the whole pocket made up of hydrophobic residues wherein the binding pocket is kept open by the alanine residues at the top of the pocket and bulky residues at the bottom (Stachelhaus *et al*., 1999).

Stachelhaus *et al*., (1999) found that the activation of small amino acids involves binding pockets which are more flexible towards the bottom of the pocket while larger amino acids are activated within binding pockets that are more flexible within the top of the binding pocket. This work was carried out on the incoming Phe within the binding pocket of gramicidin synthetase. They also identified motifs within the binding pocket of the A-domain, which were used to check the alignment between differing substrates. The motifs are found near the interacting residues within the binding pocket (Figure 3.1). The following amino acids were found to be the most flexible in amino acid usage; 239, 278, 299, 322 and 331 (Stachelhaus *et al*., 1999). These residues as well as the identified motifs in the area were used in this study in a structural alignment using PROMALS3D.

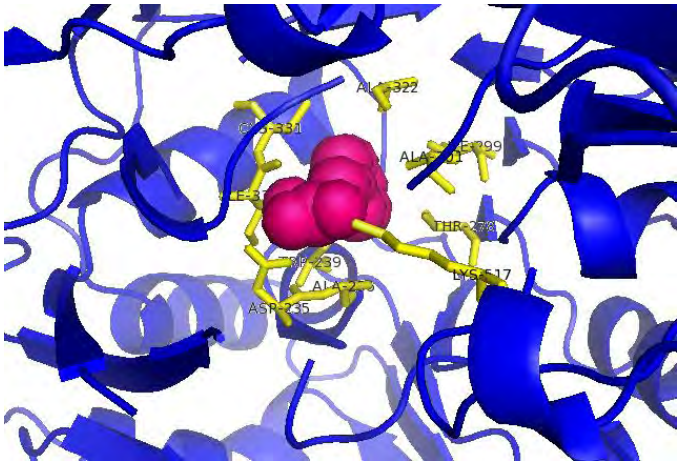


Figure 3.1: The interactions within the A-domain of gramicidin synthetase (PDB ID: 1AMU) (blue) and the incoming amino acid, Phenylalanine (magenta). Interacting residues are shown in yellow.

PROMALS3D is a profile multiple sequence alignment program which uses predicted local structures and local constraints to perform a multiple sequence alignment of either protein sequences or structures. Data such as secondary structure prediction and 3D structures are used to enhance the overall result through performing a structural alignment of the sequences/ structures (Pei & Grishin, 2007). A fast initial alignment is performed using a Blosum62 scoring matrix, which is followed by a PSI-BLAST search to determine homologs and a secondary structure prediction from UNIREF90 and PSIPRED respectively. A hidden Markov model (HMM) is then used to determine probabilities of the matched residues. The probabilities determined, homologs containing 3D structures and any user-defined constraints are then applied to determine the consistency of the scores. Each sequence is then progressively aligned and combined into a final multiple sequence alignment and is used to show an alignment using structural elements due to the structure being more conserved than the sequence (Pei & Grishin, 2007).

3.2.2. TE-domains:

The TE-domains are contained within the final module/ at the C-terminal of the NRPS and are involved in product release (Chapter 1 Section 4.1). The TE-domains contain a catalytic triad consisting of Ser-His-Asp contained within the middle of the substrate channel and belong to the serine hydrolase superfamily (Du & Lou, 2010). The TE-domain is known to form a hydrophobic bowl-shaped

pocket wherein the NRP is folded into a cyclic structure. It has been found in two different conformations depending on the state of the enzyme (open/ closed). Du and Lou (2010) also found that the folding of the NRP within the TE-domain is reliant on specific interactions with residues within the lining of the active site pocket.

3.2.3. Homology Modeling:

Reasoning and motivation:

The mechanism in which non-ribosomal peptides are assembled by non-ribosomal peptide synthetases is not yet fully understood. The 3D structure of the A-domain of the NRPS could give some insight in to the binding pocket of the incoming substrate as well as the specificity of substrate and the residues involved in its interaction. Structural biology and the analysis of a proteins structure are usually done with the intention of linking the structure to the function of the protein.

The test strain was found to have increased phytopathogen activity in comparison to other strains in laboratory tests (Personal correspondence). The NRPs assembled by the NRPSs encoded in the strain are thought to be the root of this phytopathogenic activity. The A-domain structure of these NRPSs will give more information on the active site to where the amino acids of the NRP product bind as well as the residues involved in this interaction.

Homology modeling is an important technique that can be used to predict the 3D structure of a protein by using homolog proteins of known structure. Known structures refer to those determined previously via physical means such as x-ray crystallography and nuclear magnetic resonance (NMR). Homology modeling is however a completely computation process and is therefore not limited by the same constraints as physical experiments. There are reasons why certain protein structures are not yet available, such as the actual protein cannot be sufficiently purified or crystalized thereby making the protein unable to be rendered using methods such as x-ray crystallography. The method is therefore very useful for

predicting the tertiary structure of a protein when its structure has not yet been determined experimentally. The process is reliant on the fact that the structure of a protein is usually very well conserved in comparison to that of the sequence (Xiong, 2006; Tastan Bishop et al., 2008).

Steps Involved:

The main steps of homology modeling are; 1) template identification, 2) alignment, 3) model building and refinement and 4) validation. Step 1 identifies homologues to the queried protein of unknown structure, which are used as templates for the modeling process. Important elements to consider in selecting an ideal template are high resolution, high percentage identity, and an available structure with the required ligands. During the second step the template sequence and the target sequence of unknown structure are aligned to ensure accuracy of the final model. Model building and refinement are the third step in the process and involves the building of the target model using the chosen template. This is done using programs such as Modeller and SwissModel. Loop areas are the most difficult areas within the structures to model, which can usually be improved by loop modeling servers or by using energy minimization techniques. The final step within the homology modeling process is validation. This step involves the checking of the structure in relation to known structures to access the correctness of the predicted structure. Popular validation programs include; Verify3D, ANOLEA, Prosa and MetaMQAP (Xiong, 2006; Bishop et al., 2008).

Template Identification and Modeling:

HHpred is used in protein homology detection and structure prediction. This is done through searching many different databases such as PDB, SCOP and Pfam. The input can be either in the form of a queried sequence or a multiple sequence alignment. It offers the choice of carrying out either a global or a local alignment. HHpred is also connected to Modeller (Šali *et al.*, 1995) making it possible to generate 3D structural models with the homologs detected by HHpred on the same webserver interface (MPI Toolkit). HHpred is able to function equally well with both single and multi-domain query sequences. It has been deemed one of

the most sensitive severs for the purposes of homology modeling through its high performances in the annual CASP competitions (Söding *et al.*, 2005).

HHpred makes use of a PSI-BLAST search within the first step of the alignment process in search of homologs to the queried sequence. Sequence identity, E-value and coverage are some of the many parameters, which can be altered or specified in the searches on the server. Finally the sequence is annotated with PSIPRED derived confidence values and the predicted secondary structure. A hidden Markov model (HMM) is then derived from the multiple sequence alignment generated, which is used to calculate the probabilities of the viewed residues occurring at the particular positions. Using this model and position-specific gap penalties the database is searched and matches scored to find appropriate homologs (Söding et al., 2005). HHpred was ideal for the purposes of this project as it has a short response times while still maintaining a high model accuracy (Söding et al., 2005).

When choosing a template on which to model the target protein of unknown structure the main characteristics to take into account are the sequence identity, the quality of the template and the amount of coverage of the target sequence by the template structure (di Luccio and Koehl, 2011).

Available Structural Information:

Structures available for the different NRPSs within the Protein Data Bank (PDB) are shown in Table 3.1. There are no structures for myc synthetase or bct synthetase within the PDB. There are two structures for feng synthetase; for feng synthetase thioesterase domain cluster (PDB ID: 2CB9) as well as the feng thioesterase domain inhibited by PMSF (PDB ID: 2CBG), both of which are from *B. subtilis.* There are four structures for bcb, all from *B. subtilis*; trisatecholate sideophore binding protein FeuA (PDB ID: 2W18), FeuA complexed with ferri-bcb (PDB ID: 2WHY), ferri-enterobactin (PDB ID: 2XUZ) and ferri MECAM (PDB ID: 2XV1), all within *B. subtilis*. There are six recorded structures for srf synthetase within the PDB; the external thioesterase of srf synthetase (PDB ID:

2RON), external thioesterase of srf synthetase in complex with a carrier domain (PDB ID: 2K2Q), the structural basis for cyclization of lipopeptide antibiotic srf by thioesterase domain SrfTE (PDB ID: 1JMK), termination module of srf subunit A (PDB ID: 2VSQ), srf in micellar media (PDB ID: 2NPV) and the 4'phosphopantetheinyl transferase SFP-co-enzyme A complex (PDB ID: 1QR0).

Table 3.1: Available crystal structures in the PDB of NRPS domains and subunits for the *B. subtilis* group. 1AMU is a key structure used in this study. It is not from the *B. subtilis* group referred to in this study but does form part of the group on a broader scale and is a crystal structure from *Brevibacillus brevis*.

Crystal Structure	PDB ID	Reference
Srf synthetase thioesterase subunit	2K2Q, 2RON, 1JMK	(Koglin *et al*., 2008)
Recombinant fragment of the NRPS synthetase fenb thioesterase domain (residues 1043-1274)	2CB9, 2CGB	(Samel *et al*., 2006)
Termination module of srf a biosynthesis cluster (residues 1-1009,1015-1274)	2SVQ	(Tanovic *et al.,* 2008)
Dhbe in complex with dhb and amp	1MD9	(May *et al*., 2002)
Dhbe in complex with dhb-adenylate	1MDB	(May *et al.,* 2002)
Dhbe in absence of substrate	1MDF	(May *et al.,* 2002)
D-alanine-poly(phosphoribitol) ligase subunit 1 (A-domain of NRPS)	3E7W	(Yonus *et al*., 2008)
Phe A-domain of gramicidin synthetase 1 in a complex with amp and Phe	1AMU	(Conti *et al.,* 1997)

The quality of a model determined by the crystallographic data is termed the R-value. This is a measure of the match between what is seen experimentally when the structure is solved in relation to the simulated diffraction pattern determined beforehand. The R-free value is used to reduce the bias introduced during the refinement process. This is done by removing 10% of the observations before refinement and which is then used to determine the accuracy of the model in the prediction of the remaining 10%. A good quality structure should therefore have R-value and R-free values, which are close to each other (Kleywegt and Jones, 1997).

Validation:

The evaluation and validation of models is a vital step in homology modeling/ structure prediction. There are many validation methods available and it is usually recommended to use a combination of different methods to account for any bias or inaccuracy in any one particular program. This study used three validation methods; Verify3D, ANOLEA and MetaMQAPII.

Verify 3D and ANOLEA each produce evaluations on every individual residue. Verify3D functions through the assignment of environments to the proteins and thereby the determination of a score of the amount to which the environment in which the protein is found is reliable or likely in relation to each amino acid (Bowie *et al.,* 1991). The score generated is then averaged over a long stretch of 21 amino acids (Liithy et al., 1992). Verify3D has been found to only detect major errors within the structure, which are generally related to misalignments while not detecting non-physical errors such as errors in bong length, angles and steric clashes (Pawlowski et al., 2008). ANOLEA works using a distance-dependent formulae that calculates the potential through the non-local environment (NLE) of the heavy atom within the structure and thereby recognizes errors, which are not usually identified by Verify3D (Melo & Feytmans, 1998).

MetaMQAPII is a model quality assessment program (MQAP) that uses a multivariate regression model thereby controlling minor parameters which are often ignored in other MQAPs. It is able to predict the deviation of the C-atoms of the model in relation to that of known structures on an individual as well as global level, which are then averaged using a 5-residue window. MetaMAQPII displays the potential errors within a predicted model through coloring techniques wherein the correctness is represented on the spectrum from blue to red with blue being correct and red representing areas of concern where the structure is less accurate (Pawlowski *et al*., 2008). This is done through the incorporation of techniques from eight other validation programs; Verify3D,

ProSA, ANOLEA, BALA-SNAPP, TUNE, REFINER and PROQRES (Melo & Feytmans, 1998).

MetaMQAP has been found to be superior as other MQAPs such as Verify3D since it uses the results of 8 other MQAP methods (including; Verfiy3D, PROSA, BALA, ANOLEA, PROVE, PQORES, REFINER and TUNE) in its prediction of deviation of local residues from known structures. It is usually however recommended to use more than one validation method/ program to validate the results obtained.

3.2.4. Motif Identification:

To determine the amount of conservation specifically within the A-domains of the modules of the NRPSs the sequences were analyzed for conserved areas/ motifs. MEME is a motif search program, which searches for several different motifs within a sequence (Bailey *et al*., 2009). Motifs found in a set of sequences show that the subsequence does not occur by chance and could indicate a shared biological function between the sequences (Bailey & Elkan, 1994).

3.3. Methodology

3.3.1. Homology modeling

Full Modules:

Template Identification and Modeling:

HHpred was used to determine the known structure with the best secondary structure match to that of the target being modeled which was then used as the template against which the unknown structure could be modeled against (Söding et al., 2005) using the default parameters. PIR files were generated by HHpred (Šali et al., 1995) and edited manually to remove areas for which templates could not be found. This was generally in areas where multiple templates were used for generating different areas of one structure where overlapping areas were removed and in the terminal regions of the structures. Structures were generated using Modeller on the MPI Toolkit (Biegert *et al.,* 2006).

Structures of full modules were generated for plp/ feng synthetase modules 6 and 9 using HHpred and Modeller on the MPI Toolkit (Biegert *et al.,* 2006). Templates used for the modeling can be seen in Table 3.2 and in Figure 3.2. Of the two structures built, three templates were used two of which were 2VSQ. 2VSQ is the third subunit of srf synthetase. This shows that the sequences within the two synthetase modules are highly conserved within the synthetase.

2VSQ is the C-terminal module of srf synthetase of *B. subtils* it has a resolution of 2.6 Å and an R-value of 0.215 and R-free value if 0.272. This template has a good resolution and the quality of the structure is quite good as the R-fee value is very similar to the R-value. 2XHG is the E-domain of tyrocidine initiation module of *Brevibacillus brevis*, which has a very good high resolution of 1.5 Å with an R-value and R-free value of 0.16 and 0.179 respectively. This template has a very good, high resolution and is of a high quality with a low R-value and a R-free value very close to that of the R-value.

T**able 3.2: Templates used for homology modeling obtained from PDB**. Areas of models constructed with 2VSQ as the template are shown in green. Lengths of the template sequences are shown in brackets in the template column.

Synthetase	Module	Template(s)		Query range	Resolution (Å)	E-value	Identity (%)
Plipstatin/ Fengycin synthetase	6	2VSQ	244-1044(1304)	1-800	2.6	3.0e-114	37
		2XHG	9-457 (466)	802-1243	1.5	1.6-36	40
	9	2VSQ	1-1272(1304)	1-1266	2.6	1.0e-164	39

Validation:

Models constructed for full modules of plp/ feng synthetase modules 6 and 9 were validated using MetaMQAPII (Figure 3.3). Domain locations, determined in Chapter 2, were used to visualize the domains of the constructed models in PyMOL by coloring the domains and the linker regions (Figure 3.3).

TE-domains:

Template Identification and Modeling:

The TE-domains are only found on the final module of the synthetase and are involved in the release of the final NRP product. The structures of the TE-domains of plp/ feng synthetase and myc synthetase from the tests strain were constructed using HHPred and Modeller on the Bioinformatics Toolkit (Figures 3.5). These structures were then superimposed and compared.

The TE-domains of myc synthetase and plp/ feng synthetase were modeled using the templates shown in Table 3.3. 3FLA is the type II TE-domain of RifR from *Amycolatopsis mediterranei* and has a good, high resolution of 1.8 Å and consists of two chains, however, only the first chain was used as a template. 3FLA has an R-value of 0.174 and an R-free value of 0.202. The templates resolution is high, with a low R-value and an R-free value close to that of the R-value thereby indicating it is a good choice for a template. There were no residues found matching in the template for the first 13 residues of the myc synthetase TE-domain. 3QMV is the TE-domain of RedJ from *Streptomyces coelicolor*, with a resolution of 2.12 Å and R-value and R-free values of 0.203 and 0.249 respectively. This templates resolution is slightly lower than the previous template but it is still an acceptable resolution with an R-value within the normal accepted range and an R-free value close to that of the R-value (Kleywegt and Jones, 1997). 3QMV consists of four chains but only the first was used in modeling the TE-domain of the plp/feng TE-domain.

Table 3.3: Templates used to model the TE domains of myc and plp/ feng synthetases.

Synthetase	Subunit	Module	Template		Query Range	Resolution (Å)	E-value	Identity (%)
Myc synthetase	C	2	3FLA	13-163 (267)	13-154	1.8	1.0e-13	18
Plp/ Feng synthetase	1	10	3QMV	53-265 (280)	2-186	2.12	5.1e-28	16

Models were superimposed using PyMOL to demonstrate the conservation between the TE-domains of the two different synthetases (Figures 3.5).

Validation:

Models constructed for TE-domains were validated using MetaMQAPII (Pawlowski *et al.,* 2008) and visualized in PyMOL (The PyMOL Molecular Graphics System) (Figure 3.5). Templates used to construct the models were validated by MetaMQAPII (Figure 3.2).

A-domain structures:

Template Identification and Modeling:

Since templates are not currently available for building 3D structures of the full modules of these NRPSs, the A-domains, which specify the amino acid incorporated, where templates could be found, were therefore investigated. Models were built for the following modules and domains; plp/ feng synthetase A-domains from the following modules; 2, 3, 5, 6, 7, 8, 9, 10, myc synthetase; subunit A modules 1 and 2, subunit B modules 1 and 3, and subunit C modules 2 (Figures 3.6-7). Templates used in the modeling of these A-domains can be seen in Table 3.4. Templates were identified using HHpred and used to build the models with Modeller on the MPI Toolkit. Areas for which adequate templates could not be found were removed before modeling the domains. These were usually areas at the terminals of the structures.

3KXW is a fatty acid AMP ligase *Legionella pneumophilia subsp. Pneumophila str.* Philadelphia with a good high resolution of 1.85 Å, an R-value of 0.192 and a R-free value of 0.230. This template has a good, high resolution with a low R-value and an R-free value very close to that of the R-value making it a favorable template to be chosen. 2VSQ is the C-terminal module of Srf synthetase of *B. subtils* it has a resolution of 2.6 Å, R-value of 0.215 and a 0.272 R-free value. With an acceptable high resolution, low R-value and an R-free value close to that of the R-value this is a good template to use. The resolution could be higher and would thereby have produced a better quality target model but is still within the acceptable range. 3TSY is a stilbene synthase fusion protein from *Arabidopsis*

thallana with a low resolution of 3.1 Å and R-value and R-free value of 0.179 of 0.208 respectively. This template has a very low resolution but is of a high quality shown by the low R-value and the closeness of the R-value and R-free value. 1AMU is the Phe activating domain of gramicidin synthetase of *Brevibacillus brevis* consisting of two identical chains, only one of which was used, with a good high resolution of 1.9 Å, an R-value of 0.213 and a R-free value of 0.246. This template therefore has a high resolution and is of a good quality with low R-value and a small range between the R-value and R-free value. 3R44 is the mycobacterium tuberculosis fatty acyl CoA synthetase with a favorable high resolution of 1.8 Å and a low, favorable R-value of 0.164 with a very close R-free value of 0.199.

Table 3.4: Templates used to construct 3D models of the A-domains myc synthetase and feng synthetase. Models constructed using the same templates are shown in corresponding colors. All models build using 2VSQ as a template is highlighted in green, all models build using 1AMU as the template is highlighted in purple and all models built using 3R44 as the template are highlighted in blue.

Synthetase	Sub-unit	Module	Template(s)		Query Range	Resolution (Å)	E-value	Identity (%)
Mycosubtilin Synthetase	A	1	3KXW	9-550 (590)	5-537	1.85	7.9e-64	28
		2	2VSQ	424-939 (1304)	1-514	2.6	2.1e-69	33
	B	1	2VSQ	421-939 (1304)	1-516	2.6	2.3e-73	35
		3	2VSQ	421-939 (1304)	1-528	2.6	1.1e-69	36
	C	2	2VSQ	5-1278(1304)	18-1292	2.6	3.9e-69	35
Plipstatin/ Fengycin synthetase	1	2	1AMU	18-527 (563)	1-525	1.9	7.6e-69	37
		3	3R44	1-511 (517)	18-510	1.8	8.5e-83	23
		5	3R44	11-375 (517)	19-383	1.8	1.7e-60	20
				17-375 (517)	1-359	1.8	3.4e-62	20
		6	3R44	13-514 (517)	20-518	1.8	5.6e-84	22
		7A	2VSQ	429-954 (1304)	1-528	2.6	4.4e-70	35
		7B	3R44	98-513 (517)	29-421	1.8	6.5e-73	23
		8	3R44	16-460 (517)	24-470	1.8	3.0E-74	23
		9	1AMU	17-526 (563)	1-511	1.9	1.7e-72	39
		10	3R44	16-510 (517)	23-510	1.8	6.0E-80	23

Validation:

Models constructed for A-domains were validated using Verify3D (Bowie *et al.,* 1991, Roland *et al.,* 1992) (Appendix 1.12), ANOLEA (Melo *et al.,* 1997) (Appendix 1.12) and MetaMQAPII (Pawlowski *et al.,* 2008) and visualized in PyMOL (The PyMOL Molecular Graphics System) (Figure 3.6-7). Templates used to construct the models were validated by MetaMQAPII (Figure 3.2).

3.3.2. A-domain Motif Analysis:

A-domain sequences were compared in a sequence alignment and a motif search using MEME (Bailey and Elkan, 1994, Bailey and Gribskov, 1998). This was also carried out for A-domains, which coded for the same amino acid. Glu is contained 3 times within plp/ feng synthetase and Asn three times within myc synthetase. Tyr is contained twice in plp/ feng synthetase and once in myc synthetase. Pro is contained once in each of the NRPs, plp/ feng synthetase and myc synthetase. These similar domains were also superimposed using PyMOL to view their correlation and overlap (Tables 3.5-10) (Figure 3.8-9).

3.3.3. A-domain Substrate Pocket:

Stachelhaus *et al.,* (1999) determined the interacting residues within the binding pocket through examination of the crystal structure of the A-domain in complex with Phe. This was then used to determine the motifs surrounding the active site residues. These motifs were then used in a sequence alignment to determine the interacting residues within other A-domains. This procedure was applied in a reverse manner in this study. The motifs identified by Satchelhaus *et al.,* (1999) were used to identify the interacting residues within the some of the A-domains of the test strain which was then compared through superimposition of the crystal structure determined by Stachelhaues *et al*., (1999) with the A-domain models constructed in this research project.

A multiple sequence alignment of the A-domains, which coded for the same amino acids, was performed using PROMALS3D and sequences were viewed and

annotated using JalView. This was performed on the A-domains which code for Tyr; myc synthetase subunit B module 1, plp/feng modules 3 and 9 (Figure 3.10) and those which code for Glu (with the exception of plp/feng synthetase module 1 as no structure was successfully constructed for this A-domain); myc synthetase subunit B module 3, plp/ feng synthetase modules 5 (a and b) and 8 (Figure 3.12). The closest match in each case then mapped to the structure in PyMOL; plp/feng synthetase module 9 (Figure 3.11) and myc synthetase subunit B module 3 (Figure 3.13) to visualize the interacting residues within the A-domain binding pocket.

3.4. Results and Discussion

3.4.1. Homology modeling

Full Modules:

Template Identification and Modeling:

Initially we attempted to build structures of the full modules of myc synthetase and plp/ feng synthetase. There were in most cases very few reliable templates available for the module sequences. In many cases the only templates matching the queried module sequence had very low identities and high E-values. These fragments were only covering small fragments of the sequence and therefore only a few structures of the full modules could be generated. Full module structures were only built for plp/feng synthetase modules 6 and 9. Templates used to construct the models were validated using MetaMQAPII (Figure 3.2).

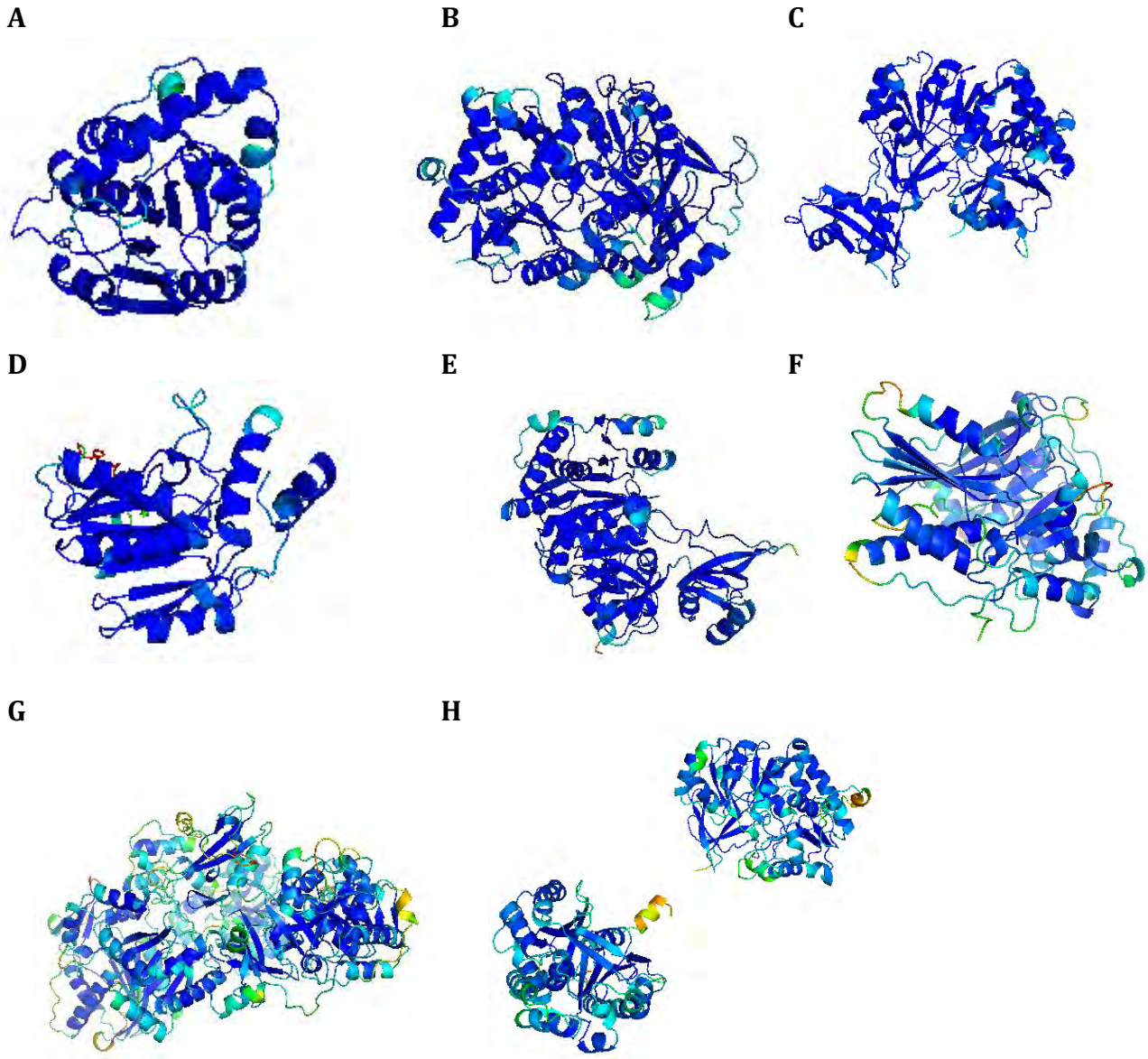


Figure 3.2: Templates used for homology modeling construction colored according to MetaMQAPII scores. Template names; A) 3FLA, B) 3KXW, C) 3R44, D) 3QMV, E) 1AMU, F) 2XHG, G) 2VSQ, and H) 3TSY. Areas shown in blue are calculated by MetaMQAPII to be correct decreasing in reliability of the model as the color moves through the spectrum towards red.

The models constructed of the full modules were colored according to the domains of the module as predicted by the SBSPKS webserver (Chapter 2) (Figures 3.3). C-domains were colored in pink, A-domains in blue/ cyan, T-domains purple, E-domains in green and all linker and terminal regions were shown in yellow. In most cases models generated did not contain the last few residues of the E-domains and the C-terminal regions. This was either due to there being no C-terminal region detected in the sequence submitted to the SBSPKS webserver or there not being templates available for those regions. Linker regions are in each case found to be quite short regions, usually ranging

from 1 residue to around 50 residues, in comparison to the terminal regions which were significantly longer often around 200 residues or more in length, when present. This could possibly be due to the location of the module within the gene when submitted to the SBSPKS webserver thereby not taking into account regions further beyond those regions as well the terminal regions being defined as the sequence flanking the domains of the module. Therefore the further along the module is within the gene the longer the terminal region will appear. In the Plp/feng module 6 constructed (Figure 3.3a) there was no C-domain detected by the SBSPKS webserver. The largest regions in each case were found to be the A-domains followed by the E-domains. The linker regions and T-domains are quite short in length.

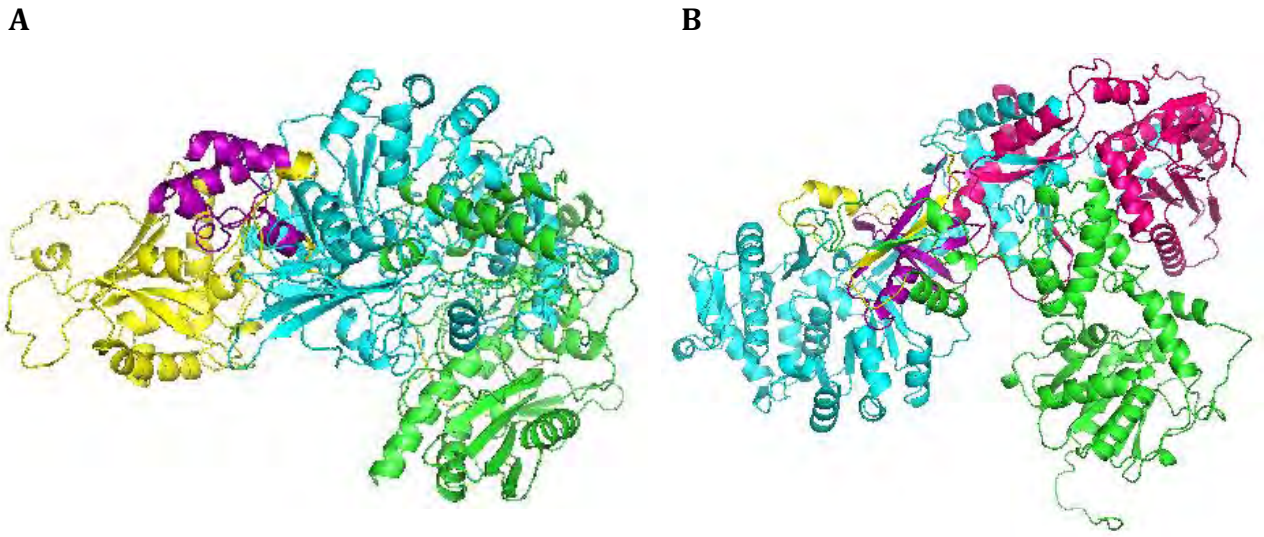


Figure 3.3: Plp/feng full module models colored according to domains. A) Module 6, B) Module 9. C-domain - pink, A-domain - cyan, T-domain - purple, E-domain - green, Linker and terminal regions – yellow.

Within the feng 6 module model the A-domain appears within the center of the structure and the structure appears to curve in a moon shape with the A-domain at the curve of the structure. However, with module 9 the C-domain seems to be more centered within the structure and at the height of the curved shape.

Validation:

The plp/feng synthetase module 6 model when validated using MetaMQAPII was found to have a few areas highlighted in red indicating there may be areas within

the model which are considered incorrect (Figure 3.4a). These regions were in most cases found to be between 1 to 3 residues in length and contained within a loop region of the structure. There was one region within a helix of 5 residues/ inserts in length at 1045 to 1049 residues where the template used did contained a deletion/ gap and thereby did not provide template on which the target could be modeled against. This shows why this area was not seen as reliable as was the remainder of the structure determined using the template structure available. This area was within the E-domain of the module. The structure was generated using two templates; 2VSQ for residues 1 to 800 and 2XHG for residues 802 to 1243 (Figure 3.2 f and g). Both templates were relatively good with acceptable to good resolutions. Each did however have regions within their structures where they were considered by MetaMQAPII to be inaccurate. These areas are shown within Figure 3.2 in areas shown towards the red end of the color spectrum. None of the areas were shown as red thereby indicating there were no major areas within the model that were inaccurate. The area however in plp/ feng synthetase module 6 between 1045 to 1049 shown to be incorrect was modeled using the second template, 2XHG and this area within the template was not shown by MetaMQAPII to be an area of concern within the templates structure. This concern within the module 6 structure could therefore be attributed to the low identity within that area of the alignment between the template and query sequences. The identities were not very high however they were above 30% for the whole structure.

This was repeated with the full module 9 of plp/ feng synthetase (Figure 3.4b). There was only one residue found to be within the red region my MetaMQAPII at residue 1216 within a loop region of the E-domain. The remainder of the structure seems rather reliable and is mostly shown in a dark blue. This structure was modeled using only one template, 2VSQ. The template did show a few areas for concern by MetaMQAPII (Figure 3.2 g) shown as areas in yellow. The loop shown as a concerning are within module 9 on the template used to model the structure was also shown by MetaMQAPII as a possible incorrect region of the structure which could explain the incorrectness within the module 9 structure.

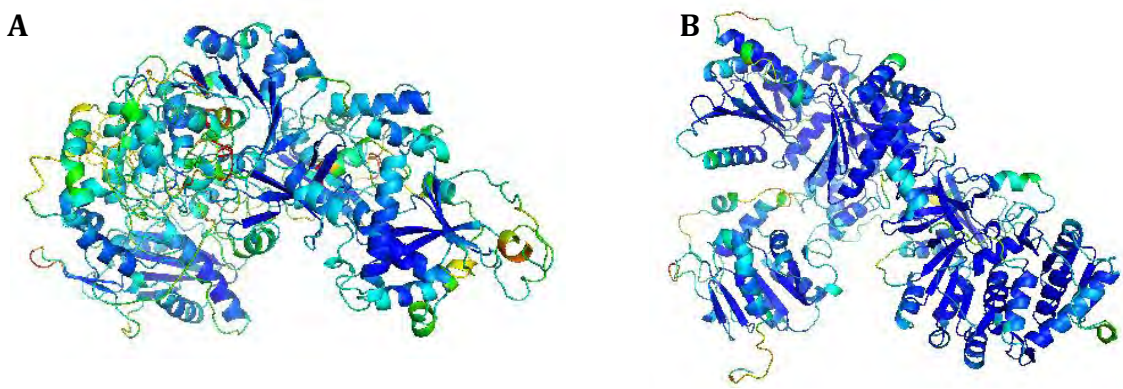


Figure 3.4: Homology model of Plp/ Feng synthetase full modules colored using MetaMQAPII scores. A) Module 6, B) Module 9. Areas shown in blue are calculated by MetaMQAPII to be correct decreasing in reliability of the model as the color moves through the spectrum towards red.

TE-domains:

Template Identification and Modeling:

The TE-domains of plp and myc were modeled using the templates seen in Table 3.3 identified using HHpred. Templates used were validated using MetaMQAPII (Figure 3.2).

Validation:

The plp/feng module 10 contains the TE-domain (Figure 3.5a) and had two areas of two residues each; 37-38 within a loop region and 158-159 within the end of a helix region which were identified by MetaMQAPII as areas for concern. These areas were however not shown as incorrect by MetaMQAPII within the template used to model the structures (Figure 3.2). The only area within the template shown to be of concern were in the first 17 residues which were not used as a template in the modeling of the structure. The second module of myc synthetase subunit C contains the synthetases TE-domain and contained two areas also of two residues each within a loop region that were not confidently modeled; residues 87-88 and 108-110 (Figure 3.5b). The template used to model this structure, 3FLA, however, did not show any areas that were not calculated to be correct with all residues being within the blue are of the color spectrum as calculated by MetaMQAPII (Figure 3.2a). The remainders of the structures were more reliable and shown in blues and greens by MetaMQAPII.

The TE-domains from feng/ plp and myc were superimposed to visualize the overlap between the structures of the two TE-domains (Figure 3.5c). The two TE-domains are similar and have some overlapping areas. There are however more regions which are different between the structures than regions that are the same. This can therefore show that the TE-domains between the different synthetases are different and perhaps specific for the NRP they facilitate release for.

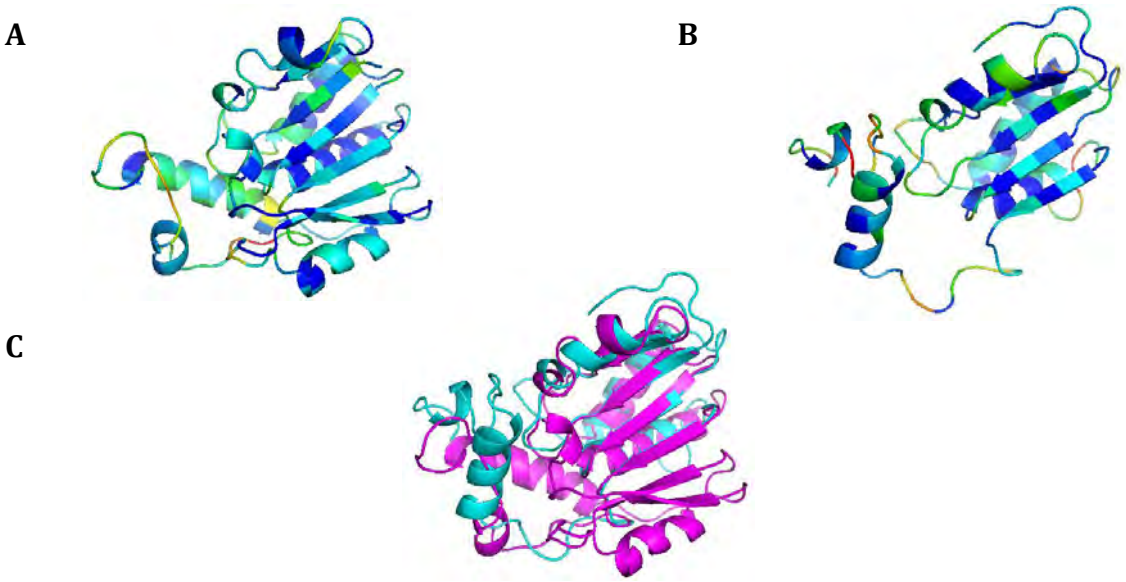


Figure 3.5: TE-domain models generated for A) plp/feng synthetase and B) myc synthetase and C) TE-domains superimposed. Structures in A) and B) are colored according to their scores determined by MetaMQAPII where areas shown in blue are calculated to be correct decreasing in reliability of the model as the color moves through the spectrum towards red. C) Shows the two TE domains of the two different NRPSs superimposed; myc synthetase shown in cyan and plp/ feng synthetase shown in magenta.

A-domain structures:

Template Identification and Modeling:

A-domain structures were constructed for the myc synthetase and plp/feng synthetase modules, with the exception of plp/feng synthetase modules 1 and 4 and myc synthetase subunit B module 2 and 4 and subunit C module 1. The templates used are shown in Table 3.4. Templates used to construct the models were validated using MetaMQAPII (Figure 3.2). Out of the 14 domains modeled only 4 templates were used. 2VSQ was used in the modeling of 5 of these

structures; 3R44 was used as a template in 6 of the constructed domains while 1AMU was used to model 2 of the domains. This showed the conservation between the A-domain sequences.

Validation:

Models constructed were then validated using MetaMQAPII and visualized in PyMOL (Figures 3.6-7). Anolea and Verify3D were also used to evaluate the models constructed (Appendix 1.12) (Eisenberg et al., 1997; Melo et al., 1997).

Plp/ feng synthetase module 2 A-domain structure constructed (Figure 3.6a) has two residues, 527-528 at the terminal region, which were identified by MetaMQAPII as deviating from known protein structures due to there not being an adequate/ appropriate template available. This area was therefore removed from the structure, which improved the overall structure. The rest of the structure is reliable colored in shades of green and blue showing it is in line with known protein conformation. Verify3D did not find any areas, which scored in the very bad region, and the majority of the structure was in the favorable zone. ANOLEA showed a small region around 210 residues to be unfavorable. This was not a region shown by MetaMQAPII to be unfavorable. The template used to model this structure was 1AMU and did not display any incorrect regions by MetaMQAPII other than the first few residues at the N-terminal which were not used in the creation of the model (Figure 3.2e).

Plp/feng synthetase module 3 A-domain model constructed (Figure 3.6b) had a loop at the N-terminal region, residues 1- 9, which did not have a reliable template for construction and was therefore not included in the model creation. The remainder of the structure was reliably constructed being mainly blue in color due to its scores obtained and calculated by MetaMQAPII. Verify3D found three small areas to be unfavorable including an area around residue 80-100 and 185 while ANOLEA found a small area around 305 to be unfavorable (Appendix 1.12) (Eisenberg et al., 1997; Melo et al., 1997). ANOLEA, Verify3D and MetaMQAPII identified differing areas to each other that were considered by the

different programs to indicate an unreliable structure/ model. 3R44 was the template used to model this structure. When this template was analyzed by MetaMQAPII (Figure 3.2c) and found to be a very reliable structure being completely represented in blue. The areas of concern within the plp/feng module 3 A-domain structure are therefore not due to faults within the template and are more likely attributed to the low percentage identity between the template and the target sequences.

Plp/feng synthetase module 5 A-domain models were constructed for both of the sequences identified to contain an A-domain by the SBSPKS (Chapter 2) (Figure 3.6c and d). The template identified to be the closest homolog to the sequences was 3R44 in both cases. This was a very reliable structure as analysed by MetaMQAPII and was therefore used as the template (Figure 3.2c). Both of the A-domain structures constructed had areas identified as a concern. The two structures even though representative of the same A-domain within the module appear very different and more similar to the other A-domains than each other. Verify3D did not indicate any areas of specific concern within the structures however ANOLEA identified several areas within the structure to be of concern within structures 5 a and b (Appendix 1.12). The model of plp/feng synthetase module 5b A-domain was more reliable and had less areas of concern than the module 5a A-domain model.

Plp/feng synthetase module 6 A-domain model (Figure 3.6e) had a few problem areas where the template was not reliable and did not match the target sequence resulting in a red colored region calculated by MetaMQAPII. There was no reliable template for the N-terminal region and it was therefore not used for the modeling of the structure. The following area was not in line with known protein conformation; residues, 147-149 within a loop region. Verify 3D identified an area that was unfavorable within the structure (residues around 150-160) while ANOLEA identified a few more smaller regions at residues around 120, 150, 260, 320 and 460 (Appendix 1.12). These regions seem to mostly correlate between all three programs. The template used in the modeling of this structure, 3R44,

was very reliable and the areas indicated as areas for concern are not due to the template (Figure 3.2c).

For the two models constructed for plp/feng synthetase module 7 (A and B) A-domains (Figure 3.6f and g) the first seems more reliable than the second (B) with less red/ problematic areas within the structure. This could be attributed to the templates used. Two different templates were used in the modeling of these structures; 2VSQ for 7A and 3R44 for 7B. 3R44 is a more reliable/ correct structure identified by validation through MetaMQAPII (Figure 3.2c) whereas 2VSQ had a few areas that MetamQAPII identified as in less correct (Figure 3.2g). The second A-domain of module 7 (7B) (Figure 3.6g) had more problem areas that were not correctly modeled. Residues 49-51, 53-54 and 61 all contained within the large loop were not reliably modeled and represented in red in the structure. Residues 176, 419-421 were contained at the ends of a helix region. These structures were also validated using Verify3D and ANOLEA each finding different areas, which they identified, as unfavorable (Appendix 1.12). Verify3D identified residue 160 to be unfavorable and ANOLEA identified areas around these residues to be of high energy; 220 and 350 in plp/ feng synthetase module 7A A-domain. Plp/feng module 7B A-domain had a small area around residue 120 identified by Verify3D as unfavorable and two small areas around residues 90 and 220 to be high in energy (Eisenberg et al., 1997; Melo et al., 1997). Again these areas identified by the three programs differed.

The A-domain of the 8th module of plp/feng synthetase modeled was mostly reliable with only 2 problematic areas within the N-terminal due to a lack of template and was therefore not included in the model construction (Figure 3.6h). The majority of the structure is however more reliable. ANOLEA and Verify3D were also used to identify areas that were unfavorable or of high energy in the structure (Appendix 1.12). ANOLEA identified three small areas around the following residues to be high energy; 180, 230, 325 while Verify3D identified areas around residues 145-155, and 330 to be unfavorable. 3R44 was used as the template for this model and was found to be reliable by MetaMQAPII (Figure 3.2c) therefore indicating errors in the target structure are not due to the

template but possibly due to the low sequence identity/ amount of deviation within the template and target sequences since there was only a 23% identity found.

The A-domain of plp/feng synthetase module 9 generated was scored as reliable by MetaMQAPII with only 1 region around residue 198 within a loop to being identified as loop being unreliable (Figure 3.6i). This correlated with the results generated by ANOLEA (Appendix 1.12) however the results generated by Verify3D found two areas further within the structure to be of concern, around residues 247 and 316 (Appendix 1.12). However these two areas were different in each program. 1AMU was the template used for this structure and the only areas of this template identified, as a potential problematic area was not used in the modeling was within the N-terminal. The overall structure of 1AMU is reliable as shown by MetaMQAPII (Figure 3.2e).

The A-domain of module 10 of plp/ feng synthetase was mostly reliable with only two areas shown as unfavorable by MetaMQAPII at residues 37-38 and 158-159 (Figure 3.6j). These areas were shown in red on the structure the remainder of the structure is however deemed to be correct by MetaMQAPII and shown in green and blue on the structure. ANOLEA and Verify3D were also used to validate the structure (Appendix 1.12). ANOLEA identified areas around residues 20-25 and 130 while Verify3D identified an area around residue 300 as unfavorable. The template used for the construction of this model was 3R44 which was considered reliable by MetaMQAPII (Figure 3.2c).

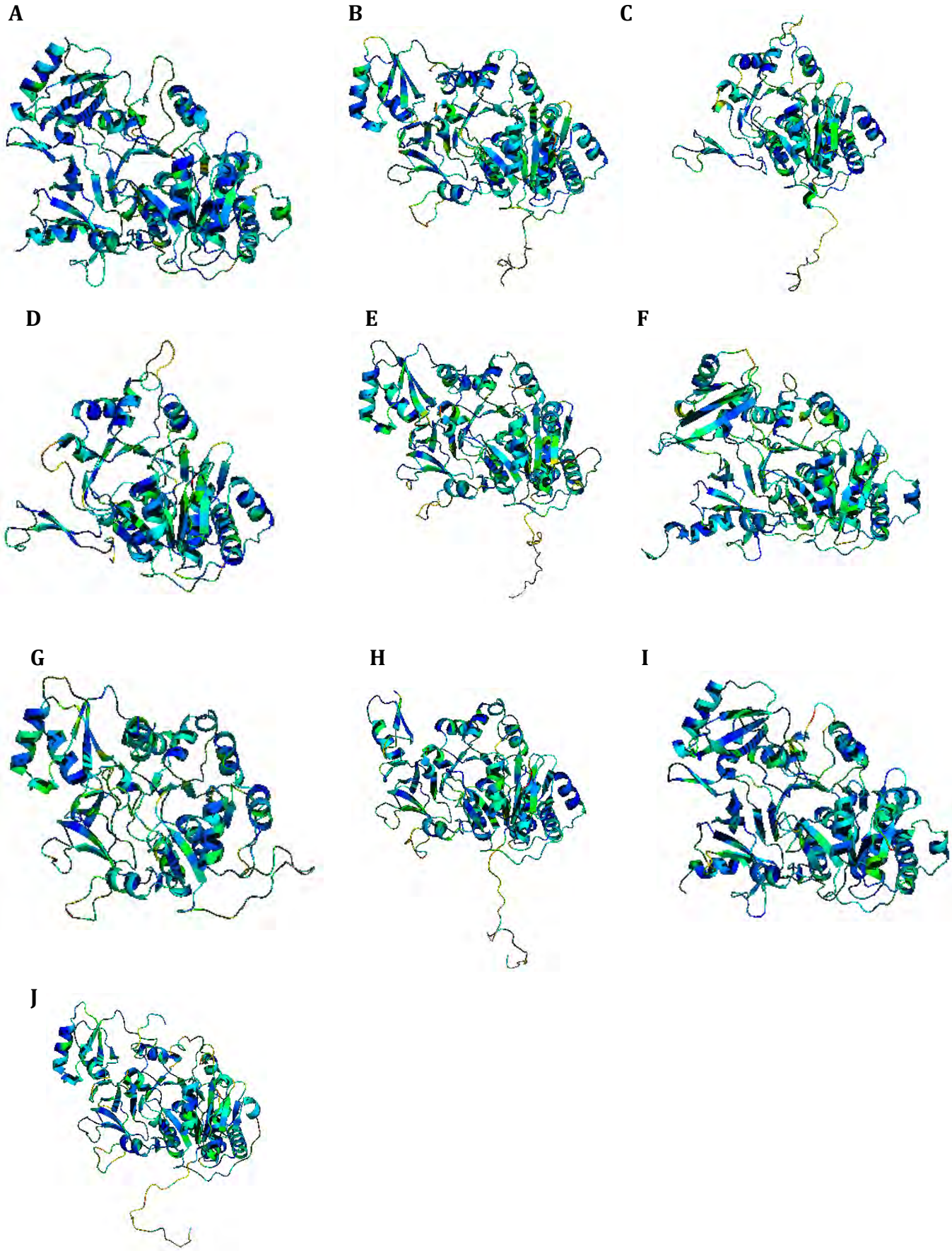


Figure 3.6: Plp/ feng synthetase A-domain module models colored according MetaMQAPII scores. Modules as follows; A) 2, B) 3, C) 5a, D) 5b, E) 6, F) 7a, G) 7b, H) 8, I) 9 and J) 10. Templates used for the model construction are shown in Table 3.2.

Myc synthetase subunit A module 1 A-domain (Figure 3.7a) had one region shown in red; residues 126-129. ANOLEA identified two regions of concern within the structure around residues 30 and 520 while Verify3D identified three areas around residues 20, 380 and 520 as unfavorable (Appendix 1.12). 3KXW was the template used to create this structure, which was found to be very reliable by MetaMQAPII (Figure 3.2b). Areas of concern within the modeled A-domain are therefore most likely attributed to the low sequence identity seen between the template and target of only 28%.

Myc synthetase subunit A module 2 A-domain (Figure 3.7b) only had one region of three residues, 36-38, that were less reliable and were contained within a loop region of the structure. ANOLEA and Verify3D were also used as a validation of the structure (Appendix 1.12). Verify3D did not identify any major concerns within the structure while ANOLEA identified 3 small areas around residues 200, 240 and 315. 2VSQ was the template this structure was modeled against which was mostly reliable but did have a few areas which MetaMQAPII identified to contain concerns (Figure 3.2g).

Myc synthetase subunit B module 1 A-domain structure (Figure 3.7c) had one region of 3 residues in length 40-42 that were shown by MetaMQAPII as not reliable. This region was contained within a loop. This was different to the areas identified to be of high energy and unfavorable by ANOLEA and Verify3D (Appendix 1.12) (Eisenberg et al., 1997; Melo et al., 1997). ANOLEA identified two small areas as high in energy around residues 20, and 510 while Verify3D identified areas around 1-15 and 370 as unfavorable. As a whole the structure seems reliable in comparison to the template with no major areas of concern. The template used for this structure was 2VSQ which was mostly reliable but did have a few areas which MetaMQAPII identified to contain concerns (Figure 3.2g).

Myc synthetase subunit B module 3 A-domain structure (Figure 3.7d) was identified by MetaMQAPII to be reliable/ correct with no major areas within the structure of concern. ANOLEA only identified one region around residue 250 and 1-30 by Verify3D to be unfavorable (Appendix 1.12). 2VSQ was the template this

structure was modeled against which was mostly reliable but did have a few areas which MetaMQAPII identified to contain concerns (Figure 3.2g).

Myc synthetase subunit C module 2 A-domain (Figure 3.7e) was shown to be correctly modeled and reliable by MetaMQAPII. ANOLEA and Verify3D identified unfavorable and areas of concern around residues; 240 and 290 respectively (Appendix 1.12). 2VSQ was the template this structure was modeled against which was mostly reliable but did have a few areas which MetaMQAPII identified to contain concerns (Figure 3.2g).

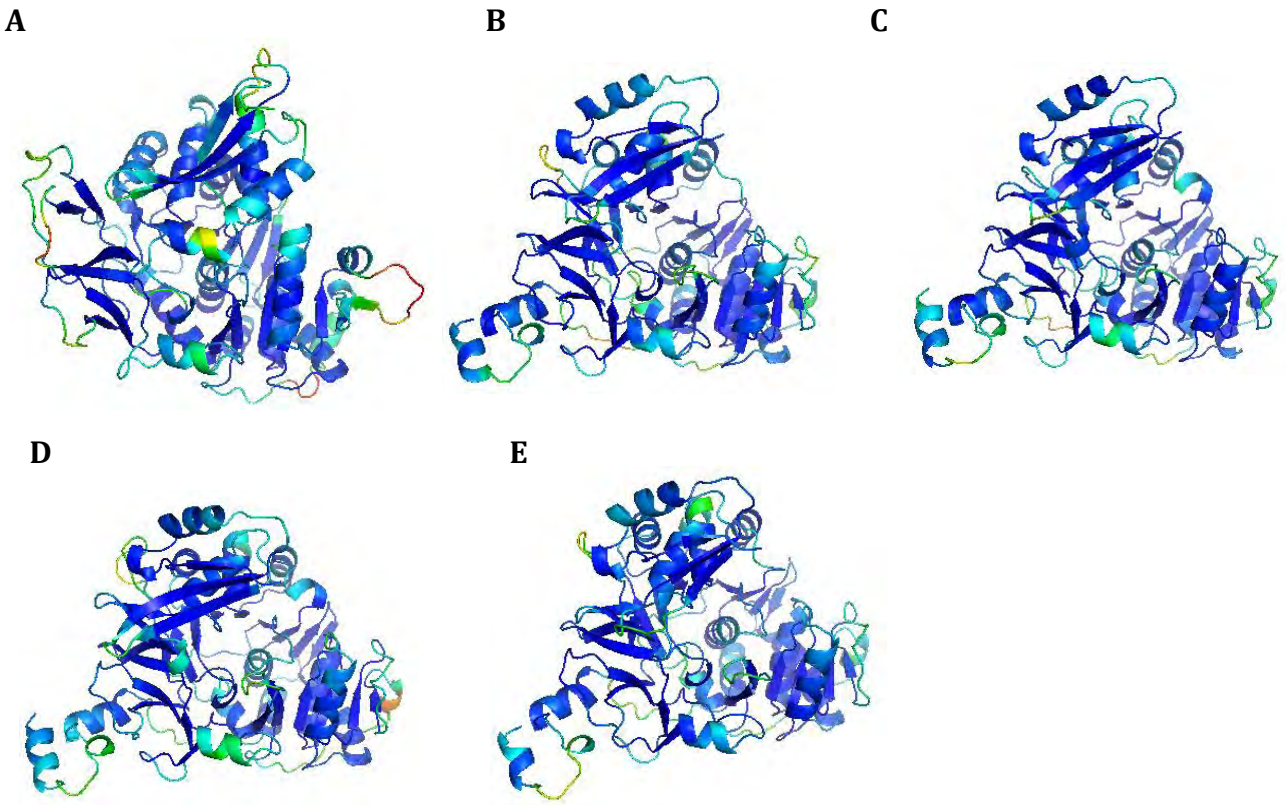


Figure 3.7: Myc synthetase A-domain module models colored according to MetaMQAPII scores. A) Myc synthetase subunit A module 1, B) Myc synthetase subunit A module 2, C) Myc synthetase subunit B module 1, D) Myc synthetase subunit B module 3, and E) Myc synthetase subunit C module 1.

3.4.2. A-domain Motif analysis:

The sequences of the A-domains which coded for the same amino acids were then compared in a sequence alignment and a motif search using MEME (Bailey *et al.*, 2009). Several motifs were found in each of the common A-domains showing the conservation between the A-domains, which specify the attachment

of the amino acids (Table 3.5-7). Glu occurs three times within the plp/ feng synthetase product. A MEME analysis was also done on these sequences to search for motifs within the domain (Table 3.8). Asn occurs three times within myc synthetase, once within each of the subunits. The sequences were run through MEME to search for motifs within the A-domains, which select for Asn (Table 3.9).

Table 3.5: Motifs identified in the A-domains of plp/ feng synthetase and myc synthetase modules by MEME.

Synthetase			Mycosubtilin								Plipstatin/ fengycin										
Subunit			A		B				C		1										
Module			1	2	1	2	3	4	1	2	1	2	3	4	5		6	7		8	9	10
Motif	E-value	Width													A	B		A	B			
1	1.4e-780	94		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓		✓	✓
2	5.1e-703	113	□	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	□	✓			✓	✓	✓
3	2.7e-309	41	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	□	✓	✓	✓	✓
4	6.5e-152	29		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	3.1e-119	113									✓				✓	□						
6	1e2-85	21	✓	✓	✓	✓	✓	✓	✓	✓	□	✓	✓	✓	□	□	✓	✓	✓	✓	✓	✓
7	1.7e-85	26		✓	✓	✓	✓	✓	✓	✓	□	✓	✓	✓	□		✓	✓	✓	□	✓	✓
8	5.6e-62	15		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓	✓	✓	✓
9	1.5e-38	21		✓	✓	✓	✓	✓	✓	✓	□	✓	✓	✓	□		✓	✓	✓	□	✓	✓
10	1.7e-26	15										✓	✓	✓			✓	✓	✓	□	✓	✓

Plp/feng synthetase and myc synthetase each contain a Pro. Plp/feng synthetase contains two Tyr's and myc synthetase contains one Tyr. The A-domains, which code for the Pro residues and those, which code for the Tyr residues were each compared and investigated in a motif search using MEME (Table 3.10).

Table 3.6: Myc synthetase A-domain motifs. Motifs shown in green are those, which are in the modules, which code for Asn alone. Motifs shown in blue are found in 7 of 8 of the A-domains.

Synthetase				Mycosubtilin							
Subunit				A		B				C	
Module				1	2	1	2	3	4	1	2
Motif	E-value	Width	Sites								
1	5.5e-282	97	7		✓	✓	✓	✓	✓	✓	✓
2	2.6e-245	106	7		✓	✓	✓	✓	✓	✓	✓
3	5.2e-141	159	3		✓		✓				✓
4	1.7e-69	41	7		✓	✓	✓	✓	✓	✓	✓
5	3.9e-36	59	4			✓		✓	✓	✓	
6	7.6e-16	21	7		✓	✓	✓	✓	✓	✓	✓
7	1.8e-8	21	7		✓	✓	✓	✓	✓		✓
8	8.6e-8	29	3		✓		✓				✓
9	2.0e-4	57	4			✓		✓	✓	✓	
10	6.3e-4	15	5	✓		✓		✓	✓	✓	

Table 3.7: Plp/ feng synthetase motifs identified in MEME search. Motifs shown in blue are those, which are in all the sequences/ A-domains of plp. Motifs shown in green are those only in the A-domains specifying Glu. Motifs shown in purple are those in all sequences except for those coding for Glu and red showing those not found in Glu with the exception of module 8.

Synthetase				Plipstatin/ fengycin											
Subunit				1											
Module				1	2	3	4	5		6	7		8	9	10
Motif	E-value	Width	Sites					A	B		A	B			
1	3.8e-344	95	9		✓	✓	✓			✓	✓	✓	✓	✓	✓
2	1.4e-302	104	10	?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	3.2e-207	80	8	✓		✓		✓	✓		✓	✓	✓	✓	
4	8.0e-75	38	8		✓	✓	✓			✓	✓	✓		✓	✓
5	1.5e-60	21	12	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	1.1e-49	20	12	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7	1.1e-38	41	3	✓				✓	✓						
8	1.0e-27	21	14	✓✓	?	✓	✓	✓✓	✓✓	✓	✓	✓	✓	✓	✓
9	2.1e-23	29	6		✓		✓			✓	✓	✓			✓
10	1.5e-18	78	4		✓		✓			✓					✓

Table 3.8: Motifs found in the A-domains of plp that code for Glu.

Synthetase				Plipstatin/ fengycin			
Subunit				1			
Module				1	5		8
Motif	E-value	Width	Sites		A	B	
1	4.6e-165	143	4	✓	✓	✓	✓
2	3.6e-64	63	4	✓	✓	✓	✓
3	1.0e-34	80	3	?	✓	✓	✓
4	2.0e-31	29	4	✓	✓	✓	✓
5	4.3e-13	20	4	✓	✓	✓	✓
6	6.5e-3	21	2	✓			✓
7	6.7e-3	8	4	✓	✓	✓	✓

Table 3.9: Motifs found in the A-domains of myc synthetase that code for Asn.

Synthetase				Mycosubtilin		
Subunit				A	B	C
Module				2	2	2
Motif	E-value	Width	Sites			
1	4.0e-187	224	3	✓	✓	✓
2	2.3e-64	112	3	✓	✓	✓
3	2.0e-23	52	3	✓	✓	✓
4	2.6e-17	35	3	✓	✓	✓
5	6.9e-4	14	3	✓	✓	✓
6	1.0e-3	17	3	✓	✓	✓
7	2.6e-2	11	3	✓	✓	✓

Table 3.10: Motifs found in the A-domains of plp and myc synthetase that code for Pro.

Synthetase				Mycosubtilin	Plipstatin/ fengycin	
Subunit				B	1	
Module				4	7	
Motif	E-value	Width	Sites		A	B
1	1.5e-98	159	3	✓	✓	✓
2	7.8e-90	187	2		✓	✓
3	8.1e-9	20	2		✓	✓
4	5.3e-4	80	2		✓	✓

Several motifs were found for each of the substrates specified (Tables 3.5-10). The individual motifs were not investigated in detail but the overall conservation was evident in the amount and motifs found as well as the lengths and repetition of the motifs within the individual A-domains as well as between the A-domains of the different modules which code attachment for the same substrate.

In certain cases where there was more than one A-domain identified by the SBSPKS (Chapter 2) it was found that the A-domain structures and sequences were not necessarily closer to that of the A-domain within the same module (i.e. in the case of plp/feng 5 a and b and 7 a and b) and often had a closer correlation to an A-domain within a different module. This was seen by superimposing the structures of the A-domains, which coded for the same amino acid substrate, Pro (Figure 3.8) it was seen that there was a high level of conservation between the structures of the three models. Plp/feng 7 a A-domain (shown in green) appears to be more structurally similar to that of the myc synthetase subunit B module 4 A-domain structure (shown in magenta) than the A-domain structure of plp/feng synthetase 7b (shown in cyan). There are, however, still notable differences seen in all three of the structures. The three A-domains which code for Tyr, plp/feng synthetase module 3 and 9 and myc synthetase subunit B module 1, were superimposed in Figure 3.9 and showed correlation between all three of the structures while still clearly showing the structures are not identical and have many areas overall where they are different. There are areas where plp/feng synthetase 3 (shown in green) and myc synthetase subunit B module 1 (shown in magenta) are more similar to each other as well as areas where plp/feng synthetase module 9 (shown in cyan) and myc synthetase subunit B module 1 are more similar in structure. Plp/feng synthetase modules 3 and 9 however seem more different to each other than to myc synthetase subunit B module 1 structurally. This could indicate that there is more conservation between NRPS than within them.

Figure 3.8: Plp/ feng synthetase module 7 (A and B) and myc synthetase subunit B module 4 all specifying for Pro are superimposed. Plp 7A shown in green, Plp 7B shown in cyan and Myc B module 4 shown in magenta.

Figure 3.9: Plp/Feng synthetase modules 3 and 9 and Myc synthetase subunit B module 1 all specifying for Tyr are superimposed. Plp 3 shown in green, Plp 9 shown in cyan and Myc B module 1 shown in magenta.

3.4.3. A-domain Substrate Pocket:

A multiple sequence alignment was carried out using the 3D information of the structures by PROMALS3D. This was performed between the sequences of three A-domains that code for Tyr (myc synthetase subunit B module 1, plp/feng synthetase module 3 and 9) as well as the sequence of known structure used by Stachelhaus *et al.,* (1999), 1AMU, containing the A-domain in complex with Phe (Figure 3.10). No A-domain specific for Phe binding were found within the NRPs of myc synthetase or plp/feng synthetase. The structure was therefore aligned with the A-domains of the NRPSs, which code attachment for Tyr as it is the closest structural amino acid to Phe and both amino acids are aromatic. The closest match to the template 1AMU, determined by PROMALS3D and shown by

the order in which the end alignment is outputted, was plp/feng synthetase module 9. The homology model of this A-domain was therefore superimposed with 1AMU in PyMOL. Using the multiple sequence alignment, the motifs identified by Stachelhaus *et al.* (1999) could be used to match up the interacting residues in the other sequences. The residues then predicted to interact within the binding pocket were identified and mapped into the models by using PyMOL (Figure 3.11). From this the interacting residues of plp/feng synthetase module 9 were found to be; 216D, 217G, 220T, 259I, 284T, 286A, 310E, 318V and 319A for plp/feng synthetase module 9.

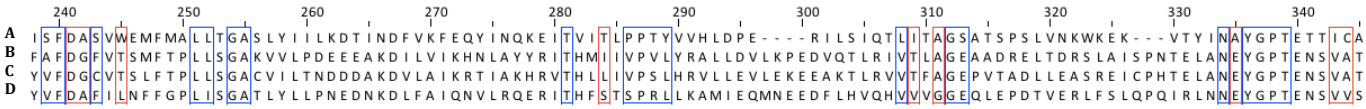


Figure 3.10: Multiple sequence alignment of the binding pockets within the A-domain of A) 1AMU, B) plp/feng module 9, C)plp/feng module 3 and D) myc synthetase subunit B module 1 carried out using PROMALS3D and visualized and colored using JalView. Motifs identified by Stachelhaus *et al.* (1999) are shown in blue boxes and interacting residues are shown in red boxes.

The motifs found between the 1AMU Phe A-domain and the A-domain modules form the test strain specific for Tyr were compared. Overall the motifs were quite conserved through all four module A-domains. They were however more conserved between the template (1AMU) and the plp/ feng synthetase module 3 and myc synthetase subunit B module 1 A-domains. Motifs 4, 5, 9 and 10 were conserved through all of the 4 sequences. Motif 6 was not conserved in the sequences.

The residues within the A-domains identified to be interacting with the incoming substrate were found to be more conserved between the template (1AMU) and the plp/ feng synthetase module A-domains especially between the plp/ feng synthetase module A-domains themselves. The residues appear to be more divergent for the myc synthetase in comparison to the template and plp/ feng synthetase residues identified. In cases where there is more conservation between the A-domains from the same NRPS this could indicate the substrate specificity is also dependent on the NRPS from which the A-domain originates. This is seen in the third, fifth, eighth and ninth interacting residue were the interacting residue is conserved between the plp/ feng synthetase A-domains

but differed in the template and the myc synthetase A-domain. There were also cases where the interacting residues differed in each of the NRPSs such as interacting residue 4. This could indicate that the module in which the A-domain occurs within the NRPS also affects the interaction with the A-domain and the substrate.

Residue Thr 220 in the plp/ feng synthetase module 9 A-domain projects away from the Phe substrate where the template (1AMU) Trp 239 aromatic rings project towards the Phe. This could be due to the hydroxyl group (OH) not present in the Phe but found within the Tyr substrate structure with which the plp/ feng synthetase module 9 interacts. Ile 259 of plp/ feng synthetase module 9 A-domain has a slightly longer side chain in comparison to the template counterpart, Thr 278, which also contains a hydroxyl group on its side chain, which the Ile 259 does not. This additional inward projection by the Ile 259 could show an additional interaction or stabilization with its specific substrate, Tyr, which near to that side chain would be the Tyr hydroxyl group.

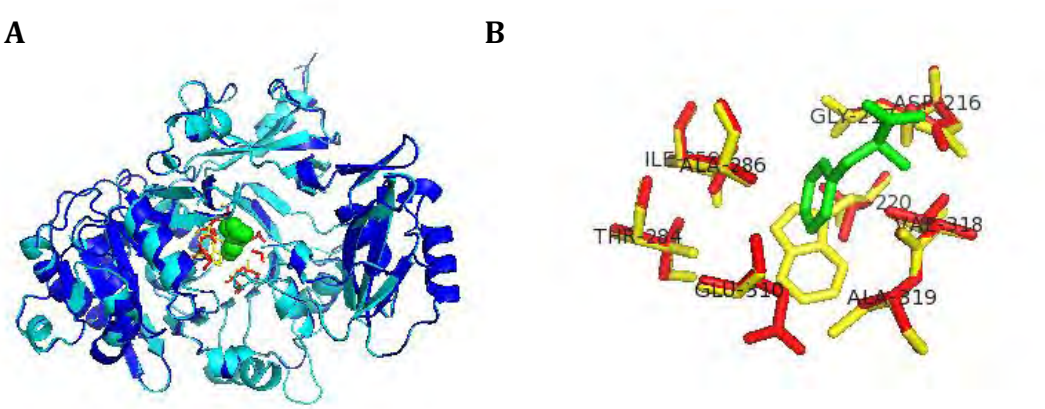


Figure 3.11: Plp/feng synthetase module 9 A-domain superimposed with 1AMU gramicidin synthetase in complex with Phe. 1AMU – blue, plp/feng 9 Adom – cyan, Phe – green, 1Amu interacting residues – yellow (identified by Stachelhaus *et al.* (1999)), plp/feng 9 A-domain interacting residues – red (identified with structural alignment with 1AMU). A) The full image of 1AMU in complex with Phe and plp/feng module 9 A-domain with interacting residues shown, B) Magnified view of the binding pocket and interacting residues.

This was repeated for the A-domains which code for an amino acid with different physiochemical properties to Phe, Glu; myc synthetase subunit B module 3, plp/ feng synthetase modules 5 (a and b) and 8 (Figure 3.12). The interacting residues within the binding pocket for myc synthetase subunit B module 3 were predicted and mapped to the structure (Figure 3.13). From this the interacting

residues of myc synthetase subunit B module 3 were found to be; 240D, 241A, 244Q, 283D 307L, 309G, 337V, 345V and 346D.

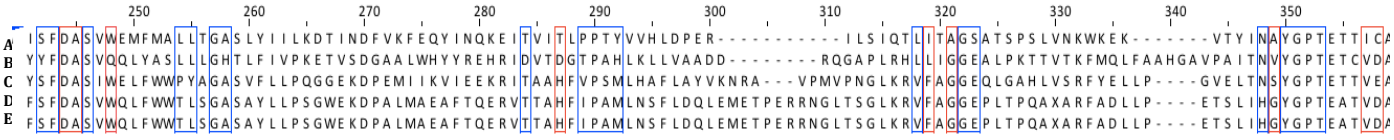


Figure 3.12: Structural alignment of the binding pockets within the A-domain of A) 1AMU, B) myc synthetase subunit B module 3, C) plp/feng module 8, D) plp/feng module 5a and E) plp/feng module 5b. Motifs identified by Stachelhaus *et al.* (1999) are shown in blue boxes and interacting residues are shown in red boxes.

The motifs found between the 1AMU Phe specific A-domain pocket and the A-domain modules form the test strain specific for Glu were compared. Overall the motifs were quite conserved though all five A-domain modules. They were however more conserved between the template (1AMU) and the plp/ feng synthetase A-domains. Motifs 2, 5 and 10 were conserved through all of the 5 sequences. Motifs 3 and 6 were not conserved in most of the sequences.

The residues within the A-domains identified to be interacting with the incoming substrate were found to be more conserved between the template (1AMU) and the plp/ feng synthetase module A-domains especially between the plp/ feng synthetase module A-domains themselves. The residues appear to be more divergent for the myc synthetase in comparison to the template and plp/ feng synthetase residues identified. In cases where the residues are conserved across all the A-domains, especially the first and second interacting residues, a relation between the specificity for the specific substrate can be seen. In certain cases there is more conservation between the A-domains from the same NRPS, which could indicate the substrate specificity is also dependent on the NRPS from which the A-domain originates. This is seen in the third interacting residue were a Glu is seen in all of the A-domains from plp/ feng synthetase and the template but differed in the myc synthetase A-domain. Interacting residues 4 and 5 show conservation between the plp/ feng synthetase residues and which is not seen in the myc synthetase or template interacting residues.

Residues 244 and 283 are smaller than the amino acid residues at the corresponding position in the alignment of 1AMU. The 1AMU corresponding residues, Trp 239 and Thr 278, are very different to the myc synthetase subunit B module 3 A-domain interacting residues at that point, Glu 244 and Asp 283, however there was no movement seen towards the substrate within the binding pocket seen by these residues. Ala 301 in 1AMU projects towards the Phe substrate, however, the myc synthetase subunit B module 3 A-domain corresponding residue at that point Gly 309 side chain projects away from the Phe substrate. This could be due to the additional branch found on the Glu substrate for which myc synthetase subunit B module 3 A-domain is specific.

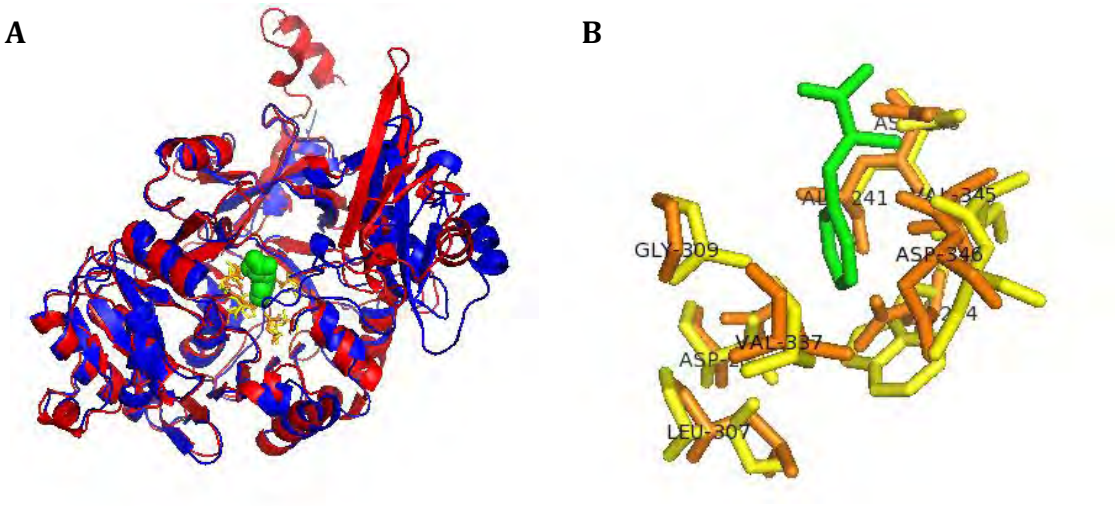


Figure 3.13: Myc synthetase subunit B module 3 A-domain module superimposed with 1AMU gramicidin synthetase in complex with Phe. 1AMU – blue, Myc subunit B module 3 A-domain – red, Phe – green, 1AMU interacting residues – yellow (identified by Stachelhaus *et al.* (1999)), Myc subunit B module 3 A-domain interacting residues – orange (identified with structural alignment with 1AMU). A) The full image of 1AMU in complex with Phe and Myc subunit B module 3 A-domain with interacting residues shown, B) Magnified view of the binding pocket and interacting residues.

3.5. Conclusions:

Full Modules:

Due to the large nature of these enzyme complexes (NRPSs) models of the full modules could only be built for plp/ feng synthetase module 6 and 9. Identities of the templates used to model the structures were low but were above 30% in each case and since structure is more conserved than sequence due to function

this was deemed as a sufficient amount for the purposes of generating a preliminary models. Structures were validated using MetaMQAPII. A few areas were found to be of concern within the structures, however, overall the structures were found to be reliable. The areas of concern were mainly within loop and terminal regions. These areas could be improved through loop refinement. The model constructed for plp/feng synthetase module 9 was found to be more reliable overall than that of plp/feng synthetase module 6. Two templates were used in modeling plp/feng synthetase module 6. The template used for the modeling of the first half of the 6th module of plp/feng synthetase was found by MetaMQAPII to contain a few areas for concern, which could therefore have influenced the generated structure. The same template was used for the modeling of the 9th module of plp/feng synthetase and the structure generated, when validated by MetaMQAPII was found to be more reliable/ correct than that of the template structure when also validated using MetaMQAPII.

TE-domains:

The TE-domains of plp/feng synthetase and myc synthetase were generated using single templates with very low percentage identities (16% and 18% respectively). The template structures were of high quality and did not show many areas of concern when validated using MetaMQAPII. The TE-domains modeled did show a few areas that were of concern and these areas were mainly located within loop regions of the structure which could be further improved through loop refinement. The plp/feng synthetase and myc synthetase TE-domains constructed were superimposed which showed a level of conservation between the two structures however the structures seemed to be more different from each other than similar. This indicates there is a level of conservation for functional purposes while still maintaining a high level of divergence and specific to the NRP for which the regulate release for.

A-domain structures:

A-domain structures were constructed for the myc synthetase and plp/feng synthetase modules, with the exception of plp/feng synthetase modules 1 and 4

and myc synthetase subunit B module 2 and 4 and subunit C module 1. Of the 14 structures modeled only 4 templates were used showing the amount of conservation between all the A-domains. Models were validated using Verify3D, ANOLEA and MetaMQAPII. Certain areas were identified within the structure to be of concern. The three validation methods however did not usually correlate in the areas in which they identified as incorrect/ high-energy areas. Overall the models generated were of a good quality but could use further refinement. The templates used in the modeling of the structures were also validated by MetaMQAPII and did not show any major areas of concern.

A-domain Motif analysis:

The sequences of the A-domains which coded for the same amino acids were compared in a sequence alignment and a motif search using MEME. Several motifs were found in each of the common A-domains showing the conservation between the A-domains, Several motifs were found for Tyr, Pro, Glu and Asn. The individual motifs were not investigated in detail but the overall conservation was evident in the amount and motifs found as well as the lengths and repetition of the motifs within the individual A-domains as well as between the A-domains of the different modules which code attachment for the same substrate.

The A-domain structures and sequences were not necessarily closer to that of the A-domain within the same module (i.e. in the case of plp/feng 5 a and b and 7 a and b) and often had a closer correlation to an A-domain within a different module. There was a some correlation between A-domains within the same NRPS but there seemed to be more between A-domains of different NRPSs. This could indicate that there is more conservation between NRPS than within them.

A-domain Substrate Pocket:

Overall the motifs were quite conserved through all module A-domains investigated for Tyr-Phe and Glu-Phe. There were, however, notable differences between the motifs in A-domain binding pocket between the different NRPSs specific to the same substrate as well as to the different modules within the same NRPSs. Certain motifs were conserved across all the modules of the different

NRPSs while other motifs were different in each module even within the same NRPS.

Interacting residues are more conserved between the template (1AMU) and the plp/ feng synthetase module A-domains especially between the plp/ feng synthetase module A-domains themselves. The residues appear to be more divergent for the myc synthetase in comparison to the template and plp/ feng synthetase residues identified. In cases where there is more conservation between the A-domains from the same NRPS that could indicate that substrate specificity is also dependent on the NRPS from which the A-domain originates. The module in which the A-domain occurs within the NRPS was also found to affects the interaction with the A-domain and the substrate by different interacting residues being involved in different modules of the same NRPS.

In cases where the residues are conserved across all the A-domains, especially the first and second interacting residues, a relation between the specificity for the specific substrate can be seen. In cases where there is more conservation between the A-domains from the same NRPS substrate specificity could be influenced by the NRPS from which the A-domain originates.

Specific residues were seen that either project away from or towards the substrate within the binding pocket to a greater degree or in an opposite direction to that seen in the template structure. These changes were attributed to the differing substrate, which would usually occur within the binding pocket of the A-domain being investigated (i.e. Tyr or Glu not Phe) such as the additional hydroxyl group seen in Tyr, which is not present in Phe. Ile 259 shows more of a projection towards the substrate than seen in the template towards Phe and could indicate an additional interaction or stabilization with its specific substrate, Tyr near to the hydroxyl group. Gly 309 side chain projects away from the Phe substrate which could be due to the additional branch on the Glu substrate for which myc synthetase subunit B module 3 A-domain is specific.

All the interacting residues identified are in a similar location to those identified by Stachelbaues *et al.,* (1999). The residues predicted to be interacting in both cases appear to be projecting in towards the Phe within the center of the binding pocket. The two A-domains modeled however do not code for Phe and therefore an exact interaction with the substrate and the residues cannot be seen. This could be done in future work using a docking study and the substrate the A-domain specifies.

Since this was a pilot study with limited time constraints only one model was generated by Modeller using the MPI Toolkit in each case discussed. . Although HHpred uses Modeller, it is not clear what calculations were done while building the models. Ideally this should be done in script base by using stand alone Modeller program and at least 50 to 100 models should be generated to determine the optimal model. The models generated were found to be of a sufficiently high and acceptable quality and were therefore found to be sufficient for this study. This could, however, be continued in another study where more models would be generated to select the most optimal structure and wherein loop refinement could be carried out.

Future work would involve the refinement of these models through processes such as loop refinement. The binding pockets of the NRPS A-domains could be investigated within a docking study with the substrate they are predicted to activate. This would also involve the investigation and comparison of all the different substrates for which the different NRPS module A-domains are specific. A more detailed investigation could be carried out in future work on the A-domain binding pocket cavity. This would give insight and clarity into the exact binding residues within the binding pocket specific to the amino acid being activated by the specific A-domain of the NRPS.

FOURTH CHAPTER: CONCLUSIONS

A newly sequenced *Bacillus* strain, found to have increased phytopathogenic activity, was investigated to gain insights to the possible reasoning behind this activity. This strain was found to contain modules for five NRPSs; myc synthetase, plp/ feng synthetase, srf synthetase, bcb synthetase and bct synthetase. The full modules known to be required in the production of srf and bct NRPs were not found in this strain and only some of the known modules were detected. This could be due to contig construction and the variability of the regions surrounding the NRPSs within the genome. The first amino acid, Dhb, of bcb was not found in this or neighboring strains. Myc synthetase and plp/feng synthetase were found to be complete and comparable to literature.

The NRPSs and the amino acids that they code attachment for in the NRP product of the test strain were compared to neighboring strains. Neighboring strains were identified in a phylogenetic study looking at other *Bacillus* strains. Neighboring strains to the test strain, *B. atrophaeus* UCMB 5137 (63Z), were found to be *B. atrophaeus*, *B. subtilis* and *B. amyloliquifaciens* strains. Topologies of the trees were found to be similar and the hypothesis that the genes were transferred horizontally was therefore rejected. The branch lengths where however found to be significantly different which led to the hypothesis that different NRPS genes are under different adaptive evolutionary pressures.

The neighboring strains used for comparison were; *B. atrophaeus* 1942, *B. subtilis* BS5N, *B. subtilis subsp. subtilis* 168 and *B. amyloliquifaciens* FZB42. When comparing the individual NRPS genes to those in the neighboring strains all were found to follow the same pattern of evolution where the test strain was closest related to the *B. atrophaeus* 1942 strain followed by the *B. subtilis* strains and

finally the *B. amyloliquifaciens* strain. The *B. subtilis* strains were not found to contain the myc synthetase modules. Only one change in the NRP amino acids in the myc produced by *B. amyloliquifaciens* FZB42 strain where the C subunit was found to code for Serine and Threonine whereas the database, test strain and *B. atrophaeus* 1942 strain myc subunit C were found to code for Ser and Asn. This strain also differed in its B subunit of Tyr(D)-Asn-Pro-Glu whereas the other strains and database showed a pattern of Tyr(D)-Asn-Glu-Pro. The *B. atrophaeus* and *B amyloliquifaciens* strains were also found to differ slightly from the *B. subtilis* strain and the database record in the case of feng. These were only slight changes in that certain of the amino acids were found to be in the L-conformation wherein the other strains were in the D-conformation.

The terminal regions between modules as well as the linker regions between domains of the NRPSs found were investigated for conservation. The conservation was investigated within the test strain overall as well as in comparison to that of the neighboring strains. All of the linker and terminal regions were found to exhibit a large amount of conservation overall. This level of conservation is possibly indicative of its importance in the correct assembly of the NRP products in correlation with previous studies done where it was found that the linker regions when interrupted lead to errors in the production of the NRPs (Chapter 1). The terminal and linker regions were found to be conserved within the strain as well as in comparison to neighboring strains.

TE-domains are similar and have some overlapping areas. There are however more regions which are different between the structures than regions that are the same. This can therefore show that the TE-domains between the different synthetases are different and perhaps specific for the NRP they facilitate release for.

Homology models were builtd to gain insight in to the structure of the NRPSs of the test stain. A-domains, which code for the same amino acid in myc synthetase and plp/ feng synthetase were compared and it was therefore evident that there is a high amount of conservation between these domains while still maintaining

unique distinct areas within the structure in each case. Since there were no A-domains found to code attachment for Phe within plp/feng synthetase or myc synthetase these predicted residues were for A-domains coding attachment for Tyr and Glu respectively. This could therefore affect the interaction seen between these residues and the Phe used in comparison to the substrate for which it is specific. Motifs were quite conserved through all module A-domains investigated for Tyr-Phe and Glu-Phe. Certain motifs were conserved across all the modules of the different NRPSs while other motifs were different in each module even within the same NRPS. Interacting residues are more conserved between the template (1AMU) and the plp/ feng synthetase module A-domains especially between the plp/ feng synthetase module A-domains themselves. The residues are more divergent for the myc synthetase in comparison to the template and plp/ feng synthetase residues identified. Substrate specificity may also be influenced by the NRPS from which the A-domain originates. The module in which the A-domain occurs within the NRPS also affects the interaction with the A-domain and the substrate by different interacting residues being involved in different modules of the same NRPS.

Specific residues were seen that either project away from or towards the substrate within the binding pocket to a greater degree or in an opposite direction to that seen in the template structure. These changes were attributed to the differing substrate, which would usually occur within the binding pocket of the A-domain being investigated (i.e. Tyr or Glu not Phe) such as the additional hydroxyl group seen in Tyr, which is not present in Phe. Ile 259 shows more of a projection towards the substrate than seen in the template towards Phe and could indicate an additional interaction or stabilization with its specific substrate, Tyr near to the hydroxyl group. Gly 309 side chain projects away from the Phe substrate which could be due to the additional branch on the Glu substrate for which myc synthetase subunit B module 3 A-domain is specific.

Future work on the structural elements would involve the refinement of these models by loop refinement and the generation of multiple models from which the most optimal predicted structure could be selected. The binding pockets of

the NRPS A-domains could be investigated within a docking study with the substrate they are predicted to activate to give insight and clarity into the exact binding residues within the binding pocket specific to the amino acid being activated by the specific A-domain of the NRPS.

Since the completion of this work further refinement has been done to the sequence of the test strain. Further investigation of this corrected assembly could possibly give more insight into the specialized phytopathogenicity of the strain. Joining and contig construction has since been reviewed and refined by O. Reva and revealed some duplications were attributed to areas of contig duplications. The final sequence now contains 24 contigs of the total length of 4,075,528 bp, which is closer to the expected sequence length. This new information and refined sequence could be used in future work on the genome of *B. atrophaeus* UCMB 5137 (63Z) in the investigation of its extraordinary phytopathogenic ability. These changes in the sequence due to the duplication removal will possibly affect the linker region coordinates and could be further verified.

REFERENCES:

Anand, S., Prasad, M. V. R., Yadav, G., Kumar, N., Shehara, J., Ansari, M. Z., & Mohanty, D. (2010). SBSPKS: structure based sequence analysis of polyketide synthases . Nucleic Acids Research , 38 (suppl 2), W487–W496.

Aziz, R., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R., Formsma, K., *et al*. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. Biomed Central Genomics, 9(1), 75.

Bachmann, B. O., & Ravel, J. (2009). Chapter 8 Methods for In Silico Prediction of Microbial Polyketide and Nonribosomal Peptide Biosynthetic Pathways from DNA Sequence Data. In D. A. H. B. T.-M. in Enzymology (Ed.), Complex Enzymes in Microbial Natural Product Biosynthesis, Part A: Overview Articles and Peptides (Vol. Volume 458, pp. 181–217). Academic Press.

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings International Conference on Intelligent Systems for Molecular Biology ISMB International Conference on Intelligent Systems for Molecular Biology, 2(6), 28–36.

Bailey, T. L., & Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics, 14 (1), 48–54.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., *et al*. (2009). MEME Suite: tools for motif discovery and searching. Nucleic Acids Research , 37 (suppl 2), W202–W208.

Barbe, V., Cruveiller, S., Kunst, F., Lenoble, P., Meurice, G., Sekowska, A., Vallenet, D., *et al*. (2009). From a consortium sequence to a unified sequence: the *Bacillus* subtilis 168 reference genome a decade later. Microbiology, 155 (6), 1758–1775.

Bishop, A. O. T., De Beer, T. A. P., & Joubert, F. (2008). Protein homology modelling and its use in South Africa. Academy of Science of South Africa.

Bowie, J. U., Lüthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253(5016), 164–170.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution , 17 (4), 540–552.

Challis, Gregory L, & Naismith, J. H. (2004). Structural aspects of non-ribosomal peptide biosynthesis. Current Opinion in Structural Biology, 14(6), 748–756.

Chen, X. H., Koumoutsi, A., Scholz, R., Eisenreich, A., Schneider, K., Heinemeyer, I., Morgenstern, B., *et al*. (2007). Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus* amyloliquefa- ciens FZB42. Nature Biotechnology, 25, 1007–1014.

Christiansen, G., Philmus, B., Hemscheidt, T., & Kurmayer, R. (2011). Genetic Variation of Adenylation Domains of the Anabaenopeptin Synthesis Operon and Evolution of Substrate Promiscuity . Journal of Bacteriology , 193 (15), 3822–3831.

Deleu, M., Paquot, M., & Nylander, T. (2005). Fengycin interaction with lipid monolayers at the air–aqueous interface—implications for the effect of fengycin on biological membranes. Journal of Colloid and Interface Science, 283(2), 358–365.

di Luccio, E. & Koehl, P. (2011). A Quality Metric for Homology Modelling: The H-Factor. Bioinformatics, 12, 48.

Duitman, E. H., Hamoen, L. W., Rembold, M., Venema, G., Seitz, H., Saenger, W., Bernhard, F., *et al*. (1999). The mycosubtilin synthetase of *Bacillus* subtilis ATCC6633: A multifunctional hybrid between a peptide synthetase, an amino transferase, and a fatty acid synthase. Proceedings of the National Academy of Sciences , 96 (23), 13294–13299.

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

Eisenberg, D., Lüthy, R., & Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. Methods in enzymology, 277, 396–404.

Finking, R., & Marahiel, M. (2004). Biosynthesis of nonribosomal peptides 1. Annual Review of Microbiology, (58), 453–488.

Fritze, D., & Pukall, R. diger. (2001). Reclassification of bioindicator strains *Bacillus* subtilis DSM 675 and *Bacillus* subtilis DSM 2277 as *Bacillus* atrophaeus. International Journal of Systematic and Evolutionary Microbiology, 51, 35–37.

Fuma, S., Fujishima, Y., Corbell, N., D'Souzal, C., Nakano, M. M., Zuberl, P., & Yamane, K. (1993). Nucleotide sequence of 5' portion of srfA that contains the region required for competence establishment in *Bacillus* subtilus. Nucleic Acids Research, 21(1), 93–97.

Gokhale, R. S., Tsuji, S. Y., Cane, D. E., & Khosla, C. (1999). Dissecting and Exploiting Intermodular Communication in Polyketide Synthases. Science, 284(5413), 482–485.

Grunewald, J., & Marahiel, M., (2006). Chemoenzymatic and Template-Directed Synthesis of Bioactive Macrocyclic Peptides. Microbiology and Molecular Biology Reviews 70, 121–146.

Hahn, M., & Stachelhaus, T. (2004). Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains. Proceedings of the National Academy of Sciences 101(44), 15585–15590.

Harrison, C., & Langdale, J. (2006). A step-by-step guide to phylogeny reconstruction. Plant Journal, (45), 561–572.

Heathcote, M. L., Staunton, J., & Leadlay, P. F. (2001). Role of type II thioesterases: evidence for removal of short acyl chains produced by aberrant decarboxylation of chain extender units. Journal of Chemical Biology 8, 207–220.

Hillson, N., & Walsh, C. (2003). Dimeric structure of the six-domain VibF subunit of vibriobactin synthetase: mutant domain activity regain and ultracentrifugation studies. Biochemistry, (42), 766–775.

Hoffmann, D., Hevel, J. M., Moore, R. E., & Moore, B. S. (2003). Sequence analysis and biochemical characterization of the nostopeptolide A biosynthetic gene cluster from Nostoc sp. GSV224. Gene, 311(0), 171–180.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., *et al*. (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Research, 40 (D1), D306–D312.

Jenke-Kodama, H., & Dittmann, E. (2009). Bioinformatic perspectives on NRPS/PKS megasynthases: Advances and challenges. Natural Product Reports, 26(7), 874–883.

Kinsinger, R., Shirk, M., & Fall, R. (2003). Rapid surface motility in *Bacillus* subtilis is dependent on extracellular surfactin and potassium ion. Journal of Bacteriology (185), 5627–5631.

Kleywegt G. J. and Jones T. A., (1997). "Model Building and Refinement Practice". Methods in Enzymology 277, 208-230.

Koglin, A., Löhr, F., Bernhard, F., Rogov, V. R., Frueh, D. P., Strieter, E. R., Mofid, M. R., (2008). Structural basis for the selectivity of the external thioesterase of the surfactin-synthetase. Nature, 454, 907–911.

Konz D, Klens A, Schörgendorfer K, M. M. (2005). The bacitracin biosynthesis operon of *Bacillus* licheniformis ATCC 10716: molecular characterization of three multi-modular peptide synthetases. BioTechniques.

Koumoutsi, A., Chen, X. H., Henne, A., Liesegang, H., Hitzeroth, G., Franke, P., Vater, J., (2004). Structural and Functional Characterization of Gene Clusters Directing Nonribosomal Synthesis of Bioactive Cyclic Lipopeptides in *Bacillus* amyloliquefaciens Strain FZB42. Journal of Bacteriology, 186(4), 1084–1096.

Lai, J. R., Fischbach, M. A., Liu, D. R., & Walsh, C. T. (2006). A protein interaction surface in nonribosomal peptide synthesis mapped by combinatorial mutagenesis and selection. Proceedings of the National Academy of Sciences 103(14), 5314–5319.

Leclère, V., Béchet, M., Adam, A., Guez, J.-S., Wathelet, B., Ongena, M., Thonart, P., (2005). Mycosubtilin Overproduction by *Bacillus* subtilis BBG100 Enhances the Organism's Antagonistic and Biocontrol Activities . Applied and Environmental Microbiology , 71 (8), 4577–4584.

Liithy, R., Bowie, J., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature. 356:83-85.

Lin, T.-P., Chen, C.-L., Chang, L.-K., Tschen, J. S.-M., & Liu, S.-T. (1999). Functional and Transcriptional Analyses of a Fengycin Synthetase Gene, fenC, from *Bacillus* subtilis . Journal of Bacteriology , 181 (16), 5060–5067.

Maget-Dana, R., Thimon, L., Peypoux, F., & Ptak, M. (1992). Surfactin/iturin A interactions may explain the synergistic effect of surfactin on the biological properties of iturin A. Biochimie, 74(12), 1047–1051.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., Gonzales, N. R., (2013). CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Research, *41* (D1), D348–D352.

May, J. J., Wendrich, T. M., & Marahiel, M. A. (2001). The dhb Operon of Bacillus subtilisEncodes the Biosynthetic Template for the Catecholic Siderophore 2,3-Dihydroxybenzoate-Glycine-Threonine Trimeric Ester Bacillibactin. Journal of Biological Chemistry , 276 (10), 7209–7217.

Meier, J. L., & Burkart, M. D. (2009). The chemical biology of modular biosynthetic enzymes. Chemical Society Reviews 38, 2012–2045.

Melo, F., & Feytmans, E. (1998). Assessing protein structures with a non-local atomic interaction energy. Journal of Molecular Biology, 277(5), 1141–1152.

Melo, F., Devos, D., Depiereux, E., & Feytmans, E. (1997). ANOLEA: a www server to assess protein structures. Proceedings/International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology, 5, 187–190.

Menkhaus, M., Ullrich, C., Kluge, B., Vater, J., Vollenbroich, D., & Kamp, R. M. (1993). Structural and functional organization of the surfactin synthetase multienzyme system. . Journal of Biological Chemistry , 268 (11), 7678–7684.

Minowa, Y., Araki, M., & Kanehisa, M. (2007). Comprehensive Analysis of Distinctive Polyketide and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes. Journal of Molecular Biology, 368, 1500–1517.

Moszer, I. (1998). The complete genome of *Bacillus* subtilis: from sequence annotation to data management and analysis. Federation of European Biological Societies (FEBS) Letters, (430), 28–36.

Nagórska, K., Bikowski, M., & Obuchowski, M. (2007). Multicellular behaviour and production of a wide variety of toxic substances support usage of *Bacillus* subtilis as a powerful biocontrol agent. The Journal of the Polish Biochemical Society and of the Committee of Biochemistry and Biophysics Polish Academy of Sciences, 54(3), 495–508.

Nguyen, K. T., Ritz, D., Gu, J.-Q., Alexander, D., Chu, M., Miao, V., Brian, P., (2006). Combinatorial biosynthesis of novel antibiotics related to daptomycin . Proceedings of the National Academy of Sciences, 103 (46), 17462–17467.

Nielsen, T. H., & Sorensen, J. (2003). Production of cyclic lipopeptides by Pseudomonas fluorescens strains in bulk soil and in the sugar beet rhizo- sphere. Applied and Environmental Microbiology 69, 861–868.

O'Connor, S.., Walsh, C. T., & Liu, F. (2003). Biosynthesis of Epothilone Intermediates with Alternate Starter Units: Engineering Polyketide–Nonribosomal Interfaces, Angewandte Chemie International Edition 42(33), 3917–3921.

Ongena, M., & Jacques, P. (2008a). *Bacillus* lipopeptides: versatile weapons for plant disease biocontrol. Trends in Microbiology, 16(3), 115–25.

Ongena, M., Jourdan, E., Adam, A., Paquot, M., Brans, A., Joris, B., Arpigny, J.-L., *et al*. (2007). Surfactin and fengycin lipopeptides of *Bacillus* subtilis as elicitors of induced systemic resistance in plants. Environmental Microbiology, 9(4), 1084–1090.

Pawlowski, M., Gajda, M., Matlak, R., & Bujnicki, J. (2008). MetaMQAP: A meta-server for the quality assessment of protein models. Biomed Central Bioinformatics, 9(1), 403.

Pei, J., & Grishin, N. (2007). PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics. 23(7):802-8.

Peypoux, F., Bonmatin, J. M., & Wallach, J. (1999). Recent trends in the biochemistry of surfactin. Applied Microbiology and Biotechnology 51:, 553–563.

Phae, C. G., Shoda, M., & Kubota, H. (1990). Suppressive effect of *Bacillus* subtilis and it's products on phytopathogenic microorganisms. Journal of Fermentation and Bioengineering, 69(1), 1–7.

Quadri, L. E. N. (2000). Assembly of aryl-capped siderophores by modular peptide synthetases and polyketide synthases. Molecular Microbiology, 37(1), 1–12.

Ratledge, C., & Dover, L. G. (2000). IRON METABOLISM IN PATHOGENIC BACTERIA. Annual Review of Microbiology, 54(1), 881–941.

Romero, D., de Vicente, A., Rakotoaly, R. H., Dufour, S. E., Veening, J.-W., Arrebola, E., Cazorla, F. M., *et al*. (2007). The Iturin and Fengycin Families of Lipopeptides Are Key Factors in Antagonism of *Bacillus* subtilis Toward Podosphaera fusca. Molecular Plant-Microbe Interactions, 20(4), 430–440.

Rowland, B. M., Grossman, T. H., Osburne, M. S., & Taber, H. W. (1996). Sequence and genetic organization of a *Bacillus* subtilis operon encoding 2,3-dihydroxybenzoate biosynthetic enzymes. Gene, 178(1–2), 119–123.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., & Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics, 16(10), 944–945.

Šali, A., Potterton, L., Yuan, F., van Vlijmen, H., & Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. Proteins: Structure, Function, and Bioinformatics, 23(3), 318–326.

Samel, S. a, Schoenafinger, G., Knappe, T. a, Marahiel, M. a, & Essen, L.-O. (2007). Structural and functional insights into a peptide bond-forming bidomain from a nonribosomal peptide synthetase. Structure, 15(7), 781–92.

Samel, S. A., Wagner, B., Marahiel, M. A., & Essen, L.-O. (2006). The Thioesterase Domain of the Fengycin Biosynthesis Cluster: A Structural Base for the Macrocyclization of a Non-ribosomal Lipopeptide. Journal of Molecular Biology, 359(4), 876–889.

Schwarzer, Dirk, Mootz, H. D., Linne, U., & Marahiel, M. a. (2002). Regeneration of misprimed nonribosomal peptide synthetases by type II thioesterases. Proceedings of the National Academy of Sciences of the United States of America, 99(22), 14083–8.

Sieber, S., & Marahiel, M. (2003). Learning from nature's drug factories: nonribosomal synthesis of macrocyclic peptides. J Bacteriol, (185), 7036–7043.

Siezen, R. J., & Khayatt, B. I. (2008). Natural products genomics. Microbial Biotechnology, 4, 275–82.

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research , 33 (suppl 2), W244–W248.

Souto, G., Correa, O., Montecchia, M., Kerber, N., Pucheu, N., Bachur, M., & Garcia, A. (2004). Genetic and functional characterization of a *Bacillus* sp. strain excreting surfactin and antifungal metabolites partially identified as iturin-like compounds. Journal of Applied Microbiology, (97), 1247–1256.

Stachelhaus, T., Mootz, H. D., & Marahiel, M. A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Journal of Chemical Biology., 6, 493–505.

Stachelhaus, T., Schneider, A., & Marahiel, M. A. (1995). Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. Science, 269(5220), 69–72.

Starcevic, A., Zucko, A. S. J., Simunkovic, J., Long, P. F., Cullum, J., & Hranueli, D. (2008). ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Research,, 36(21), 6882–6892.

Stein, T. (2005). *Bacillus* subtilis antibiotics: structures, syntheses and specific functions. Molecular Microbiology, 56(4), 845–857.

Stevens, B. W., Lilien, R. H., Georgiev, I., Donald, B. R., & Anderson, A. C. (2006). Redesigning the PheA domain of gramicidin synthetase leads to a new understanding of the enzyme's mechanism and selectivity. Biochemistry, 45(51), 15495–15504.

Stothard P (2000) The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102-1104.

Strieker, M., Tanović, A., & Marahiel, M. A. (2010). Nonribosomal peptide synthetases: structures and dynamics. Current Opinion in Structural Biology, 20(2), 234–240.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods . Molecular Biology and Evolution. 28 (10), 2731–2739.

Tanovic, A., Samel, S. a, Essen, L.-O., & Marahiel, M. A. (2008). Crystal structure of the termination module of a nonribosomal peptide synthetase. Science 321(5889), 659–63.

The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.

Toure, Y., Ongena, M., Jacques, P., Guiro, A., & Thonart, P. (2004). Role of lipopeptides produced by *Bacillus* subtilis GA1 in the reduction of grey mould

disease caused by Botrytis cinerea on apple. Journal of Applied Microbiology, (96), 1151–1160.

Tsuji, S. Y., Wu, N., & Khosla, C. (2001). Intermodular Communication in Polyketide Synthases. Biochemistry, 40(8), 2317–2325.

Vallari, D. S., Jackowski, S., & Rock, C. O. (1987). Regulation of pantothenate kinase by coenzyme A and its thioesters. The Journal of Biological Chemistry, 262(6), 2468–2471.

Walsh, C. (2008). The chemical versatility of natural-product assembly lines. Accounts of Chemical Research (41), 4–10.

Weber G, Schorgendorfer K, Schneiderscherzer E, L. E. (1994). The peptide synthetase catalyzing cyclosporine production in Tolypocladium niveum is encoded by a giant 45.8-kilobase open reading frame. Current Genetics (26), 120–125.

Weissman, K.J., & Leadlay, P. F. (2005). Combinatorial biosynthesis of reduced polyketides. Nature Reviews Microbiology, 3, 925–936.

Wipat, A., & Harwood, C. R. (1999). The *Bacillus* subtilis genome sequence: the molecular blueprint of a soil bacterium. FEMS Microbiology Ecology 28, 1–9.

Xiang, Z. (2006). Advances in Homology Protein Structure Modelling. Current Protein & Peptide Science, 7(3), 217-27.

Xiong, J. (2006). Essential Bioinformatics. Cambridge University Press.

Yadav, G., Gokhale, R. S., & Mohanty, D. (2003). Computational Approach for Prediction of Domain Organization and Substrate Specificity of Modular Polyketide Synthases. Journal of Molecular Biology, 328(2), 335–363.

Yeh, E., Kohli, R. M., Bruner, S. D., & Walsh, C. T. (2004). TYPE II thioesterase restores activity of a NRPS module stalled with an aminoacyl-S-enzyme that cannot be elongated. Chembiochem A European Journal Of Chemical Biology, 5, 1290–1293.

Yonus, H., Neumann, P., Zimmermann, S., May, J. J., Marahiel, M. A., & Stubbs, M. T. (2008). Crystal structure of DltA. Implications for the reaction mechanism of non-ribosomal peptide synthetase adenylation domains. The Journal of Biological Chemistry, 283(47), 32484–91.

Appendix

1.1 nrps.py

1.1.1 nrps.py script

```
#Designed by O.Reva
import sys, os, string
path = os.getcwd()
sys.path.append(os.path.join(path,"lib"))
import lib

try:
   import psyco
   psyco.profile()
except:
   pass

def getCriteriaFromFile(fname="signatures.txt"):
   file = open(os.path.join("lib",fname),'r')
   line = file.readline()
   if string.find(line,"overlap") == 0:
      overlap = int(string.split(line,":")[1])
   else:
      overlap = 0
   criteria = {}
   while line != None:
      line = file.readline()
      if line == '':
         break
      word,before,after = string.split(line,"\t")
      criteria[word] = [int(before),int(after)]
   file.close()
   return (overlap,criteria)

def save_report(report):
   path = os.path.join("output","report.txt")
   if os.path.exists(path):
      action = "X"
      while action not in ("R","A","F","Q"):
```

```
        action = raw_input("\nFile %s already exists!\nType R - to replace;
A - to append; F - to write to another file; Q - to skip.\n?"
          % os.path.basename(path))
        action = action.upper()
        if action == "Q":
          return
        elif action in ("R","F"):
          if action == "F":
            fname = raw_input("\nEnter new file name: ")
            if not fname or wrong_symbols(fname):
              print "File name %s is not correct\n" % fname
            path = os.path.join("output",fname)
            if os.path.exists(path):
              continue
          f = open(path,"w")
          f.write(report+"\n")
          f.close()
        elif action == "A":
          f = open(path,"a")
          f.write(report+"\n")
          f.close()
        else:
          print "Wrong command %s, try again" % action
    else:
      f = open(path,"w")
      f.write(report+"\n")
      f.close()

def wrong_symbols(fname):
    for symbol in ("\n","\t","<",">",":","|","\\","\"","?","/"):
      if fname.find(symbol) > -1:
        return 1

##################################################
############################

if __name__ == "__main__":
    oLookup = lib.lookup()
    overlap,criteria = getCriteriaFromFile()
    for fname in os.listdir("input"):
      if not oLookup.oIO.is_sequence_file(fname):
        continue
      print fname

oLookup.oIO.saveFasta(oLookup(fname,criteria,overlap),os.path.join("ou
tput",fname[:-3]+"fasta"))
    report = oLookup.get_report()
    if report:
      save_report(report)
```

1.1.2 Example outputs by nrps.py:

Bacillus atrophaeus UCMB-5137, 63Z, UCMB5137

[1298603-1299962]	dir	gene_BacAtrUCMB5137_13580, polyketide synthase of type I
[1477661-1479020]	rev	gene_BacAtrUCMB5137_15000, acyl-CoA dehydrogenase, short-chain specific
[1483779-1485138]	rev	gene_BacAtrUCMB5137_15030, plipastatin synthetase
[1488442-1489801]	rev	gene_BacAtrUCMB5137_15040, unknown
[1489313-1490645]	rev	gene_BacAtrUCMB5137_15050, plipastatin synthetase; gene_BacAtrUCMB5137_15060, plipastatin synthetase
[1491560-1493747]	rev	gene_BacAtrUCMB5137_15060, plipastatin synthetase
[1496589-1498344]	rev	gene_BacAtrUCMB5137_15070, plipastatin synthetase
[1498560-1500927]	rev	gene_BacAtrUCMB5137_15080, plipastatin synthetase; gene_BacAtrUCMB5137_15090, plipastatin synthetase
[1504254-1505586]	rev	gene_BacAtrUCMB5137_15100, plipastatin synthetase
[1506939-1508700]	rev	gene_BacAtrUCMB5137_15100, plipastatin synthetase
[1560504-1561863]	rev	gene_BacAtrUCMB5137_15550, Mycosubtilin synthase subunit C
[1566064-1567402]	rev	gene_BacAtrUCMB5137_15560, Mycosubtilin synthase subunit B
[1570624-1571962]	rev	gene_BacAtrUCMB5137_15560, Mycosubtilin synthase subunit B
[1576675-1578034]	rev	gene_BacAtrUCMB5137_15560, Mycosubtilin synthase subunit B
[1579091-1580423]	rev	gene_BacAtrUCMB5137_15570, Mycosubtilin synthase subunit A
[1589378-1590737]	rev	gene_BacAtrUCMB5137_15580, Malonyl CoA-acyl carrier protein transacylase
[2125842-2127171]	rev	gene_BacAtrUCMB5137_21990, putative transcriptional regulator
[2521884-2523213]	dir	gene_BacAtrUCMB5137_25940, YtmB; gene_BacAtrUCMB5137_25950, putative hydrolase
[2790738-2792505]	rev	gene_BacAtrUCMB5137_28740, siderophore 2,3-dihydroxybenzoate-glycine-threoninetrimeric ester bacillibactin synthetase
[2793654-2795733]	rev	gene_BacAtrUCMB5137_28740, siderophore 2,3-dihydroxybenzoate-glycine-threoninetrimeric ester bacillibactin synthetase
[3153737-3155099]	dir	gene_BacAtrUCMB5137_32530, putative membrane bound transcriptional regulator
[3881513-3882872]	dir	gene_BacAtrUCMB5137_40250, bacitracin synthetase 1
[3883043-3884375]	dir	gene_BacAtrUCMB5137_40250, bacitracin synthetase 1
[4049446-4050805]	dir	gene_BacAtrUCMB5137_41820, surfactin synthetase
[4051021-4052353]	dir	gene_BacAtrUCMB5137_41820, surfactin synthetase
[4054138-4055476]	dir	gene_BacAtrUCMB5137_41820, surfactin synthetase
[4055728-4057087]	dir	gene_BacAtrUCMB5137_41820, surfactin synthetase
[4057157-4059317]	rev	gene_BacAtrUCMB5137_41840, unknown;

		gene_BacAtrUCMB5137_41850, unknown; gene_BacAtrUCMB5137_41860, unknown
[4061864-4063202]	rev	gene_BacAtrUCMB5137_41890, unknown; gene_BacAtrUCMB5137_41900, unknown

Bacillus atrophaeus 1942 chromosome, NC_014639

[1256653-1258012]	dir	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1445911-1447270]	rev	GO_function: GO:0008470 - isovaleryl-CoA dehydrogenase activity, Isovaleryl-CoA dehydrogenase
[1452031-1453390]	rev	hypothetical protein
[1456694-1458731]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1456694-1459589]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1461359-1462691]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1465533-1467288]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1467504-1469517]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1467504-1470375]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1473702-1475034]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1476387-1478148]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1482974-1484993]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1482974-1485851]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1521103-1522462]	rev	
[1526663-1528001]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1531223-1532561]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1537274-1538633]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1539689-1541021]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1549976-1551335]	rev	GO_function: GO:0004314 - [acyl-carrier-protein] S-malonyltransferase activity, Malonyl CoA-acyl carrier protein transacylase
[2559869-2561198]	dir	unknown; Dipeptidyl aminopeptidases/acylaminoacyl-peptidase
[2616585-2617914]	rev	hypothetical protein
[2706030-2707368]	rev	Siderophore biosynthesis non-ribosomal peptide synthetase modules @ Bacillibactin synthetase component F
[2708517-2710596]	rev	Siderophore biosynthesis non-ribosomal peptide synthetase modules @ Bacillibactin synthetase component F
[3065233-3066589]	dir	Cell envelope-associated transcriptional attenuator LytR-

		CpsA-Psr, subfamily F2 (as in PMID19099556)
[3832987-3834346]	dir	Surfactin synthetase subunit 1
[3834517-3835849]	dir	Surfactin synthetase subunit 1
[4011127-4012486]	dir	hypothetical protein
[4012702-4014034]	dir	hypothetical protein
[4015819-4017157]	dir	hypothetical protein
[4017409-4018768]	dir	hypothetical protein
[4023476-4024814]	dir	surfactin synthetase
[4027361-4029557]	dir	surfactin synthetase

Bacillus amyloliquefaciens FZB42, NC_009725

[343289-344648]	dir	hypothetical protein
[344864-346196]	dir	hypothetical protein
[347981-349319]	dir	hypothetical protein
[354092-355451]	dir	Surfactin synthetase subunit 1
[355631-356963]	dir	Surfactin synthetase subunit 1
[360338-361697]	dir	Surfactin synthetase subunit 1
[1711491-1712850]	dir	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1877569-1878928]	rev	
[1887691-1889029]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1893742-1895101]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1896159-1897491]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1906452-1907811]	rev	GO_function: GO:0004314 - [acyl-carrier-protein] S-malonyltransferase activity, Malonyl CoA-acyl carrier protein transacylase
[1933051-1934410]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1937671-1940566]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1946489-1948244]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1948460-1951325]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1954610-1955942]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1956167-1959056]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1963860-1966737]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[2874085-2875417]	rev	peptide synthetase
[2877229-2878561]	rev	peptide synthetase
[2880373-2881705]	rev	peptide synthetase
[2899434-2900763]	dir	unknown; Dipeptidyl aminopeptidases/acylaminoacyl-peptidase
[2934095-2935424]	rev	hypothetical protein
[3020855-3022934]	rev	Polymyxin synthetase PmxB; Siderophore biosynthesis non-ribosomal peptide synthetase modules @ Bacillibactin synthetase component F
[3024083-3026156]	rev	Siderophore biosynthesis non-ribosomal peptide

		synthetase modules @ Bacillibactin synthetase component F

Bacillus subtilis BSn5 chromosome, NC_014976

[81339-82698]	rev	COG1960 Acyl-CoA dehydrogenases, acyl-CoA dehydrogenase
[92562-95481]	rev	COG1020 Non-ribosomal peptide synthetase modules and related proteins, plipastatin synthetase
[101827-103159]	rev	COG1020 Non-ribosomal peptide synthetase modules and related proteins, plipastatin synthetase
[103375-106264]	rev	COG1020 Non-ribosomal peptide synthetase modules and related proteins, plipastatin synthetase
[109532-110864]	rev	potential frameshift: common BLAST hit: gi|255767429|ref|NP_389716.2| plipastatin synthetase, plipastatin synthetase
[118796-121265]	rev	
[120358-121690]	rev	
[1287715-1289791]	rev	COG3251 Uncharacterized protein conserved in bacteria, hypothetical protein; COG1020 Non-ribosomal peptide synthetase modules and related proteins, siderophore 2,3-dihydroxybenzoate-glycine-threonine trimeric ester bacillibactin synthetase
[1290037-1293013]	rev	COG1020 Non-ribosomal peptide synthetase modules and related proteins, siderophore 2,3-dihydroxybenzoate-glycine-threonine trimeric ester bacillibactin synthetase
[2592120-2593479]	dir	potential frameshift: common BLAST hit: gi|255767106|ref|NP_388231.2| surfactin synthetase, surfactin synthetase
[2593689-2595021]	dir	potential frameshift: common BLAST hit: gi|255767106|ref|NP_388231.2| surfactin synthetase, surfactin synthetase
[2596806-2598144]	dir	potential frameshift: common BLAST hit: gi|255767106|ref|NP_388231.2| surfactin synthetase, surfactin synthetase
[2602917-2604276]	dir	
[2604456-2605788]	dir	
[2608323-2610519]	dir	
[4006066-4007425]	dir	COG0318 Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II, polyketide synthase of type I

Bacillus subtilis subsp. subtilis str. 168 chromosome, whole, NZ_CM000487

[377176-378535]	dir	hypothetical protein
[378745-380077]	dir	hypothetical protein
[381856-383194]	dir	hypothetical protein
[387976-389335]	dir	surfactin synthetase
[389514-390846]	dir	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[393392-395585]	dir	
[1794738-1796097]	dir	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1954517-1955876]	rev	GO_function: GO:0008470 - isovaleryl-CoA

		dehydrogenase activity, Isovaleryl-CoA dehydrogenase
[1965750-1968669]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1975014-1976346]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1976562-1979451]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1982719-1984051]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[1991983-1994878]	rev	GO_function: GO:0004467 - long-chain-fatty-acid-CoA ligase activity, Long-chain-fatty-acid--CoA ligase
[3129243-3130572]	dir	unknown; Dipeptidyl aminopeptidases/acylaminoacyl-peptidase
[3279056-3281132]	rev	Polymyxin synthetase PmxB; Siderophore biosynthesis non-ribosomal peptide synthetase modules @ Bacillibactin synthetase component F
[3281378-3284354]	rev	Siderophore biosynthesis non-ribosomal peptide synthetase modules @ Bacillibactin synthetase component F

1.2 nrps.py library (lib.py)

```
#Developed by O. Reva
import os, string, seq_io

class lookup:
   def __init__(self):
      self.fname = ""
      self.criteria = {}
      self.overlap = 0
      self.oIO = seq_io.IO()
      self.report = []

   def __call__(self,fname,criteria,overlap):
      self.fname = os.path.join("input",fname)
      self.criteria = {}
      self.criteria.update(criteria)
      self.overlap = overlap
      return self._get_sequences()

   def _get_sequences(self):
      if not self.criteria:
         return ""
      file_type = self.oIO.is_sequence_file(self.fname)
      if file_type == "GBK":
         self.oIO.openGBK(self.fname)
      elif file_type == "FASTA":
         self.oIO.openFasta(self.fname)
      else:
         return ""
      wordInstances = self.oIO.amcWords(self.criteria.keys())
```

```
      fragments = self._processFragments(wordInstances)
      return self._fasta_format(fragments)

   def _fasta_format(self,fragments):
      if not fragments:
         return ""
      fasta = []
      strands = ("dir","rev")
      seqname = self.oIO.getName()
      if not seqname or seqname == "Unknown sequence":
         seqname = os.path.basename(self.fname)
      self.report.append(seqname)
      for lb,rb,s,seqname in fragments:
         if not seqname:
            seqname = self.oIO.getName()
         fasta.append(">%s|%d-%d|%s" % (seqname,lb,rb,strands[s]))
         fasta.append(self.oIO.getSequence(lb,rb,s,seqname))
         self.report.append("\t[%d-%d]\t%s\t%s" %
(lb,rb,strands[s],self.oIO.getGeneNames(lb,rb)))
      return "\n".join(fasta)

   def _processFragments(self,wordInstances):
      fragments = []
      if not wordInstances:
         return fragments
      for word in self.criteria:
         if len(wordInstances[word]):
            for item in wordInstances[word]:
               strand = item['strand']*2-1
               first = item['position']+self.criteria[word][0]*strand
               second = item['position']+self.criteria[word][1]*strand
               lb = min(first,second)
               rb = max(first,second)
               if lb >= 0 and rb <= self.oIO.getSeqLength(item['seqname']):

fragments.append([lb,rb,item['frame'],word,item['seqname']])
      overlapped = []

      if len(fragments):
         fragments.sort()
         currItem = fragments[0]
         words = [currItem[3],]
         j = 0
         for i in range(1,len(fragments)):
            if (fragments[i][0]<currItem[1] and
fragments[i][2]==currItem[2] and fragments[i][3] not in words):
               words.append(fragments[i][3])
               if (fragments[i][1]>currItem[1]):
                  currItem[1] = fragments[i][1]
```

```
            j = j + 1
          elif j >= self.overlap:
            strand = 0
            if currItem[2] > 2:
              strand = 1

overlapped.append([currItem[0],currItem[1],strand,fragments[i][4]])
            j = 0
            words = []
          else:
            currItem = fragments[i]
            words = [fragments[i][3],]
            j = 0
      return overlapped

  def get_report(self):
    return "\n".join(self.report)
```

1.3 SVG genome mapper (mapper.py)

```
#Designed by O. Reva
import sys, os, string
path = os.getcwd()
sys.path.append(os.path.join(path,"lib"))
import blast,seq_io
from Bio import SeqIO

oIO = seq_io.IO()
svg = []
Y = 25
height = 200
width = 900
'''
for fname in os.listdir("gi"):
  print fname
  for db in ("plasmids","mge","phages"):
    oBlast =
blast.BLAST("blast","protein","bin",os.path.join("db","blastdb",db),os.path
.join("gi",fname),"gi")
    oBlast.execute()
    oIO.save(oBlast.tostring(),os.path.join("output","%s_%s.txt" %
(fname[:-4],db)))

#oIO.save(oBlast.svg(25,Y,width,height),os.path.join("output",seqname+".
svg"))
  #svg.append(oBlast.svg(25,Y,width,height,False))
  #Y += height

for fname in os.listdir("gi"):
```

```
   print fname
   for sbjct in os.listdir("input"):
      print "\t",sbjct
      oBlast =
blast.BLAST("bl2seq","genome","bin",os.path.join("gi",fname),os.path.join
("input",sbjct),"gi")
      oBlast.execute()

oIO.save(oBlast.svg(25,Y,width,height),os.path.join("output","%s_%s.svg"
% (fname[:-4],sbjct[:-4])))
'''

for fname in os.listdir("gi"):
   print fname
   for sbjct in os.listdir("input"):
      if sbjct[:3] != fname[:3]:
         continue
      outfname = "%s_%s.svg" % (fname[:-4],sbjct[:-4])
      print "\t",sbjct
      query_seq = SeqIO.read(os.path.join("gi",fname),"fasta")
      query_seq = query_seq.seq.translate()[:-1]
      sbjct_seq = SeqIO.read(os.path.join("input",sbjct),"fasta")
      sbjct_seq = sbjct_seq.seq.translate()[:-1]
      oBlast =
blast.BLAST("bl2seq","dna","bin",query_seq.tostring(),sbjct_seq.tostring()
,"gi")
      oBlast.execute()

oIO.save(oBlast.svg(25,Y,width,height),os.path.join("output",outfname[:-
4]+"_dna.svg"))
      5/0
'''
svg.insert(0,"<svg xmlns=\"http://www.w3.org/2000/svg\"
viewbox=\"0 0 %d %d\">" % (width,Y+height))
svg.append("</svg>")

oIO.save("\n".join(svg),"pSD_88.svg")
'''
```

1.4 Motifs in linker regions of test strain

The first motif identified within the linkers of the test strain alone was found at 17 different sites and has a width of 9 residues. In each case it was found at the beginning of the very short sequence with the exception of the third linker in srf where it was found at the end of the sequence, which was longer than all the rest of the linker regions where the motif

was identified. Motif 2 identified in the test strain linkers between the domains of all the modules was found in 7 sites and has a width of 16 residues (Appendix Table 1.4.1). The third motif identified in these regions was a long motif covering 44 residues occurring at 3 sites within the srf linkers.

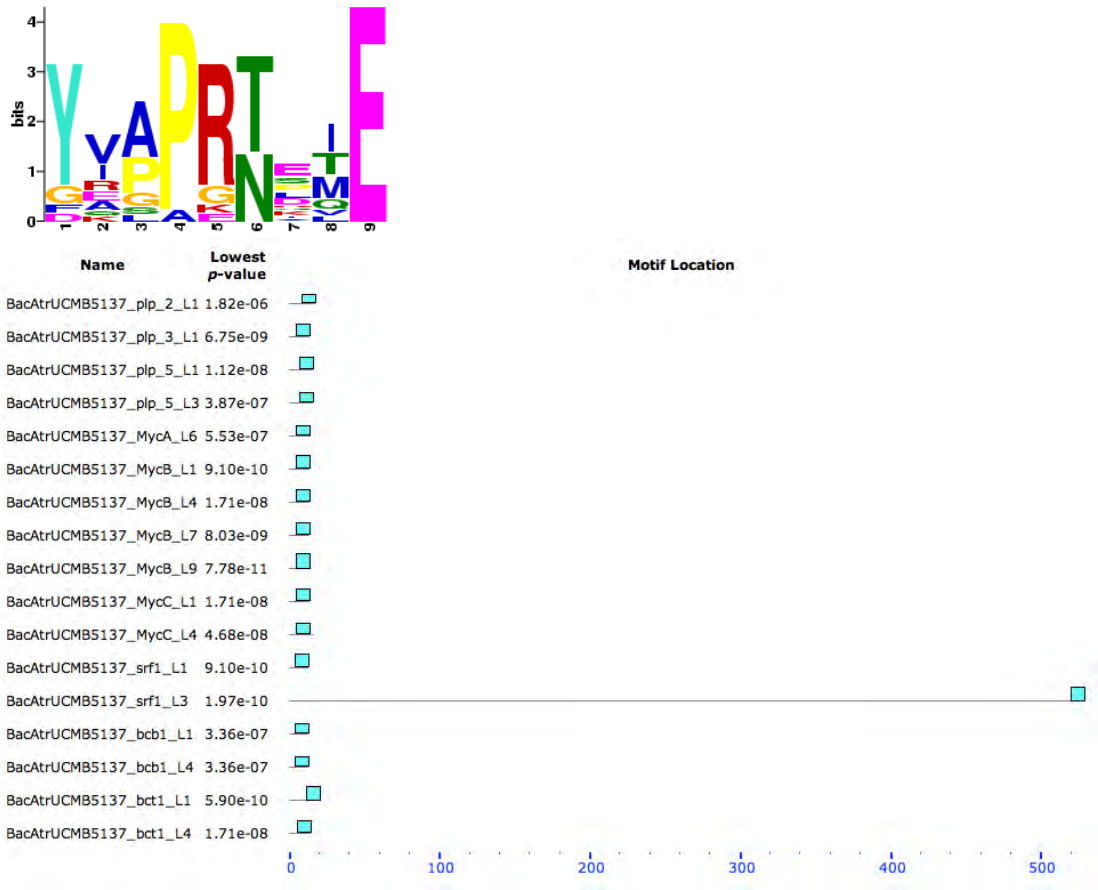
Appendix Table 1.4.1: Motifs identified in the linkers of the test strain in all the NRPSs over all the domains.

Module	Linker	Motif		
		1	2	3
plp 2	1	X		
plp 3	1	X		
	2		X	
plp 5	1	X		
	2		X	
	3	X		
mycA	6	X		
	7		X	
mycB	1	X		
	4	X		
	7	X		
	8		X	
	9	X		
	10		X	
mycC	1	X		
	4	X		
srf	1	X		

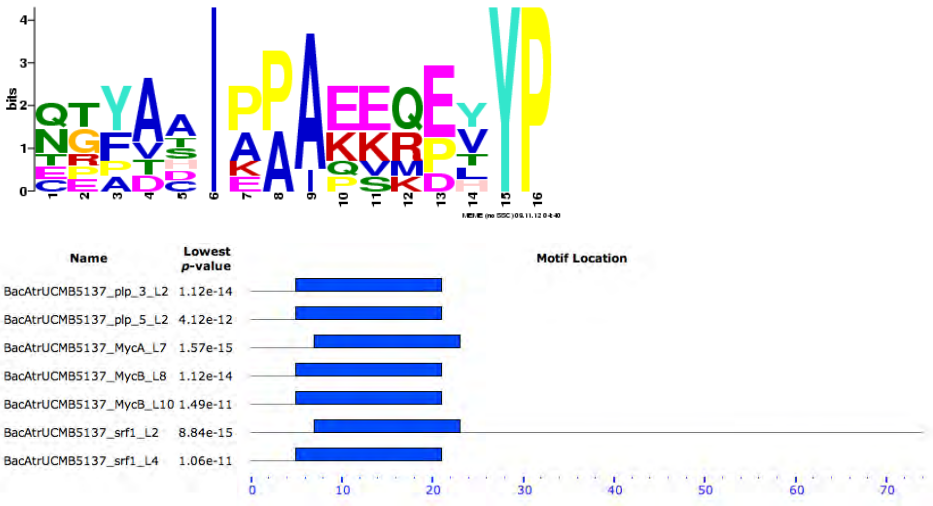
	2		X	
	3	X		X
	4		X	
	5			X(2)
bcb	1	X		
	4	X		
bct	1	X		
	4	X		

Motif 1 gave a regular expression of: YV[AP]PR[TN]E[IMT]E. The motif had an E-value of $2.5x10^{-24}$ and was found in 17 different linkers, which all individually had a p-value below $1x10^{-6}$. The motif was found in the following linkers; srf (Linker 1 &3), mycA (Linkers 4 & 6), mycB (Linkers; 1,4,7 &9), mycC (Linkers 1 &4), plp2 (Linker 1), plp3 (Linker 1), plp5 (linkers 1 & 5), bct (Linkers 1 & 4) and bcb (Linkers 1 & 4). With the exception of linker 3 in srf all the linkers contained the motif at the start of the sequence within the first 5 to 13 residues.

A second motif, occurring at 7 different sites, was found within the test strain linker regions. The sequences were also found within the first few residues of the sequence, within the first 6 to 8 residues. The motif was found in the following sequences; mycA (linker 7), mycB (linkers 8 & 10), srf (linkers 2 & 4), plp3 (linker 2) and plp5 (linker 2). The motif scored and E-value of $9.8x10^{-5}$ and is a width of 16 residues resulting in a regular expression of; [NQ][GT][YF]AAI[PA][PA]A[EK][EK][QR][EP][VY]YP. The p-value of each of the sequences was in all cases below $1x10^{-11}$.

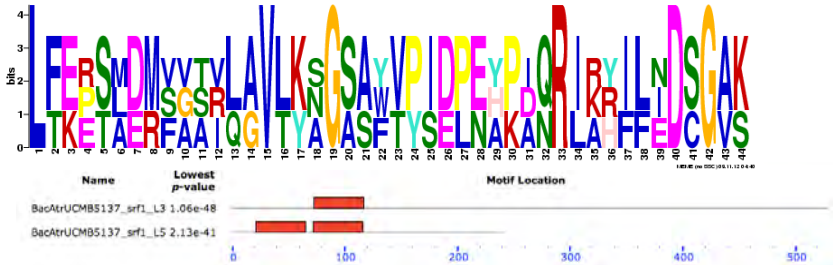


Motif 1 of the linker regions found in the test strain with an E-value of $2.5x10^{-24}$ and found in 17 different linker sequences. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence showing how all are located at the beginning of the sequence with the exception of srf linker 3.



Motif 2 identified in the linker regions of the test strain with an e-value of 9.8x10-5. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.

Finally a third motif was identified in the linkers of the test strain at 3 sites within two linkers; srf linker 3 (at 74 residues) and linker 5 (at 22 & 73 residues) (Figure 2.12). The motif has a width of 44 residues and scored and e-value of 2.4×10^{-2} with each individual occurrence having a p-value lower than 1×10^{-29}. The resulting regular expression of the motif is; L[FT][EK][EPR][ST][ALM][DE][MR][FSV][AGV][AST][IRV][LQ][AG]V[LT][KY][ANS]G[SA][AS][FWY][VT][PY][IS][DE][PL][EN][AHY][PK][ADI][QN]R[IL][AKR][HRY][IF][LF][EIN]D[SC]G[AV][KS].



Motif 3 identified in the linker regions of the test strain within srf synthetase with an e-value of 2.4×10^{-2}.Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.

1.5 Motifs in linker regions in comparison to bcb synthetase

The linker regions of bcb between the test and neighboring strains were compared using MEME and 4 motifs were found with significant e-values (Appendix Table 1.5). The first motif found was found in several linkers in all of the 5 strains. This motif was found to have a width of 12 residues and to have an e-value of 1.2×10^{-38} with the individual p-values all being 2.85×10^{-5} and smaller. The second motif found in the linkers of bcb had an e-value of 3.1×10^{-14} with the individual p-values being 3.99×10^{-11} and smaller. The width of the motif is 16 residues starting at 4th or 6th residue. This motif was found at six sites in four of the strains, the only strain which did not contain it was *B. amyloliquifaciens* FZB42. It was found in linker 2 in the four strains and in linker 5 in the two *B. subtilis* strains. Motif three within the bcb linkers was found in only the two *B. subtilis* strains with an e-value of 9.4×10^{-4} with p-values of 2.02×10^{-16} (*B. subtilis* BS5N) and 7.95×10^{-17} (*B. subtilis subsp. subtilis* 168). The motif spans 12 residues starting at residue 2 of linker 3 in both cases. The final motif was

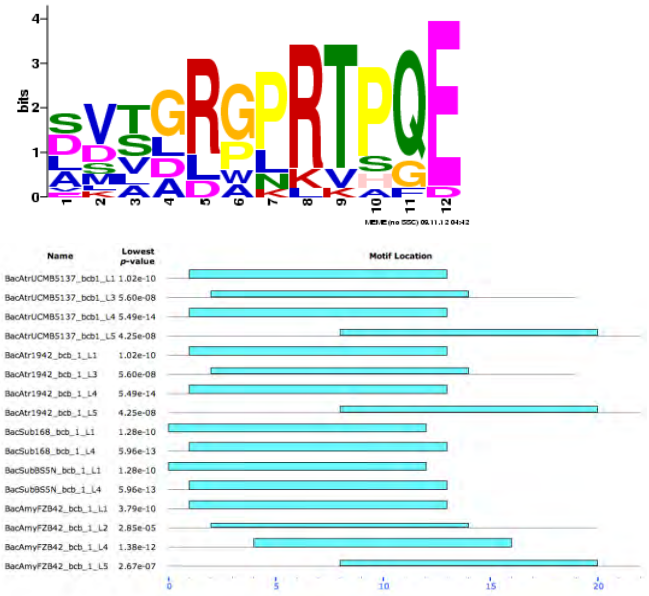
found in linker 5 of four of the strains with the exception being *B. amyloliquifaciens* FZB42. The e-value of the motif was returned as $5x10^{-2}$ with all individual sequences obtaining a p-value of $7.43x10^{-7}$. The motif starts at the first residue in each case and is 5 residues in length.

Appendix Table 1.5: Motifs identified in the linkers between domains of bcb of the test and neighboring strains. Linkers in which the test strain is found is highlighted in yellow and cases where all 5 of the strains display the motif are highlighted in blue. 5137 – *B. atrophaeus* UCMB 5137 (63Z), 1942 - *B. atrophaeus* 1942, 168 – *B. subtilis subsp. subtilis* 168, BS5N- *B. subtilis* BS5N, FZB42 – *B. amyloliquifaciens* FZB42.

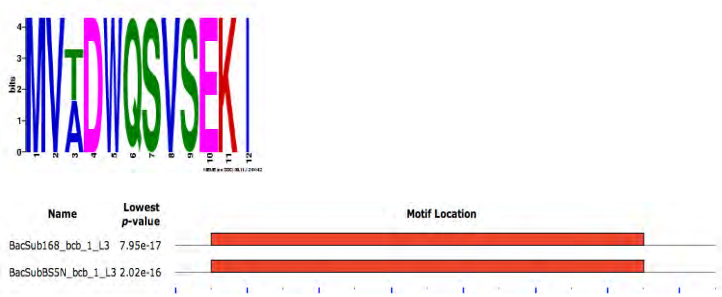
Module			bcb		
Linker	1	2	3	4	5
Motif					
1	5137, 1942, 168,BS5N,FZB42	FZB42	5137, 1942	5137, 1942, 168,BS5N,FZB42	5137, 1942, FZB42
2		5137, 1942, 168,BS5N			168,BS5N
3			168,BS5N		
4					5137, 1942, 168,BS5N

From Appendix Table 1.5 it can be seen that the most conserved motifs between the strains are motif 1 appearing in the first and forth linker of bcb in all 5 strains. It can also be seen that the two B. subtilis strains (168 and BS5N) are more similar containing the motifs in the same areas

whereas the remaining three strains of B. atrophaeus (5137 & 1942) and B. amyloliquifaciens appear to be more closely related with motifs appearing in the same linkers and locations.



Motif 1 of the linkers of the test strain bcb in comparison to the neighboring strains. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



Motif 3 between the linker regions of bcb of the test strain in comparison to the neighboring strains. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.

1.6 Motifs in linker regions in comparison to plp/feng synthetase

Several motifs were detected in the linker regions of the plp/ feng synthetase modules in the test and neighboring strains. Only those found to include the test strain are discussed and for clarity purposes are summarized in the table below. The first motif with an e-value of 1.8×10^{-138} was found in all 5 strains in 22 different sites. It has a width of 16

residues and a regular expression of; [SN]P[YF][AE]AI[EK]PAEK[QR][DE]TYP. This motif starts at either 6 or 9 residues of the sequence with individual p-values being $5.42x10^{-10}$ or smaller. Motif 2 was found 19 times within the linker regions in all 5 of the strains. The motifs e-value was $1.9x10^{-81}$ with individual p-values being $1.56x10^{-9}$ and smaller. The motif with regular expression; [RT][KP][SA][EV][RS][VA][IA][DS]Q[GA]PVEG[ED][VA], is 16 residues in length and starts at the first residue in 18 out of the 19 cases with the only exception being in the case of *B. subtilis* BS5N plp1 linker 7. The third motif was found at 34 different sites within all the 5 strains with an e-value of $5.6x^{-70}$ and individual p-value scores of $5.31x10^{-5}$ and smaller. The motif is 7 residues in length and starts between the second to twelfth residue with the following regular expression; APR[NT][LVE][TL]E. The motif occurred less in the test strain in comparison to the neighboring strains. Motif 8 was the next motif to contain the test strain linker regions. The motif was found once in each of the strains near the end of the sequence at residues 95-98 with the exception of the test strain which was found at the beginning of the sequence at residue 2. The motif had an e-value of 8.4 x10-6 and p-values 8.02x10-9 and smaller. The motif is 9 residues in length and displays the following regular expression; LKAGG[AIS][ILF][VN][PR]. The motif is included in plp1 linker 1 of all the strains except for the *B. subtilis* BS5N strain where it is in plp2 linker 1. The final motif containing the test strain found to be significant was motif 9 with an e-value of $9x10^{-4}$ and individual p-values of $5.31x10^{-17}$ in the two instances in which it was found. The motif was only found in the test strain and the other *B. atrophaeus* strain. The motif is 12 residues in length and starts at residue 4 in both cases. The motif was however found in different modules and different linkers; plp4 linker 3 in *b. atrophaeus* 1942 strain and plp6 linker 1 in the test strain.

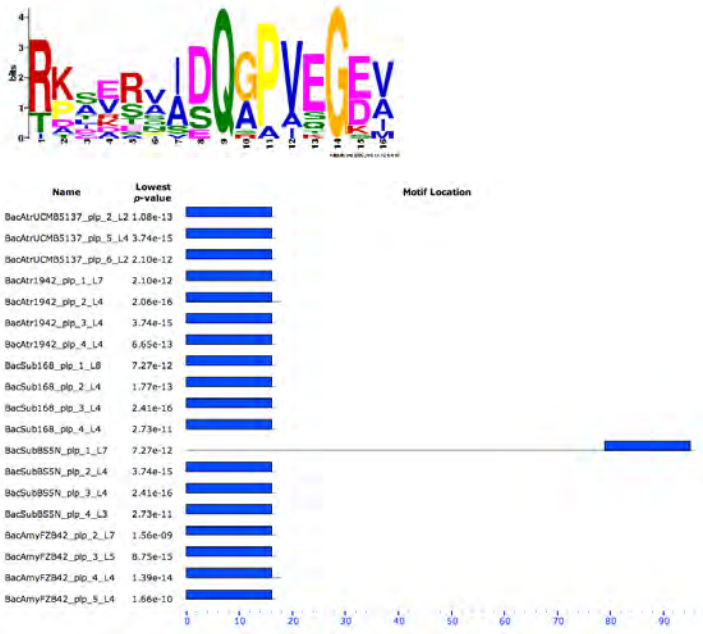
From Appendix Table 1.6 it could be seen that there were linkers wherein a motif was found within all 5 of the strains, however, it was more evident that the same motif in the test strain was often found in a different linker to that of the same motif in the other strains. In this case it was also seen that in several cases the test and *B. amyloliquifaciens* FZB42 were found in

the same location/ linker while the other strains were found in a different linker.

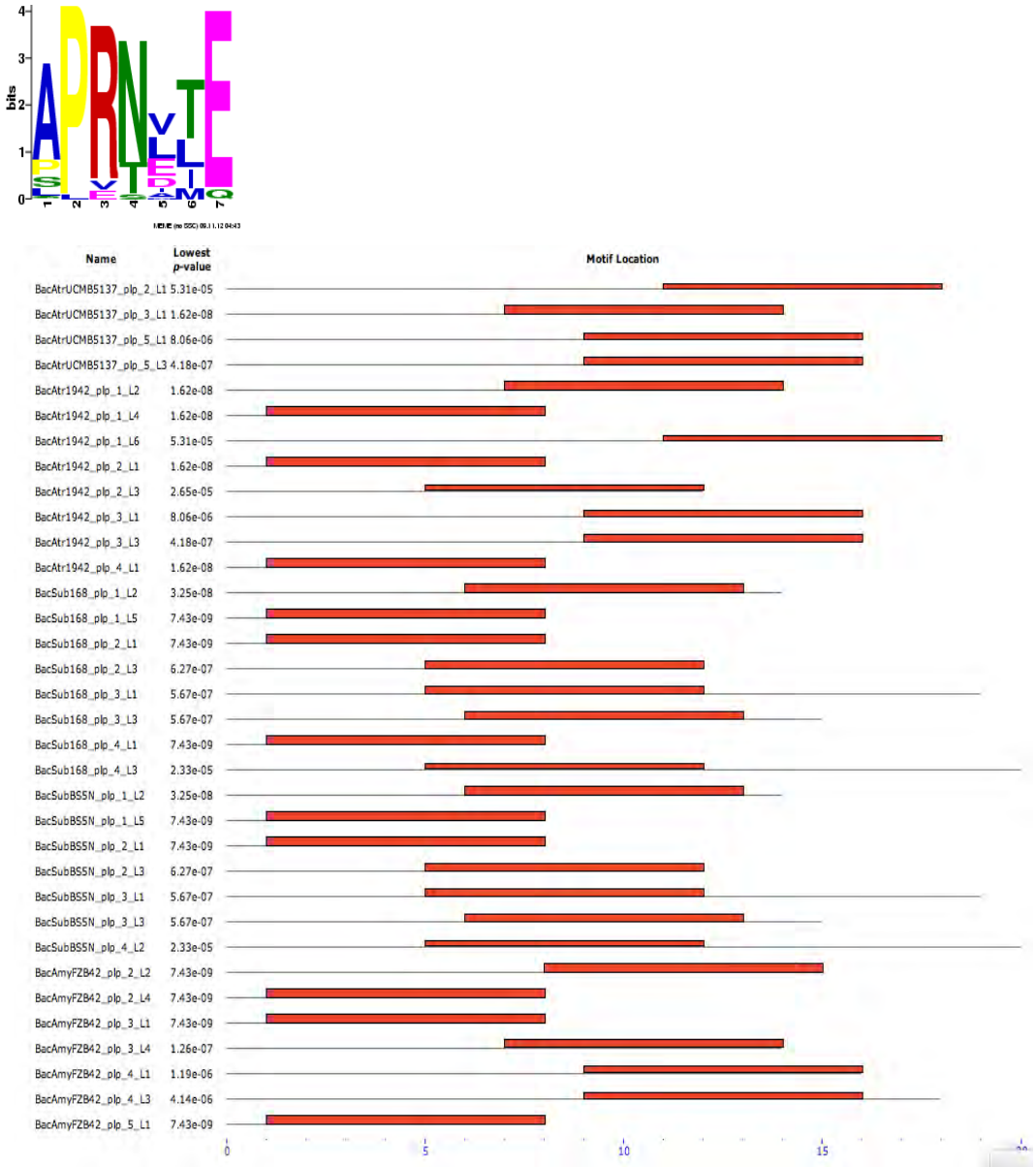
Appendix Table 1.6: Summary of plp/ feng synthetase linker motifs identified in the test and neighbor strains. The linkers in which the test strain occurs are highlighted in yellow. The linkers in which all 5 of the strains occur is highlighted in blue.

Module	1								2						3					4				5				6	
Linker	1	2	3	4	5	6	7	8	1	2	3	4	5	7	1	2	3	4	5	1	2	3	4	1	2	3	4	1	2
Motif																													
1			1942, BS5N, 168		1942	BS5N, 168				1942, BS5N, 168	FZB42		FZB42			5137, 1942, BS5N, FZB42, 168				BS5N	1942, FZB42, 168				5137, FZB42				
2							1942, BS5N	168		5137		1942,BS5N ,168		FZB42				1942, BS5N, 168	FZB42			BS5N	1942, 168, FZB42				5137, FZB42			5137
3		1942, BS5N, 168		1942	BS5N, 168	1942			1942, BS5N, 168	FZB42	1942, BS5N, 168	FZB42		5137	5137, 1942, BS5N, 168, FZB42		1942, BS5N, 168	FZB42		1942, 168, FZB42	BS5N	168, FZB42		5137, FZB42		5137				
8	5137, 1942, BS5N, 168								FZB42																				
9																						1942						5137	

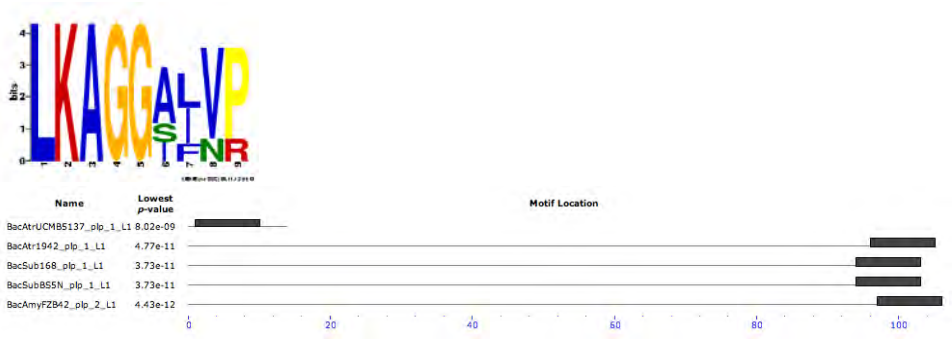
Motif 1 identified in the linker search of plp/feng synthetase in all 5 strains. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



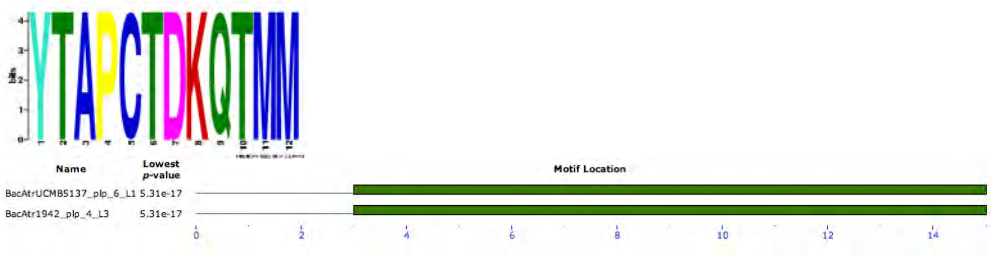
Motif 2 found in the linker regions of the plp/feng synthetase linkers of the 5 strains. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



Motif 3 found in the linker regions of plp/feng. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



Motif 8 found in the linker regions of plp/feng. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



Motif 9 found in plp/feng linker regions. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.

1.7 Motifs in linker regions in comparison to myc synthetase

The linker regions of myc synthetase were also investigated in the 3 strains found to contain myc synthetase; *B. atrpophaeus* UCMB 5137 (63Z), *B. atrophaeus* 1942 and *B. amyloliquifaciens* FZB42. Many motifs were found in the linker regions between the domains of the three subunits, twenty-four of which were found to be significant on assessment of their e and p-values.

Some motifs were identified in the B and C subunits in several different linkers however the most amount of conservation shown in the recognition of motifs was in the second and third linkers of the A subunit. None of the significant motifs were found to overlap on inspection of the data returned from the MAST data. There were several motifs identified in each of these linkers as well as several motifs appearing within both the linkers. The second and third linkers of subunit A are therefore the most conserved areas within the synthetase linkers.

The motifs represented in the table below were the motifs found to be significant. They did differ in the amount of conservation within the sequence represented with the bit scores of the generated logos. Generally in motifs which showed less conservation in terms of their bit scores a pattern was still evident and the amount of sites at which the motif was found within the strains was in most cases higher with a lower overall bit score/ conservation. In cases where the bit scores and conservation was higher the motif was generally found once in a sequence once in each of the organisms (3 times).

There are several cases within these motifs that there are extra occurrences of the motif within the *B. amyloliquifaciens* FZB42 strain that does not occur in the other two *B. atrophaeus* strains.

Appendix Table 1.7: Motifs identified in the linker region search of the myc synthetase modules of *B. atropheaus* UCMB 5137 (63Z), *B. atrophaeus* 1942 and *B. amyloliquifaciens* FZB42. Motifs identified in mycA Linker 3 are shown in blue, linker 2 shown in yellow and motifs found in both mycA linker 2 and 3 are shown in green.

Motif	E-value	Width	# of Sites	Largest p-value	Linker(s)	Bits
1	$7.2x10^{-70}$	57	3	$1.19x10^{-69}$	mycA L3	mostly 4
2	$1.4x10^{-80}$	41	13	$3.1x10^{-14}$	mycA L2 (x2) mycA L3 (x2 except in 1942), mycA L9 (x2 only in FZB42)	low bits
3	$8.4x10^{-63}$	11	27	$1.97x10^{-6}$	mycA L3, L6 (&L4 in FZB42), mycB L1, L4, L7, L9 (&L2, L3, L6 not L4 in FZB42), mycC L1, L4	low bits but pattern
4	$1.6x10^{-72}$	80	4	$1.1x10^{-31}$	mycA L2 (in all 3), mycB L9 (only in FZB42)	mostly 3 or 4
5	$1.8x10^{-134}$	78	6	$2.07x10^{-72}$	mycA L2, L3	mostly 4
6	$8.9x10^{-65}$	57	3	$4.04x10^{-64}$	mycA L2	mostly 4
7	$1.4x10^{-84}$	80	3	$7.46x10^{-96}$	mycA L3	mostly 4
8	$3.5x10^{-75}$	80	3	$7.31x10^{-84}$	mycA L2	mostly 4
9	$1x10^{-69}$	79	3	$2.55x10^{-86}$	mycA L3	mostly 4
10	$2.6x10^{-68}$	80	3	$5.09x10^{-68}$	mycA L2	mostly 4
11	$1.2x10^{-61}$	21	13	$3.64x10^{-15}$	mycA L5, L7, mycB L8, L10, FZB42 only- mycB L8, L9	low bits
12	$1.7x10^{-56}$	57	3	$2.57x10^{-60}$	mycA L2	mostly 4
13	$1.8x10^{-42}$	41	3	$2.72x10^{-50}$	mycA L3	mostly 4
14	$1.1x10^{-35}$	41	3	$4.44x10^{-44}$	mycA L3	mostly 4
15	$9.6x10^{-35}$	15	8	$3.01x10^{-15}$	mycB L2, L5 (63Z, 1942), L4 (FZB42), mycC L2 (all 3)	mostly 3
16	$1.3x10^{-33}$	41	3	$4.9x10^{-47}$	mycA L2	mostly 4
17	$2x10^{-27}$	21	6	$1.11x10^{-21}$	mycA L2, L3	mostly 2 or 3
18	$2.2x10^{-27}$	41	3	$1.69x10^{-39}$	mycA L3	mostly 4
19	$1.5x10^{-18}$	21	4	$2.31x10^{-17}$	mycC L3 (in all 3), mycB L2 (FZB42 only)	mostly 3 or 4
20	$7.5x10^{-12}$	16	3	$2.05x10^{-20}$	mycC L5	mostly 4
21	$1.7x10^{-8}$	21	3	$9.85x10^{-20}$	mycA L1	mostly 4
22	$3x10^{-7}$	20	3	$6.65x10^{-20}$	mycA L4	mostly 3 or 4
23	$2.3x10^{-4}$	11	3	$2.39x10^{-14}$	mycA L3	all 4
24	$2.6x10^{-2}$	8	3	$6.36x10^{-11}$	mycA L3	all 4

1.8 Motifs in terminal regions of test strain

Within the N-terminals of the test strain several motifs were identified by MEME, however, on inspection of the e-values of the motifs, none were found to be significant and were more than likely due to chance occurrence with e-values of 3 and larger.

Within the C-terminals of the modules of the NRPS identified in the test strain there were several motifs identified, three of which were found to be significant. The first two motifs were found in mycB and mycC C-terminals. Motif 1 was identified strating at residue 44 in both terminals with a width of 49 residues. The e-value was reported as $1.3x10^{-9}$ with the individual p-values both being lower than $1x10^{-56}$. The motif was found to have the following regular expression;
CIY[RS]SL[RS][PS][DE]VS[LQ]RIMTM[AT]N[HK]S[EP]MA[AV]Y[LM][IV]L[LM][A
V]GIECLLYKYT[DG][ER][ET][GN][IV]I[LV]G.
The second motif was also within the same sequences starting at residue 162 and 163 respectively. The individual sequences each had a p-value smaller than $1x10^{-54}$ with an overall e-value of $3.7x10^{-4}$. The motif was found to have the following regular expression;
CIY[RS]SL[RS][PS][DE]VS[LQ]RIMTM[AT]N[HK]S[EP]MA[AV]Y[LM][IV]L[LM][A
V]GIECLLYKYT[DG][ER][ET][GN][IV]I[LV]G. The third motif identified had an e-value of $2x10^{-2}$ with individual sequence p-values all being smaller than $1x10^{-13}$. The motif has a width of 12 residues and starts at the first residue of each of the sequences. The motif was found to be in the C-terminals of the following; plp3, plp4 and plp5 with the following resulting regular expression; M[TS][QK][AQ][NST][EGS]IQDIYP.

Appendix Table 1.8 showed that there were several conserved regions within the C-terminal regions of the modules within the test strain. C-terminals of the myc synthetase B and C subunits displayed high amounts of conservation with motifs within both of these regions. The C-terminal regions between the third, forth and

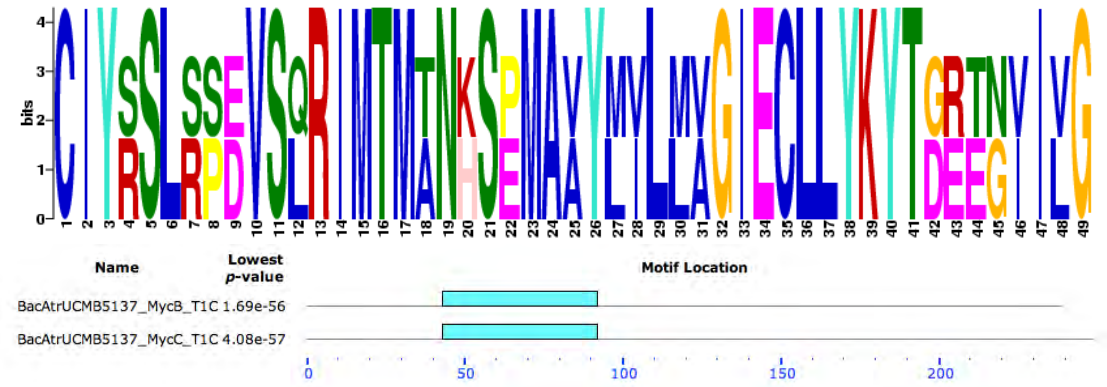
fifth modules of plp/ feng synthetase were also found to be similar to each other with conserved sequence motifs shared between them.

Appendix Table 1.8: Motifs identified in the terminal regions of the test strain, *B. atrophaeus* UCMB 5137 (63Z).

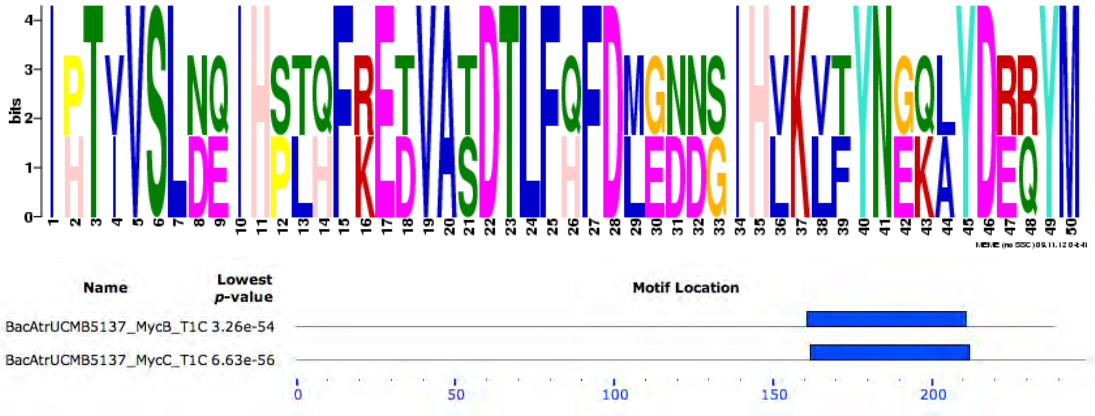
Module	Terminal	Motif		
		1	2	3
plp 3	C			X
plp4	C			X
plp 5	C			X
mycB	C	X	X	
mycC	C	X	X	

Appendix Table 1.8: Motifs identified in the linker regions of myc subunits of 5137 – *B. atrophaeus* UCMB 5137 (63Z), 1942 - *B. atrophaeus* 1942, FZB42 – *B. amyloliquifaciens* FZB42. Linkers in which the test strain and 1942 is found is highlighted in yellow and cases where all 3 of the strains display the motif are highlighted in blue. Cases where the motif was only found in FZB42 are highlighted in red. In the instance where the motif was found more than once within the linker within a strain it is indicated with the number of times it was found.

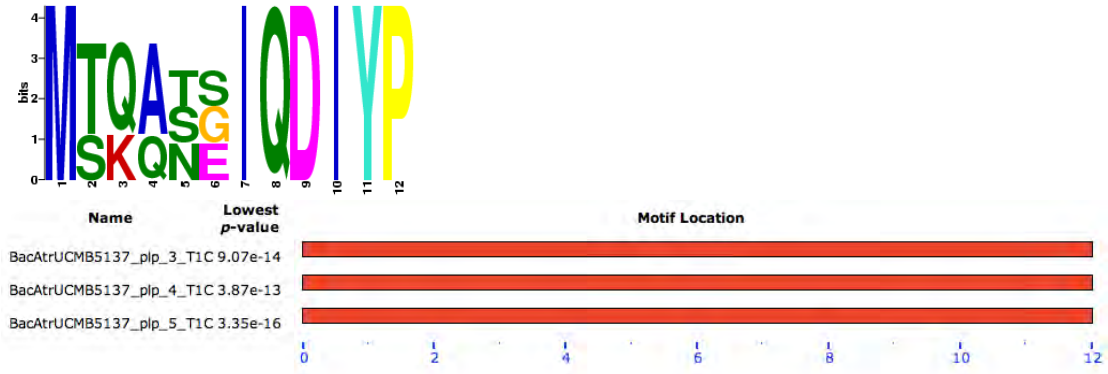
Module	mycA								mycB										mycC					
Linker	1	2	3	4	5	6	7	9	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5
Motif																							
1																							
2								2															
3																							
4																							
5																							
6																							
7																							
8																							
9																							
10																							
11																							
12																							
13																							
14																							
15																							
16																							
17																							
18																							
19																							
20																							
21																							
22																							
23																							
24																							



Motif 1 identified in the C-terminal region of the test strain. E-value of $1.3x10^{-9}$ and width of 49 residues. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



Motif 2 identified in the C-terminal region of the test strain. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence. E-value of $3.7x10^{-4}$ and width of 50 residues.



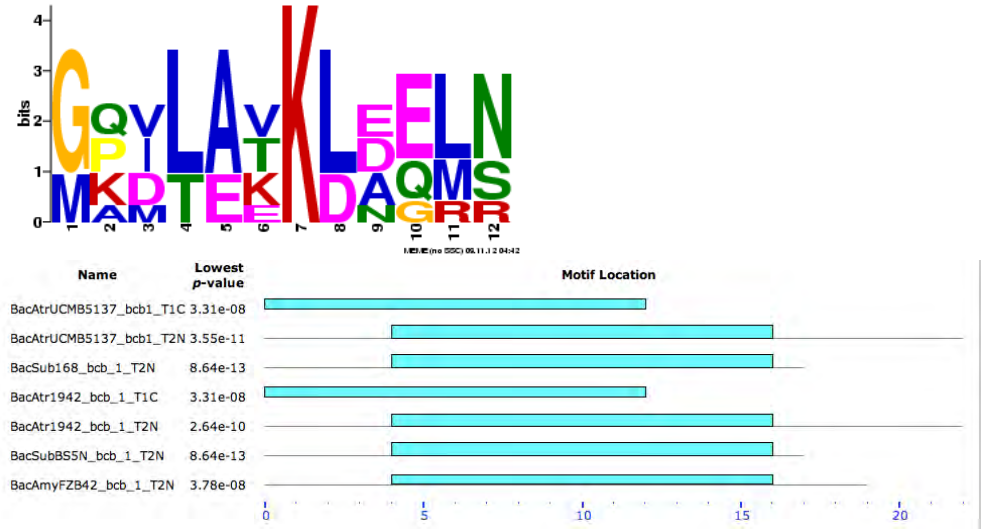
Motif 3 identified in the C-terminal of the test strain. With an e-value of $2x10^{-2}$ and a width of 13 residues in the plp3, 4 and 5. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.

1.9 Motifs in terminal regions in comparison to bcb synthetase

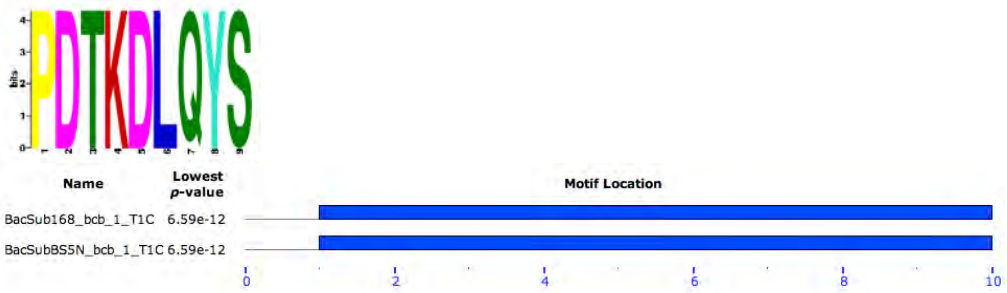
On comparing the terminals of bcb between the strains 4 motifs were found when looking at both N and C terminals. Two of these motifs were found to be significant. The regular expressions were found to be; [GM][KPQ][DIV][LT][AE][KTV]K[LD][ADE][EQ][LM][NS] and PDTKDLQYS with an e-value of $1.1x10^{-6}$ and $8.2x10^{-12}$ respectively. The first motif had a width of 12 and was found at 7 sites starting at either residue 1 or 5 within the following sequences; in the C and N terminals of; *B. atrophaeus* UCMB 5137 and *B. atropheus* 1942 and in the N-terminal of; *B. subtilis subsp. subtilis* 168, *B. subtilis* BS5N and *B. amyloliquifaciens* FZB42. Each of the sequences individually had a p-value of $1x10^{-8}$ or smaller. The second motif was found at the second residue in the C-terminals of *B. subtilis* BS5N and *B. subtilis subsp. subtils* 168 each with a p-value of $6.59x10^{-12}$ and a width of 9 residues.

There were 2 motifs identified in the C-terminals of the bcb between neighboring species. The second was identified as not significant while possibly being part of the first motif. The motif was found in 4 of the 5 strains test with *B. amyloliquifaciens* FZB42 being the only strain it was not found in. The motif has a e-value of $9.9x10^{-4}$ and a width of 8 residues starting at either 3 or 5 residues within the terminal sequence. Each of the individual sequences had a recorded p-value of lower than $1x10^{-8}$ and the overall regular expression was found to be; [DE]TKD[AL]Q[LY]S.

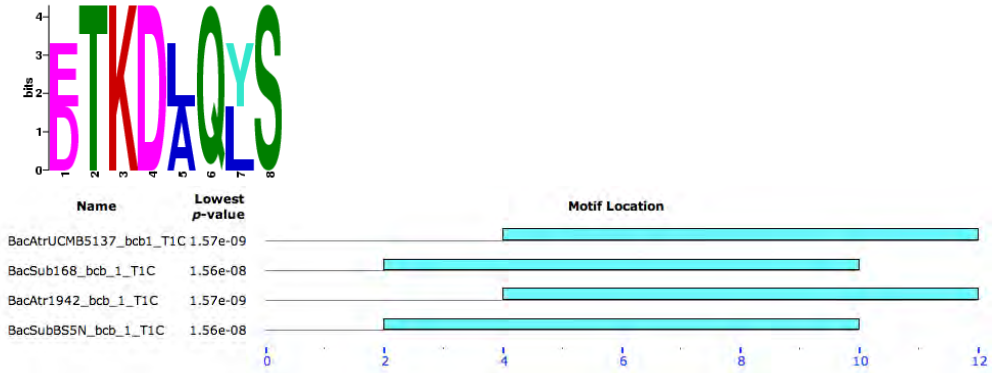
On inspection of the N-terminals of bcb within the test strain and neighboring strain 2 motifs were found, one of which was found to be significant with an e-value of $1.2x10^{-8}$ and individual p-values of $4,27x10^{-10}$ or smaller. The motif was found in all 5 of the neighboring strains N-terminals starting at residue 1 with a width of 16 residues and a regular expression of;
LA[EQD]IG[KQA][IVM]LA[KVE]KL[DEN][EG][LMR][NR].



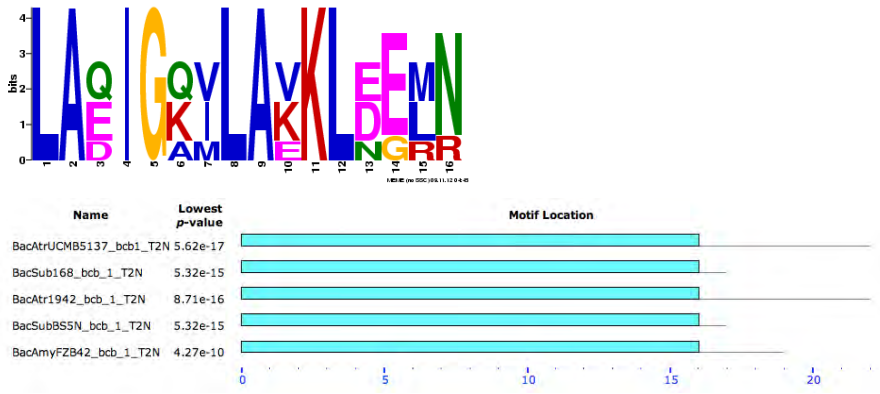
Motif 1 found in the Terminal regions of bcb between the test strain and neighboring strains. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



Motif 2 found in the terminal regions of bcb. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence



Motif 3 identified in the C-terminal region of the strains compared in bcb. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence. E-value of 9.9×10^{-4} and a width of 8 residues



Motif 4 found in the N-terminal of the test and neighboring strains for bcb. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.

Appendix Table 1.9: Motifs identified in the terminal regions of bcb in 5137 – *B. atrophaeus* UCMB 5137 (63Z), 1942 - *B. atrophaeus* 1942, 168 – *B. subtilis subsp. subtilis* 168, BS5N- *B. subtilis* BS5N, FZB42 – *B. amyloliquifaciens* FZB42.. Terminals in which the test strain is found is highlighted in yellow and cases where all 5 of the strains display the motif are highlighted in blue.

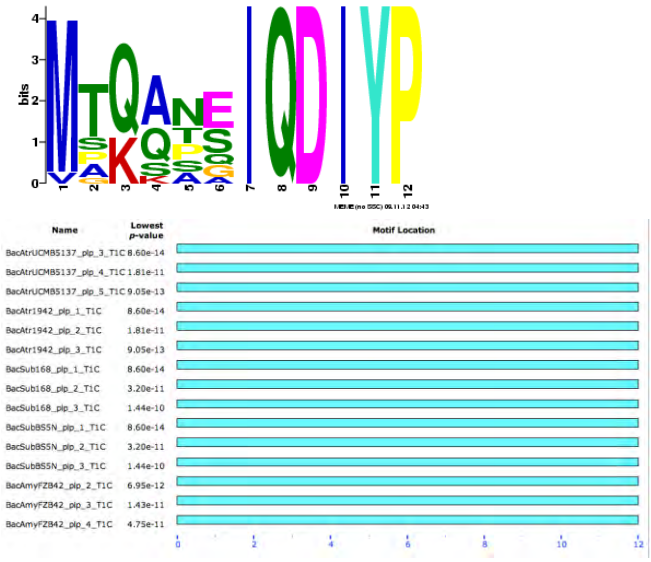
NRPS	bcb	
Terminal	C	N
Motif		
1	5137,1942	5137,1942,168,BS5N,FZB42
2	168,BS5N	
3		5137,1942,168,BS5N
4		5137,1942,168,BS5N,FZB42

Appendix Table 1.9 shows motif 2 was only found in the *B. subtilis* strains (168 and BS5N). Motif 1 and 4 showed that the N-terminal of bcb is highly conserved throughout all 5 strains with the two motifs occurring in the N-terminal of all 5 strains.

1.10 Motifs in terminal regions in comparison to plp synthetase

Four significant motifs were identified in the terminal regions of plp/ feng synthetase. The 3rd motif was only found in test strain and the forth motif was only present in the *B. subtilis* BS5N strain. The first motif in the terminal regions was found in all 5 organisms and at 15 sites but only in the N-terminals. The motif scored an E-value of 2.4×10^{-122} with individual p-values of 2.75×10^{-16} and smaller. The motif starts at the 3rd or 5th residue in each case and has a width of

22 residues. The second motif identified in the terminal search of plp/ feng synthetase terminals was found to have an e-value of $5.6x10^{-89}$ and a width of 12 residues starting in each case at residue 1. This motif was only found in the C-terminals of all 5 organisms in 15 different sites (3 times per terminal) with individual p-values being $2.06x10^{-12}$ and smaller.



Motif 2 identified in the full terminal search of plp/feng being only found in the C-terminal regions. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.

When specifically looking at the N-terminals there were 3 significant motifs found only the first of which was found in all 5 strains with motif 2 and 3 only being present in the test strain. The first motif in this case was the same as the first motif identified in the search with all the terminals however despite producing the same logo, regular expression, width and same number of sites and start different e and p-values were observed. The e-value identified for this motif in this search was found to be $7.8x10^{-98}$ with individual p-values being $3.81x10^{-16}$ and smaller.

When a search was carried out specifically on the C-terminals of plp/ feng synthetase it identified 3 significant motifs. The 2nd motif was only found in the other 4 strains and was not present in the test strain with the 3rd motif only being present in the test strain. The first motif, which was found in all 5 strains,

was the same as the second motif identified in the motif search carried out with all the terminal regions. This was also however found to display differing e and p-values as opposed to the same motif when identified in the search with all the terminal regions even though the width, number of sites and start sites were the same. The e-value in this case was recorded as 1.9×10^{-75} and the individual p-values were 1.44×10^{-10} and smaller.

Appendix Table 1.10 showed that the terminal regions of plp/ feng synthetase are quite conserved, the N-terminals slightly more so than the C-terminals, in all the strains. It is also evident that the test strain in several cases exhibits the motif in a different terminal to that of the other strains.

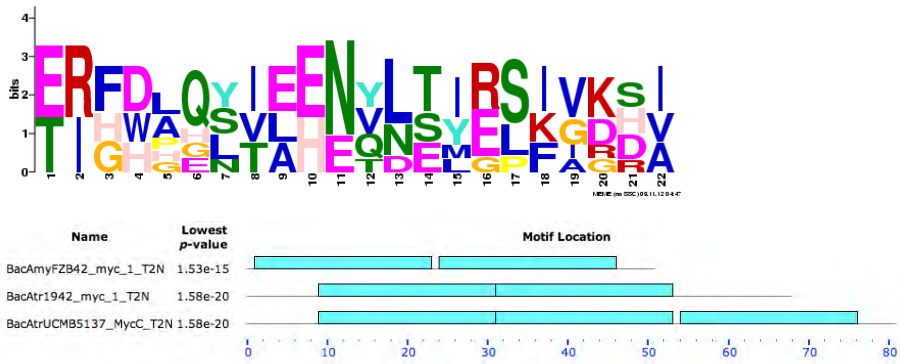
Appendix Table 1.10: Motifs identified in the terminal regions of plp/feng in in 5137 – *B. atrophaeus* UCMB 5137 (63Z), 1942 - *B. atrophaeus* 1942, 168 – *B. subtilis subsp. subtilis* 168, BS5N- *B. subtilis* BS5N, FZB42 – *B. amyloliquifaciens* FZB42.. Terminals in which the test strain is found is highlighted in yellow and cases where all 5 of the strains display the motif are highlighted in blue. Cases where the motif occurred in all the other strains but not the test strain are highlighted in green.

NRPS	plp										
Module	1		2		3		4		5		
Terminal	C	N	C	N	C	N	C	N	C	N	N
Motif											
1		1942,168,BS5N		5137,FZB42		1942,168,BS5N		1942,168,BS5N,FZB42		5137,FZB42	5137
2	1942,168,BS5N		1942,168,BS5N,FZB42		5137,1942,168,BS5N,FZB42		5137,FZB42		5137		

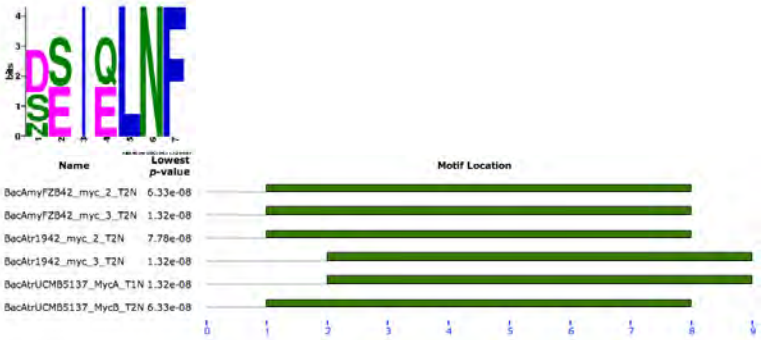
1.11 Motifs in terminal regions in comparison to myc synthetase

A MEME search was run on all the terminal regions of the three myc subunits of the 3 strains containing the myc synthetase modules; *B. atrophaeus* UCMB 5137 (63Z), *B. atrophaeus* 1942 and *B. amyloliquifaciens* FZB42. Several of the motifs identified while they had good e and p-values displayed very low bit scores and showed a large amount of variation in the motifs. Of the 9 identified motifs found to be significant 2 were found to be in the N-terminal regions only, 6 in the C-terminals only and 1 in both the C and N-terminals of the myc subunits.

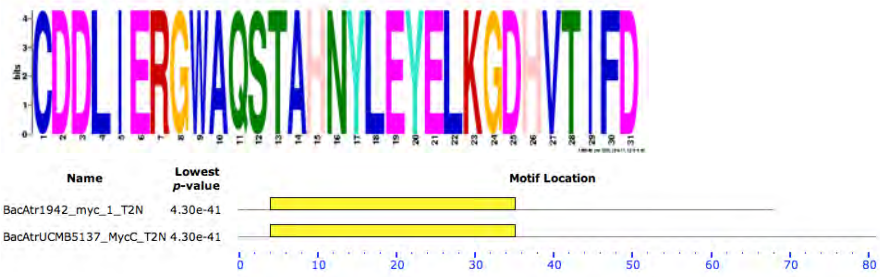
On specifically looking at the N-terminals of myc 2 significant motifs were identified. The first motif was not identified in the search when all the terminals were searched. This motif has an e-value of $5.6x10^{-14}$ with individual p-values being $2.23x10^{-13}$ and smaller. The motif was found in 7 sites in the 3 strains; 3 times in the test strain and twice in each of the neighboring strains. The motif is 22 residues in length. The second motif identified was the same as the ninth motif identified by the complete terminal search carried out. The same logo and regular expression were found with identical width, number of sites and start sites but differing e and p-values. The e-value in this search was $5.8x10^{-5}$ and the individual p-values were $2.26x10^{-7}$ and smaller as apposed to $2.4x10^{-1}$ and $7.78x10^{-8}$ in the full terminal search respectively. This motif was found in the A and B subunits of all three organsims. The remaining motif identified during the full terminal search found only in the N-terminal was not found during this search even though it displayed a bit score of 4 for the entire width of 31 residues. The e-value of this motif was found to be $5.3x10^{-13}$ with individual p-values being $4.3x10^{-41}$ in both sites. This motif was not found in the *B. amyloliquifaciens* FZB42 strain.



Motif 1 identified in the N-terminal of myc. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



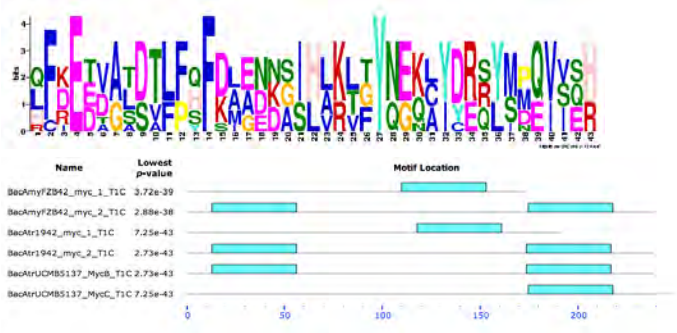
Motif 2 identified in the myc N-terminal search and the 9th motif identified in the myc full terminal search. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



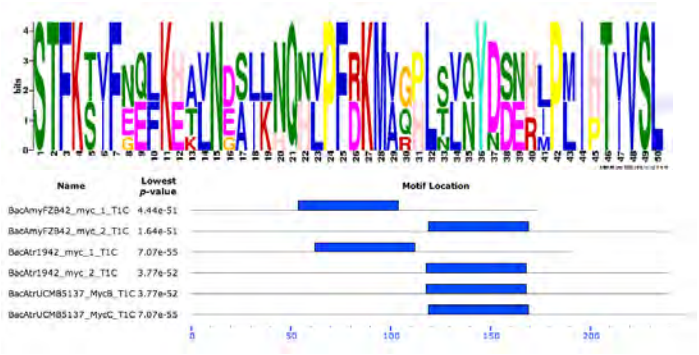
Motif 5 identified in the full terminal search found only in the N-terminal region of the myc synthetase terminals. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.

The C-terminal search identified 7 significant motifs in comparison to the 6 found in the search using all the terminals. The first motif identified in the C-terminal search is the same as the first motif identified in the full terminal search with differing e and p-values. In the full terminal search the values were found to be 1.6×10^{-120} as apposed to in the C-terminal only search where it was 6.6×10^{-119}. This motif has a width of 43 residues and is found in all 3 strains in the C-terminals of the B and C subunits. It did however have rather low bit scores

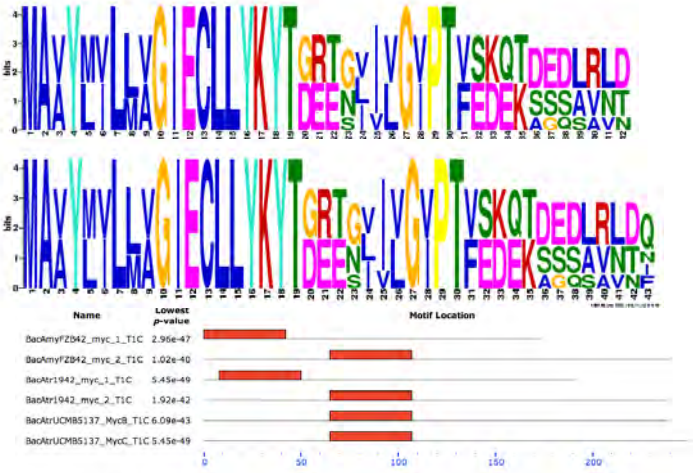
showing a large amount of variation present. The second motif was identical to that found in the full terminal search with an e-value of $1.3x10^{-110,}$ and p-values of $4.44x10^{-51}$ and smaller . The motif is 50 residues wide and is found in 6 sites in the three strains. The third motif identified was the same in the two searches with the search of all the terminals returning the same motif one residue longer (43 residues). Motif 7 identified in the full terminal search was the same as that 6th motif identified in the C-terminal only search however the e-value was different and the width in the C-terminal only search was one residue shorter than that generated in the full terminal search. The motif was found in 4 sites in the 3 organisms in each of the strains myc synthetase subunit B C-terminals as well as the myc synthetase subunit C C-terminal in the test strain. The e and p-values for this motif were differing in the two searches with $1.3x10^{-5}$ and $5.33x10^{-9}$ and smaller in the full search as apposed to $7.7x10^{-6}$ and $1.73x10^{-9}$ and smaller in the C-terminal only search. The 6th motif identified in the full terminal search was the same as the 7th motif identified in the C-terminal only search. Both motifs were found in 3 sites – the myc synthetase C subunit C-terminal in the 3 strains. The width of the motif was 11 residues starting at residue 54 in all cases. The bit scores of the logo are all 4 for the whole width. The e and p-values differed in the two searches though being $2.5x10^{-8}$ and $7.04x10^{-15}$ in the full search and $9.8x10^{-4}$ and $9.52x10^{-13}$ and smaller in the C-terminal only search. The 8th motif identified in the full terminal search, containing sequences of C-terminals only, was not found in the C-terminal search. The bit scores were mainly 4 for the width of the motif of 9 found in all 3 strains at 3 sites (once in each) within the myc synthetase subunit C C-terminals. The e-value of the motif was $1.5x10^{-2}$ and the p-values were $6.84x10^{-11}$ and smaller.



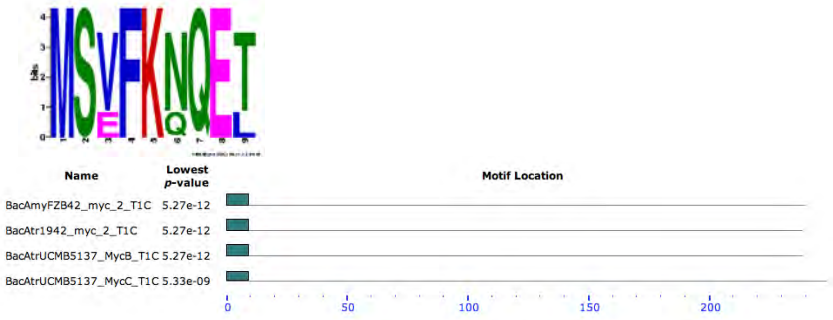
Motif 1 identified in the C-terminal motif search on myc. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



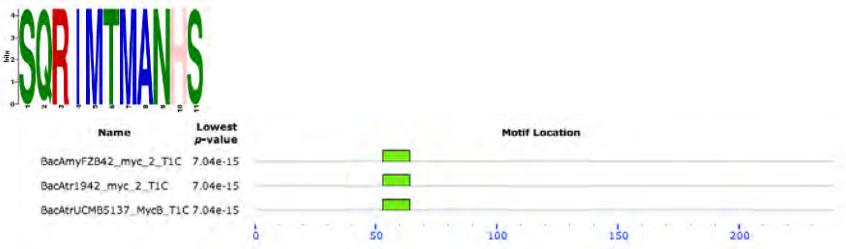
Motif 2 identified in the C-terminal of myc. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



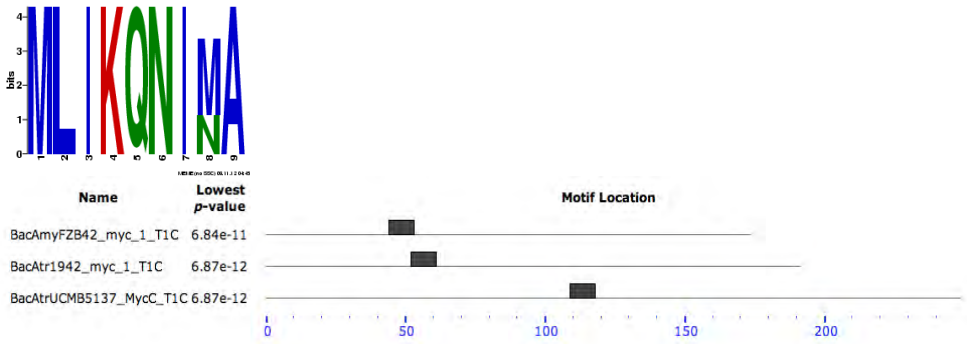
Motif 3 identified by the myc synthetase terminal search. Top: C-terminal only search motif, Middle: full terminal search motif, Bottom: Mast results showing relation in sequence.



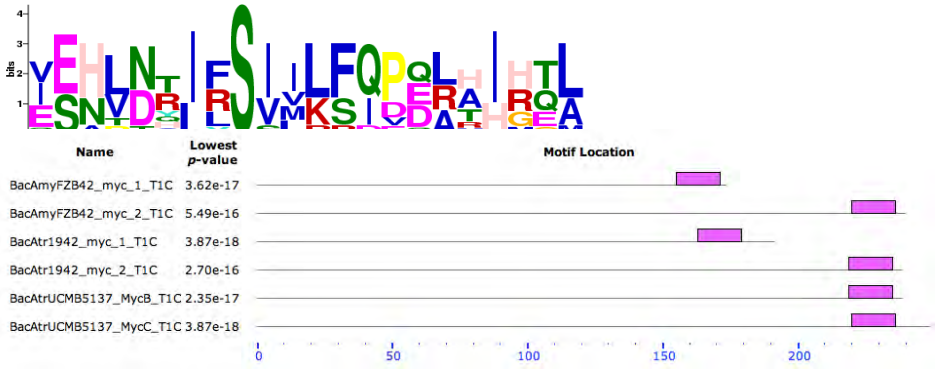
Motif 7 identified in the full terminal search and motif 6 in the C-terminal search for terminal motifs in myc. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



Motif 6 identified in the full terminal search and motif 7 identified in the C-terminal only search. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



Motif 8 identified in my C-terminal during the full terminal MEME search. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.



Motif 4 identified in the C-terminals of myc. Top: MEME logo of the motif, Bottom: MAST results of the motif in relation to the sequence.

The forth motif identified was the same as forth motif found in the full terminal search. In the full terminal search though this motif was found in the B and C myc synthetase subunits as well as in the N-terminal of the C subunit of the test strain both N and C-terminals. The e-value in both cases was 2.2×10^{-34} and the p-values 5.73×10^{-11} and smaller. The motif is 22 residues in length and was found at 22 sites in the 3 strains.

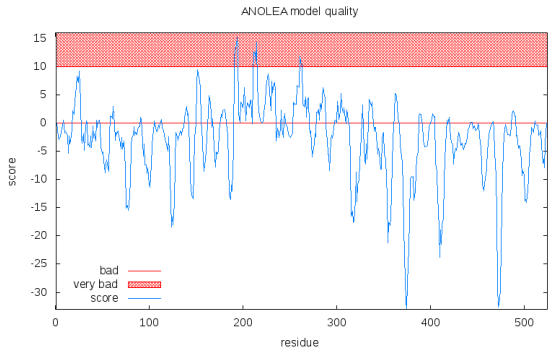
Appendix Table 1.11 showed that there is a large amount of conservation in the terminal regions of the myc synthetase subunits between the different strains as well as within the different subunits. The C-terminals of the B and C subunits appear to be the most conserved through the tested strains. The N-terminals of the A and B subunits seem to be conserved within the subunits as well as between the strains.

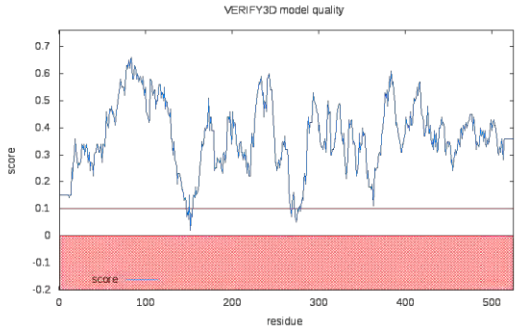
Appendix Table 1.11: Motifs identified in the terminal regions of myc synthetase subunits in 5137 – *B. atrophaeus* UCMB 5137 (63Z), 1942 - *B. atrophaeus* 1942, FZB42 – *B. amyloliquifaciens* FZB42. Linkers in which the test strain is found is highlighted in yellow and cases where all 3 of the strains display the motif are highlighted in blue.

Module	mycA	mycB		mycC	
Terminal	N	C	N	C	N
Motif					
1					5137(3),1942(2),FZB42(2)
2	5137, 1942, FZB42		5137, 1942, FZB42		
5					5137,1942
1		5137(2),1942(2), FZB42(2)			5137,1942,FZB42
2		5137,1942,FZB42		5137,1942,FZB42	
3		5137,1942,FZB42		5137,1942,FZB42	
7		5137,1942,FZB42		5137	
6		5137,1942,FZB42			
8				5137,1942,FZB42	
4		5137,1942,FZB42		5137,1942,FZB42	

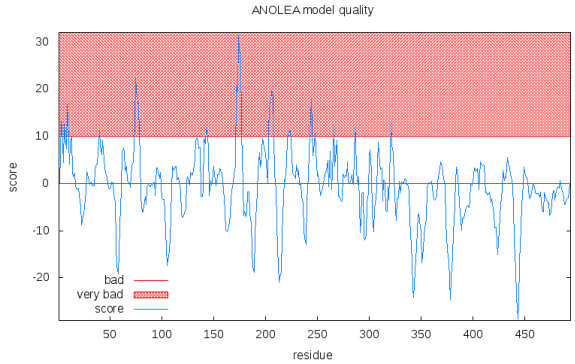
1.12 Verify3D and Anolea Results:

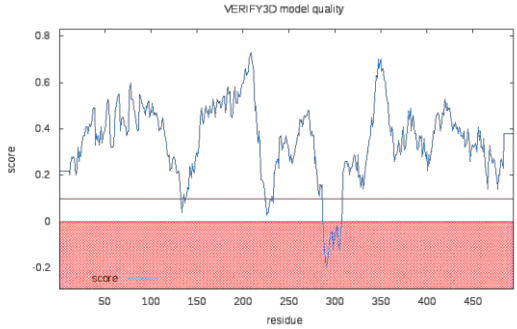
Plp/feng synthetase Module 2 A-domain model



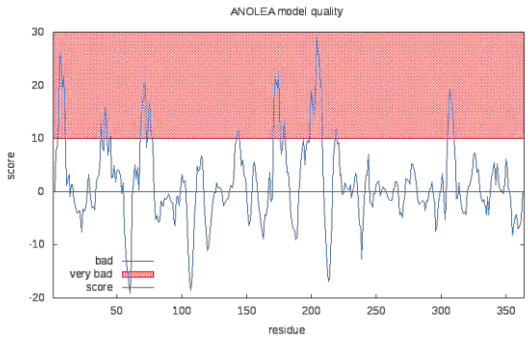


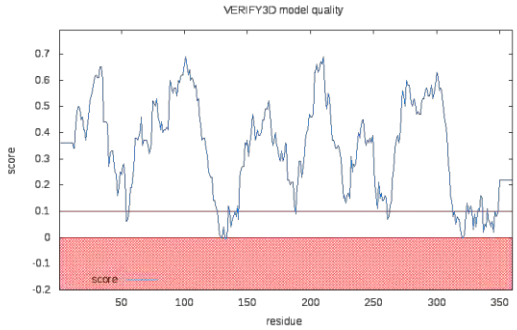
Plp/ Feng synthetase module 3 A-domain model



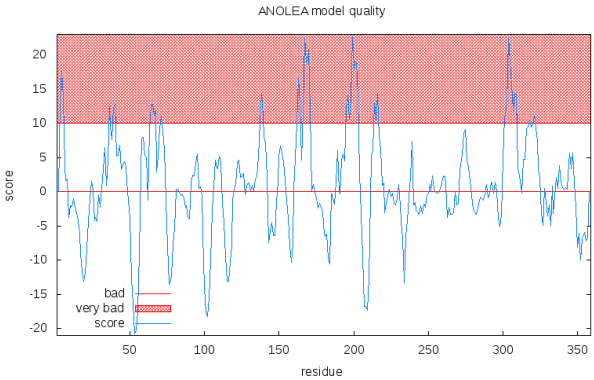


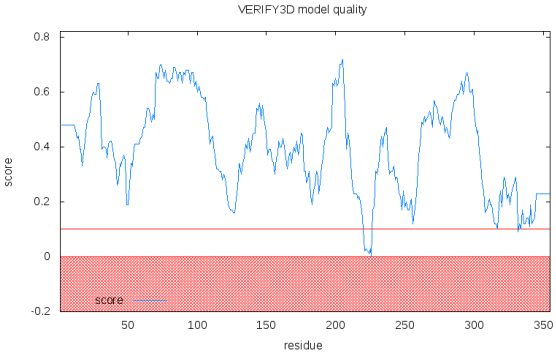
Plp/feng synthetase Module 5A A-domain model



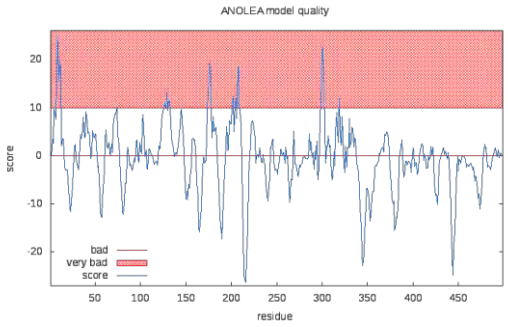


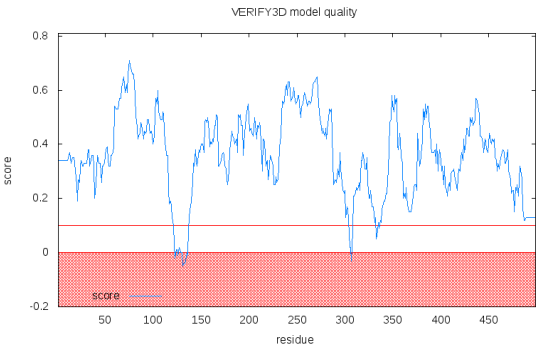
Plp/feng synthetase Module 5B A-domain model



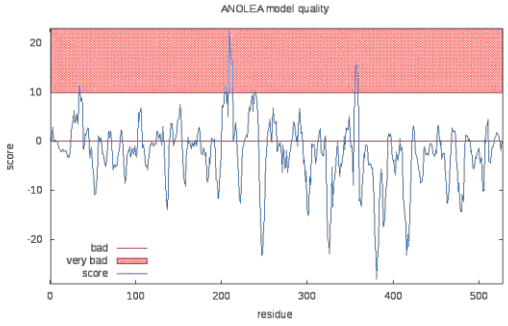


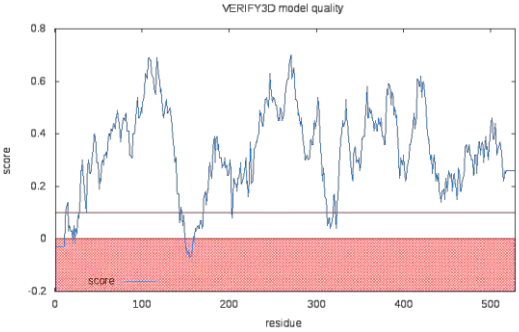
Plp/Feng synthetase module 6 A-domain model



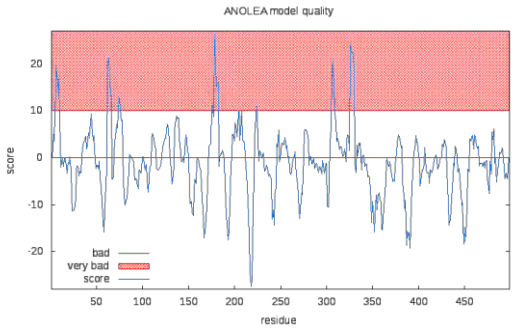


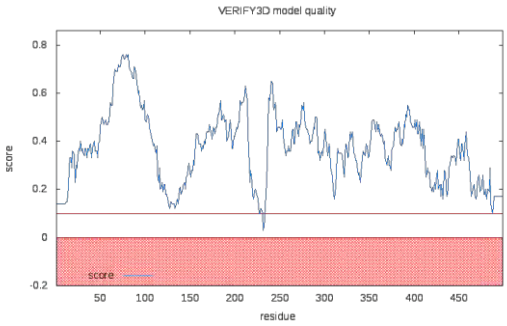
Plp/ feng synthetase module 7 (A) A-domain model.

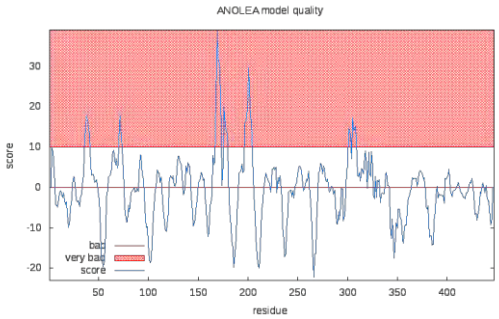


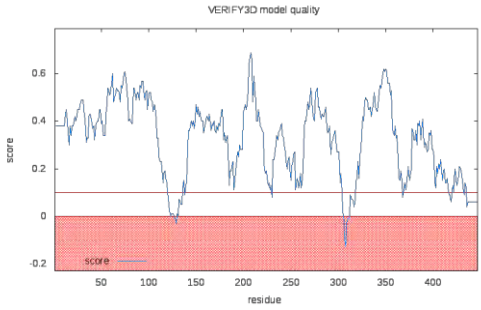


Plp/ feng synthetase module 7 (B) A-domain model



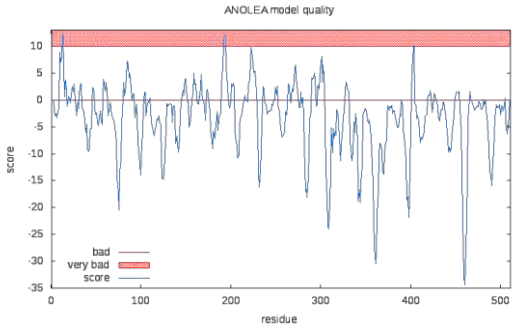


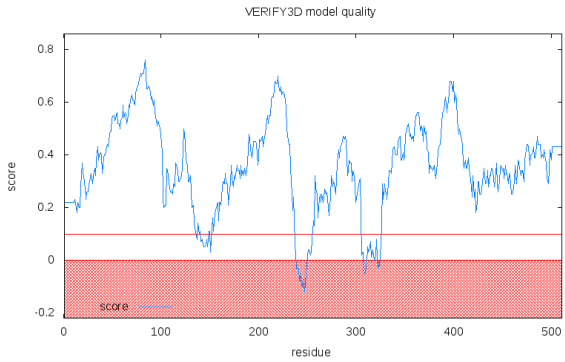




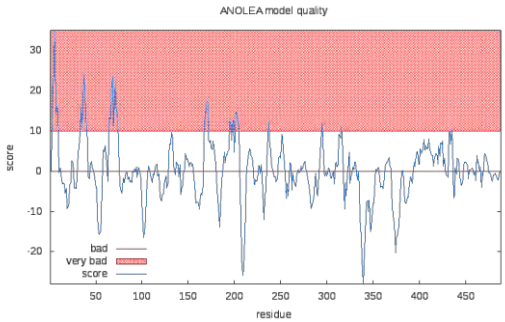
Plp/feng synthetase module 8 A-domain model

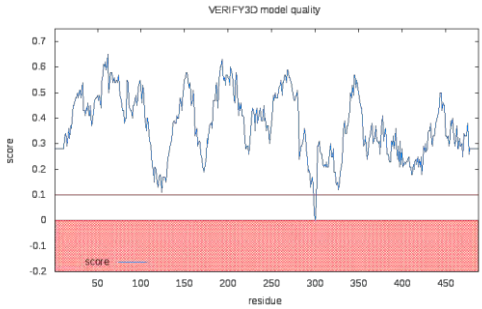
Plp/feng synthetase module 9 A-domain model



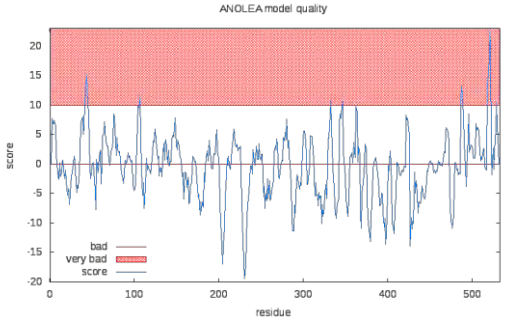


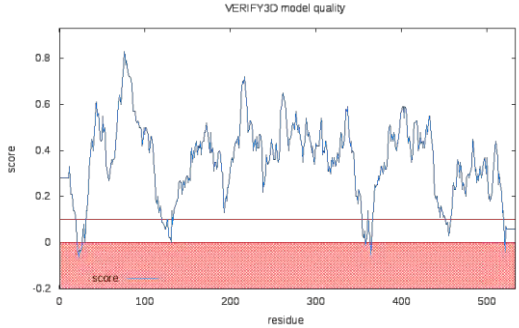
Plp/feng synthetase module 10 A-domain model



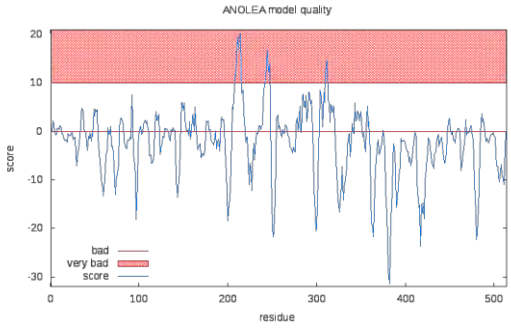


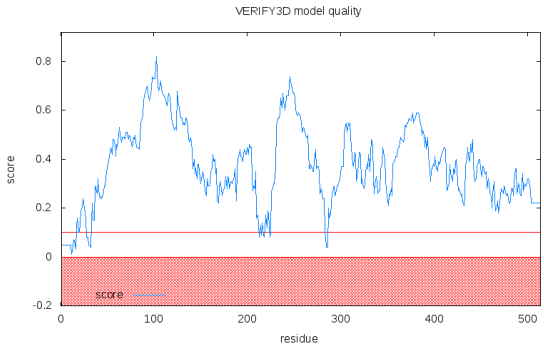
Myc synthetase subunit A module 1 A-domain model



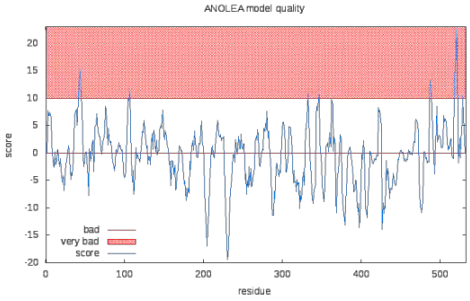


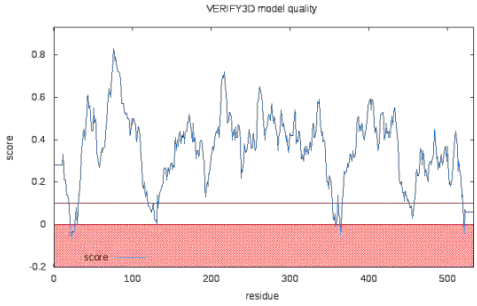
Myc synthetase subunit A module 2 A-domain model



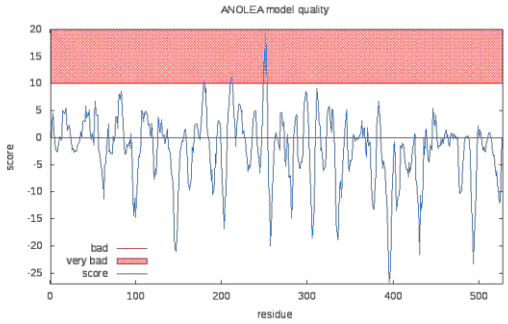


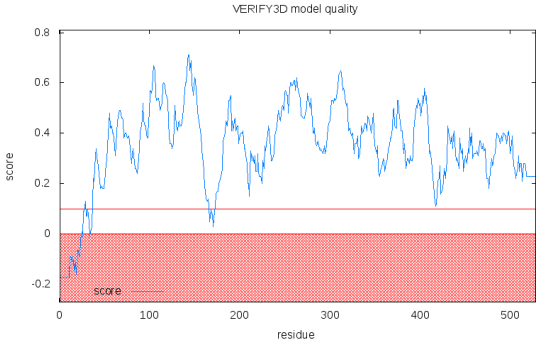
Myc synthetase subunit B module 1 A-domain model





Myc synthetase subunit B module 3 A-domain model





Myc synthetase subunit C module 2 A-domain model