Modelling polyketide synthases and related macromolecular complexes

by

Rohit Farmer

A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy (PhD)

College of Life and Environmental Sciences
School of Biosciences
The University of Birmingham
January 2015

DECLARATION

Student's declaration

This thesis presents the original work done by me at the University of Birmingham and I have written all the parts of the thesis myself. The data presented in the Chapter 3 for which the methods are described in Chapter 2 were published in Haines *et al.* (2013).

Rohit Farmer

Declaration by co-authors of Haines *et al.* (2013)

The results and the methods presented in Chapter 3 and Chapter 2 respectively were previously published in Haines *et al.* (2013). There are some strong similarities in writing between parts of Chapter 2 and 3 and Haines *et al.* (2013), however the work related to these sections was done by Rohit Farmer and he was responsible for writing the text that formed those parts of the published paper.

Dr. Anthony S. Haines	**Dr. Peter J. Winn**	**Prof. Christopher M. Thomas**
First author	Supervisor	Second supervisor

Abstract

Polyketide synthases (PKS) are the enzyme complexes that synthesise a wide range of natural products of medicinal interest, notably a large number of antibiotics. The present work investigated the mupirocin biosynthesis system, in comparison with similar pathways such as thiomarinol and kalimantacin, with the focus on the structural modelling of the protein complexes involved in antibiotic synthesis via the tools of structural bioinformatics. β- methylation in the mupirocin pathway is catalysed via enzymes encoded by the "HMG-CoA synthase (HCS) cassette"and replaces a β-carbonyl with a methyl group. To understand better, what might allow the HCS cassette to recognise β-branch associated ACPs, molecular modelling was used to explore the interaction of the ACPs with MupH (the HMG-CoA synthase homologue). Hidden Markov models (HMM) were used to classify ACPs as branching and non-branching. HMM analysis highlighted essential features for an ACP to behave like a branching ACP. A homology model of MupH was docked with the NMR structure of each of ACPs mupA3a and mupA3b. The docking results were also supported by the biological context based on analysis of phylogenetic variations in amino acid conservation and physical properties of the interface residues. Modelling and mutagenesis identified helix III of the ACP as a probable anchor point of the ACP:HCS complex. The position of this helix is determined by the core of the ACP and substituting the interface residues modulates the interaction specificity.

Although docking and mutagenesis studies performed on ACP:MupH complex laid down the general specificity rule of ACP:MupH recognition in the β-branching. BatC from closely related kalimantacin cluster cannot substitute for the function of MupH. However, Prof. Thomas' group have shown upon mutating BatC it is possible to successfully complement a Δ*mupH* mu-

tant. β-branching ACPs from the thiomarinol cluster complements the β-branching ACPs in the mupirocin cluster however, the β-branching ACP from the kalimantacin cluster does not. These observations suggests a pair wise specificity between the ACP:HCS proteins in the β-branching.

Molecular dynamics simulation of the ACP:MupH complex revealed large movements of the surface loops at the opening of the active site in MupH. These movements were found to be greater in a MupH monomer as compared to the MupH dimer structure and may assist in the accommodation of the ligand inside the MupH active site.

Molecular dynamics simulations of apo, holo and acyl forms of ACPs in the mupirocin cluster revealed that the PKS ACPs form a cavity upon the attachment of the phosphopantetheine and acyl chains similar to what is seen in the fatty acid synthase (FAS) ACPs. The cavity formed does not form a deep tunnel as in the FAS ACPs but, is rather solvent exposed cleft, enabling the polar groups on the acyl chain to hydrogen bond with the solvent. It was also observed that a bulky residue (I61 in ACP-mupA3a) in the PKS ACPs is likely to prohibit the formation of a deep tunnel as opposed to the case of smaller residue (alanine) at the equivalent position in the FAS ACPs, where a deep tunnel is seen.

Molecular docking of the cognate substrate with the ketosynthase (KS) homo dimer of module 5 of the MmpA in the mupirocin pathway revealed a loop at the dimer interface that appears to be responsible for the recognition specificity of α-hydroxylated substrate. Mutagenesis experiments showed that, upon swapping this recognition loop from a KS which does not bind an α-hydroxylated substrate, the pathway produces a full length product in the absence of MupA, the enzyme thought to be responsible for α-hydroxylation, whereas the wild type KS Δ*mupA* mutant does not. Furthermore the loop swap experiment produces a product that is slightly more hydrophobic than pseudomonic acid A, the most abundant product from the mupirocin synthesis pathway, the enhanced hydrophobicity consistent with a product lacking a hydroxyl.

This awesome thesis is dedicated to my loving family, amazing friends and the free software community

Acknowledgements

First of all I would like to thank Dr. Peter J. Winn for accepting me as his PhD student. He has not only been a very supporting and inspiring supervisor, but at times a very good friend as well. Through out three year of my work Peter has challenged me to raise my thinking capability to higher levels. He has appreciated me in my success and motivated me in my set backs and failures. At times when I felt like quitting his counselling has kept me going which has resulted in the completion of this thesis.

I am also grateful to Prof. Christopher M. Thomas for being my co-supervisor. Chris has not only inspired me to do better science, but I have also learned work life balance skills from him. I am thankful to him for being very patient with me while I was in his lab. I am also thankful to all the members of Chris's group who helped me during my stay in the lab and constantly encouraged me in learning new skills. I would like to extend my special thanks to Dr. Anthony Haines for taking so much pain in teaching me all the lab techniques. I must admit that I was a novice having never been worked in a wet lab and Tony taught me skills right from handling a micro pipette to carrying out suicide mutagenesis. I would also like to thank all the members of Centre for Systems Biology for providing me an amazing work environment.

I would also like to gratefully acknowledge Bishop Prof. (Dr.) R. B. Lal, the honourable vice chancellor of Sam Higginbottom Institute of Agriculture, Technology and Sciences, Allahabad, India for granting my study leave to pursue a PhD in the UK.

My education in the UK wouldn't be possible without the generous scholarship from the Darwin Trust of Edinburgh. I am also very grateful to all the funding bodies for providing me travel grants to attend various conferences around the world.

A special thanks to Prof. Alexandre Bonvin for his help with the HADDOCK web server. I am also thankful to everyone who answered my questions in the GROMACS mailing list and a huge thanks to everyone in the free and open source software community for providing me free software for my research.

I am especially thankful and heartfelt in the debt of my friends, Sushant, Shalabh, Utsav, Garima, Sibyl, Atul and Jasreet for their love and being with me in all endeavours of my life. I am also thankful to my colleagues in Birmingham for their cooperation and making my stay in Birmingham a delight. I would especially like to mention the support of Yousra Alsamarraie who not only helped me in the lab, but also for being my best friend in the UK. My stay in the UK would not be better without my friends from the Hope City Church, Birmingham.

Foremost, I would like to thank God for His grace and express the stupendous weight of my gratitude to my loving parents and brother for their blessings and constant encouragement to pursue my ambitions.

Rohit Farmer

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

6-MSA	6-methyl-salicylic acid
ACP	Acyl carrier protein
AIR	Ambiguous interaction restraints
AT	acyltransferase
AUFS	Absorbance units full scale
BLAST	Basic Local Alignment Search Tool
bp	Basepair
CDS	coding sequences
CHS	Chalcone synthases
CLF	Chain length factor
DEBS	6-deoxyerythronolide B synthase
DH	Dehydratase
EDTA	Ethylenediaminetetraacetic acid
ER	Enoyl reductase
ET	Evolutionary trace
FAS	Fatty acid synthase
FMN	Flavin mononucleotide
HCS	3-hydroxy-3-methyl glutryl-CoA synthase cassette
HMG-CoA	3-hydroxy-3-methyl glutryl-CoA
HMM	Hidden Markov model
HPLC	High performance liquid chromatography
IleRS	Isoleucine tRNA synthetase

INSDC	International Nucleotide Sequence Database Collaboration
KR	Ketoreductase
KS	Ketosynthase
L-agar	Luria agar
L-broth	Luria broth
MAT	Malonyl transferase
MD	Molecular dynamics
MDR	Medium chain dehydrogenase reductases
MDR	Multiple drug resistant
Mmp	Mupirocin multi functional proteins
MPT	Malonyl/palmitoyl transferase
MRSA	Methicillin-resistant *S. aureus*
MS	Mass spectrometry
MSA	Multiple sequence alignment
MT	Methyl transferase
Mup	Mupirocin cluster
NADPH	Nicotinamide adenine dinucleotide phosphate
NCBI	National Centre for Biotechnology Information
NMR	Nuclear Magnateic Resonance
NRPS	Non Ribosomal Peptide Synthase
PCP	Peptidyl Carrier Proteins
PDB	Protein Data Bank
PDR	Pan drug resistant
PHI-BLAST	Pattern Hit Initiated - Basic Local Alignment Search Tool
PIER	Protein Interface Recognition for Structural Proteomics
PKS	Polyketide synthases
PPT	Phosphopantetheinyl transferase
PSI-BLAST	Position Specific Iterative - Basic Local Alignment Search Tool

PSSM	Position specific scoring matrix
RMSD	Root Mean Square Deviation
RMSF	Root mean square fluctuation
rvET	Real value evolutionary trace
SASA	solvent accessible surface area
SDR	Short chain dehydrogenases/reductases
SNAC	Sodium 8-((2-hydroxybenzoyl)amino)octanoate
SSM	Secondary stage medium
STS	Stilbene synthases
TAE	Tris Acetate EDTA
TE	Thioesterase
TTC	Triphenyl tetrazolium chloride
XDR	Extensive drug resistant

Chapter 1

Introduction

Polyketide synthases (PKSs) are large mega-Dalton multi-domain enzyme complexes that synthesize a wide range of natural products of medicinal interest. The structure, dynamics and organization that governs these multi-domain proteins during polyketide biosynthesis is still poorly understood. Developing insight into these systems with the aim of routinely re-engineering them for the biosynthesis of novel compounds, such as anti-cancer drugs and antibiotics, is a long standing goal of many of the PKS research community.

Ever since the inception of the first antimicrobial agent penicillin, observed by Alexander Fleming in the late 1930s, the scientific community has endeavoured to discover new types of antibiotics. From the 1960s to the 1980s the world has seen a boom in the antibiotic industry with the introduction of a diversity of drug variants and classes. However, the introduction of new antibiotics from the late 1980s till the present date remained stagnant, with very few antibiotics approved by the regulating agencies and also with very few new classes of antibiotics discovered by the scientific community.

The increasing number of bacterial species acquiring drug resistance and the lack of development in antibiotic discovery has become a threat to humanity. Resistance is now found against almost all classes of antibiotics, even the last line of drugs such as vancomycin (Alanis 2005; Saga and Yamaguchi 2009). Drug resistance ranges from single drug resistance to multiple drug resistance (MDR). Recently bacterial species have been found with extensive drug resistance (XDR) i.e. resistance to all the available antibiotics except colistin, which is a highly

toxic agent and not in regular use. Even worse, the world is facing upcoming issues with pan drug resistant (PDR) organisms, which are resistant to all existing antibiotics including colistin (Powers 2004; Conly and Johnston 2005; Alanis 2005; Saga and Yamaguchi 2009).

Although resistant organisms have reached the community and can be found in outpatients the real threat is still to hospitalized patients. Patients in intensive care units, on steroids or with a compromised immune system are readily susceptible to nosocomial infections. Surgical wound infections commonly caused by methicillin-resistant *Staphylococcus aureus* can only be treated with very few antibiotics, e.g. mupirocin. Thus it has become imperative for the scientific community to discover new antibiotics with better efficacy against resistant species.

Polyketide compounds have shown a variety of medicinal uses, which includes antibiotics. In recent years PKS researchers have shown the capability to re-engineering these pathways in order to produce novel compounds (Kapur *et al.* 2012; Sugimoto *et al.* 2014). In the present study I have worked on the biosynthesis pathway of the antibiotic mupirocin. In spite of being potent against skin infections, mupirocin can only be used as a topical drug as it gets systemically metabolised (Parenti *et al.* 1987). Furthermore, bacteria such as MRSA, which mupirocin was effective against, are developing resistance. Therefore, remodeling the mupirocin biosynthetic pathway to produce variants is a possible way to increase its usability. Similar methods could also be applied to other PKS systems to enhance features such as the binding affinity, and the metabolic profile of the compounds produced. For example, type I polyketide synthases can introduce β-carbon branches into a growing polyketide chain, through tailoring enzyme, allowing introduction of for example a methyl group, halogens or cyclopropane, however it is still not understood what regulates the interaction between the tailoring enzymes and PKSs. Understanding of the underlying mechanism of β-branching can be used to instil the β-branching capability into non-β-branching systems. The dynamic behaviour of the proteins in the pathway, such as acyl carrier proteins (ACPs) that interact with several domains along the pathway, is clearly an important part of these process.

The present thesis mainly focusses on the β-branching mechanism in the mupirocin pathway and on the dynamic behaviour of ACPs. This chapter consists of a general introduction

about polyketides, polyketide synthases, challenges and success in polyketide synthase re-engineering, mupirocin biosynthesis and bioinformatics methods available for PKS research. Chapters 3 and 4 presents the results on the computational prediction of the protein-protein interaction involved in the β-branching mechanism and the experimental validation respectively. Chapter 5 presents the results on the dynamics of ACPs, and how these might affect recognition process in β-branching. Chapter 6 presents the results of two independent projects, first on the ketosynthase specificity towards α-OH substrates in the mupirocin system and the second on molecular dynamics of the loops in ACP-mupA3a + MupH complex assisting the ligand in the MupH active site. Chapter 7 gives the general discussion of the whole thesis.

1.1 Polyketide

Polyketides are natural products produced by bacteria, fungi, plants, molluscs, sponges, insects and dinoflagellates; many of these products have medicinal properties, including antibiotics, antifungal, immunosuppressants, antitumor, antituberculosis and anticholestrol agents (Figure 1.1). As the name suggests polyketides are polymers of simple ketone units synthesized by polyketide synthases (PKSs), which may or may not be further modified by the tailoring enzymes. Due to the diverse nature of modification during the synthesis process, polyketides exist in the form of macrolides, ansamycins, polyenes, polyethers, tetracyclines, acetogenins, and aromatic compounds. Polyketides are produced as secondary metabolites which are non essential for the growth of an organism but may play a crucial role in defence mechanisms, aggression or communication (Bender *et al.* 1999; Staunton and Weissman 2001; Weissman and Leadlay 2005).

Polyketide and other natural product based drugs also have a huge market share, which drives pharmaceutical companies to continuously invest in research and manufacturing. Around 10,000 natural products had been identified by the year 1998 and many of these are being used as drugs. Between 1998-2004, 21 natural product based drugs were launched worldwide and around 19 were approved by the FDA between 2005-2010 (Mishra and Tiwari 2011b; Mishra and Tiwari 2011a). Combined sales of erythromycin, FK506 and lovastatin exceeded $10

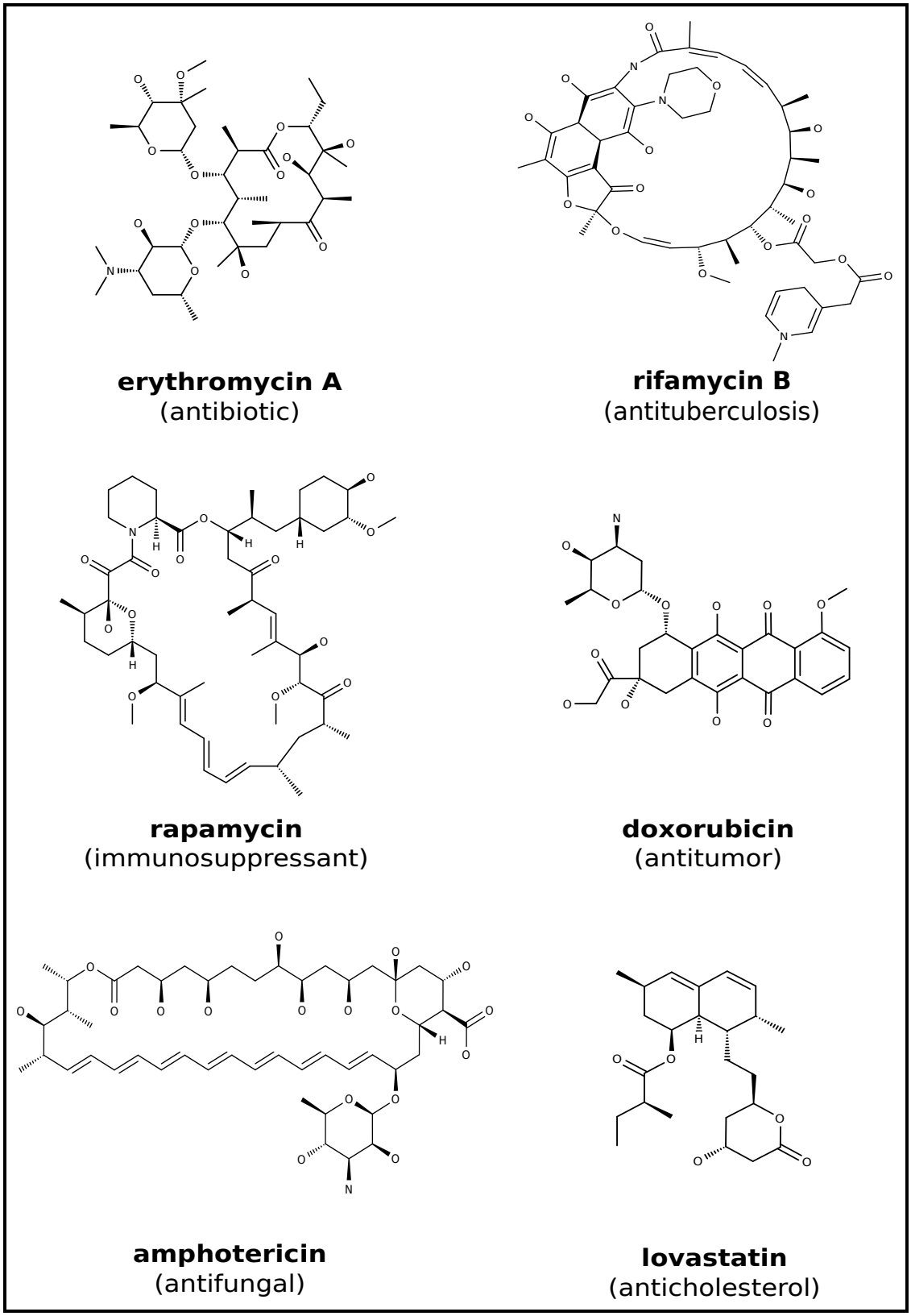

Figure 1.1: *Example polyketide compounds (*`http://www.drugbank.ca`*)*

billion a year according to a news report in Science magazine published on 1 March, 2001 (http://news.sciencemag.org/2001/03/bugs-making-drugs).

Polyketide research dates back to 1907 with the John Collie's work at London University on deducing the structure of orcinol (Collie 1907). However, the main impetus to the field came from Arthur Birch's work on 6-methyl-salicylic acid (6-MSA) produced by *Penicilium patulum* in the 1950s. Arthur Birch explained the mechanism of 6-MSA production based on data from experiments feeding radioactively labelled acetate to the producing strain (Birch *et al.* 1955). The labelled product was shown to have the same pattern as was to be expected from Birch's proposed mechanism. These observations of Collie and Birch later on came to be known as the Collie-Birch polyketide hypothesis, which explains that not only these complex natural products were a result of the condensation of simple acetate units, but also that they can be further transformed through enzymatic reactions to produce aromatic or cyclic compounds.

The detection of these compounds until the 1960s was based on deductive chemistry, where chemists break down the complex molecules into smaller recognizable moieties which were then intuitively assembled on paper (Powers 2004). Later on the advancement in spectrometric techniques such as mass spectrometry (MS) and nuclear magnetic resonance (NMR) completely transformed the research into easier and more accurate measurements. However, the detailed mechanistic view in the biosynthesis of these compounds and the genes/proteins involved was only possible after 1980s owing to the development of genetic manipulation methods. Genetic methods also enabled the scientific community to realize the potential for *de novo* production of these compounds with altered functional groups. In the present era, with the new concept of synthetic biology, researchers are aiming to design synthetic machineries for the production of novel compounds that can be used for beneficial purposes.

1.2 Polyketide synthases

Polyketides are produced by multi enzyme complexes called polyketide synthases (PKSs). Not surprisingly PKS are equally as complex as polyketide compounds. It was only after the 1980s that researchers were able to characterize the genes and proteins involved in the PKS. The very

first genes for a PKS were characterized by David Hopwood's group in 1984, for actinorhodin, an aromatic polyketide compound (Malpartida and Hopwood 1984). Later on Peter Leadlay and Leonard Katz independently identified the gene cluster responsible for the antibiotic erythromycin produced by *Saccharopolyspora erythraea* (Cortes *et al.* 1990; Tuan *et al.* 1990).

1.2.1 FAS and PKS analogous reaction mechanism

PKS share sequence and mechanistic similarity with the very well studied fatty acid synthases (FAS) . FAS synthesizes a relatively small group of saturated compounds as compared to the myriad of polyketides. Fatty acids are produced by a monotonous process of decarboxylative condensation of two ketone units followed by a series of reductive steps till a fully saturated molecule is achieved. The acyl transferase (AT) domain loads the starter (acetyl) and the extender units (malonyl) to the acyl carrier protein (ACP), which covalently tethers the ligand via a thioester bond to the long phosphopantetheine arm. The ACP transfers the starter or the extender units to the ketosynthase (KS) for the decarboxylative condensation and also carries the newly synthesized product to be passed on to the other domains for further processing. Acting successively, keto reductase (KR) reduces the β-keto group to a hydroxyl which is further reduced by a dehydratase (DH) to produce an alkene, this double bond is then reduced to a fully saturated chain by an enoyl reductrase (ER). The process of decarboxylative condensation and successive reductions continue till the fatty acid chain reaches to a required length, where upon it is then released by the thioesterase (TE) domain.

Both PKS and FAS enzyme systems can be classified into type I, II, with an additional type III for PKS (more detail on PKS types in Section 1.2.3). Type I FASs as found in mammals and fungi, are large multi domain complexes where the catalytic domains are covalently linked together in a long polypeptide chain. A single set of seven catalytic domains, as mentioned above, are utilized to produce a single molecule of fatty acid of the required length. Whereas, the type II FASs are free standing discrete mono functional catalytic units, used iteratively. Type II FASs are commonly found in bacteria, chloroplasts and mitochondria.

In comparison to FAS machinery, PKS machinery is more complex and dynamic. PKSs vary

the reductive steps on the β-keto acyl moiety after the condensation has happened, thus producing products with combinations of non modified β-keto groups, hydroxyl and enoyl groups. Several type I modular PKS were also found with modules lacking any condensation activity, for example the module 5 in the mmpA subunit of the mupirocin system. PKSs also have a choice of various starter and extender units (Khosla *et al.* 1999; Staunton and Weissman 2001), in contrast to FASs which always start with acetyl and malonyl as the starter and extender unit respectively. Among many possible examples, of polyketide antibiotic pathways which utilize non conventional starter and extender units, as compared to the FAS, two examples are aureothin synthesis, which incorporates p-nitrobenzoate as the starter unit, and the erythromycin system which utilizes (2S)-methylmalonyl-CoA as the extender units. PKS products are usually modified by various types of tailoring enzymes working *in trans* for purposes such as cyclization, beta branching, epoxidation, pyran ring formation etc. Figure 1.2 shows the generic reaction mechanism carried out by the FAS and PKS.

1.2.2 FAS and PKS models

Over the years many research groups have proposed the likely domain organization and structural models for the FAS and PKS systems. The very first model proposed and widely accepted was the head to tail model in the 1970s which was later on overtaken by a newer head to head model for the FAS system and subsequently for the PKS as well. These simple models based on the domain organisation of the gene cluster and a few crosslinking experiments, helped researchers to understand FAS and PKS assembly until augmented by more accurate atomic resolution structures for full length FAS that were determined through X-ray crystallography. Although, researchers managed to successfully determine the structures for the FAS from various organisms we still lack a complete structure of a PKS. Some recent efforts have indeed produced structures of individual or binary domains from different PKS systems but a complete picture of an entire subunit or a module with all the catalytic domains is still missing.

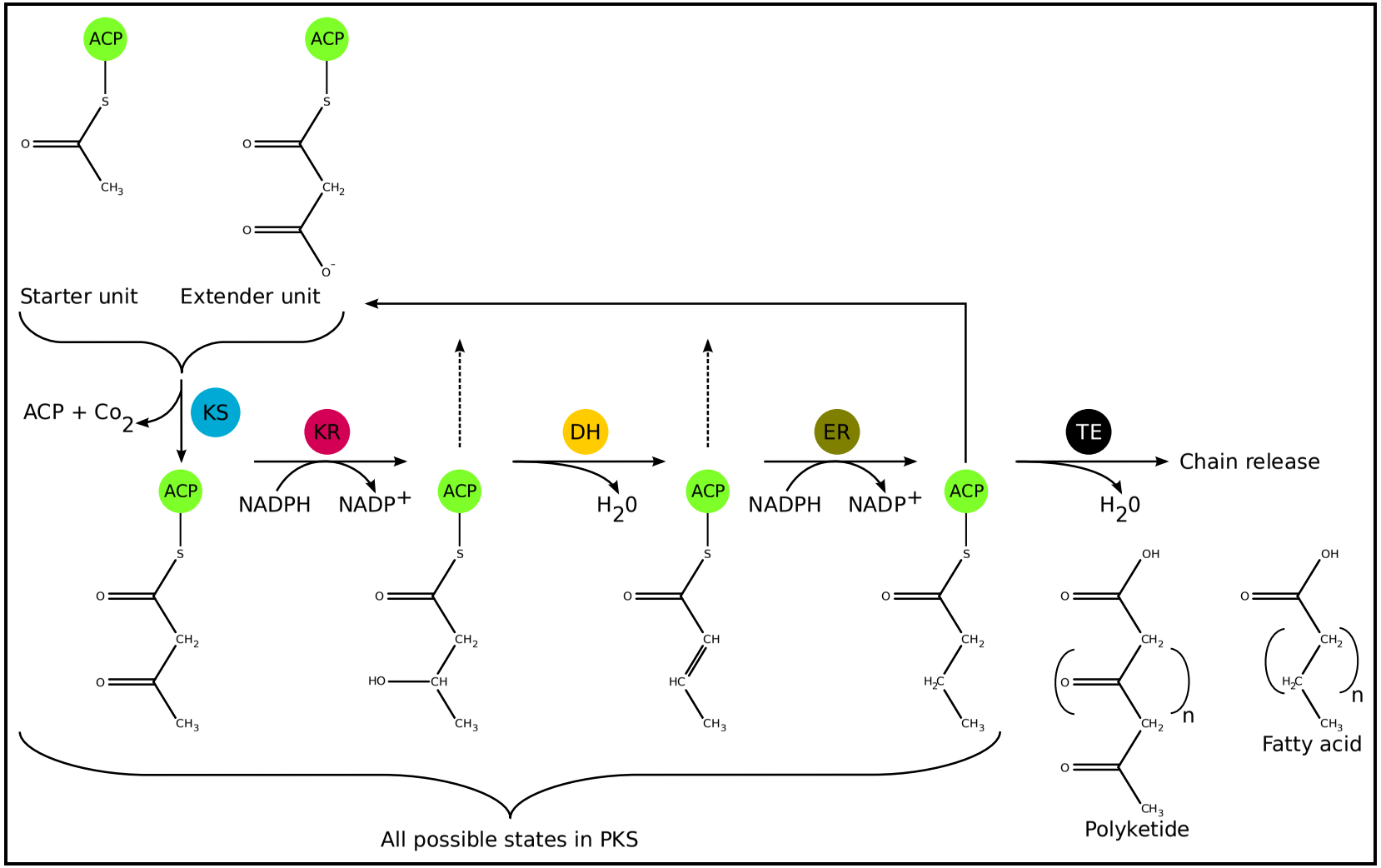

Figure 1.2: *Generic reaction mechanism for FAS and PKS. ACP: Acyl carrier protein; KS: Ketosynthase; KR: Ketoreductase; DH: Dehydrates; ER: Enoyl reductrase; TE: Tioesterase. In FAS the β-keto moiety in the elongated product is completely reduced to an acyl moiety as compared to the PKS in which it may or may not be partially or completely reduced thus producing keto, hyroxy and enoyl products.*

1.2.2.1 Head to tail/head models for FAS

Evidence in the 1970s and 1980s seemed to support the existence of a head to tail model for FAS structure and function (Figure 1.3 A), but this was eventually superseded by a head to head model (Figure 1.3 B) In the 1970s many researchers observed that FAS mono functional units fail to catalyse chain elongation. FASs can be reversibly dissociated into functional units on exposure to low ionic strength buffers at cold temperature; the function restores in high ionic media at room temperature (Kumar *et al.* 1970; Smith and Abraham 1971). The loss of elongation implied that for condensation to happen the active cite cysteine of the ketosynthase needs to be in close proximity of the phosphopantetheine arm of the ACP. Later on it was discovered that the active site cysteine of one subunit can be cross linked to the phosphopantetheine of the ACP on the other subunit by 1,3-dibromopropanone. This cross linking experiment was interpreted as meaning that the condensation required the interaction of KS and ACP from the opposite subunits (Wakil and Stoops 1983). Another experiment on the FAS dimer with blocked thioesterase, showed two hanging fatty acids (Singh *et al.* 1984). Additionally sequence analysis revealed that the ACP is located far away from the KS on the FAS cluster. Thus the above observations seemed to support the fully extended head to tail model where there were two reaction sites which were coordinated by the ACPs from the opposite subunit.

Thus, the head to tail model gained a wide acceptance till it was challenged on the basis of the mutant complementation experiments. The complementation experiments showed that in the animal FAS the KS and ACPs can function together on either subunit (Witkowski *et al.* 1996). Joshi and co-workers from the Smith group created several mutant knockouts in which different domains were inactivated. These mutant subunits were made to re-associate to create a mixed population of hetero and homo dimers. There findings revealed that homodimers formed by the KS, ACP, DH or TE mutant knockouts (at least one point mutation) and the hetero dimers formed by the ACP and TE mutant, ACP and DH mutant and DH and TE mutants were unable to carry out synthesis. However, heterodimers formed by KS and either DH, ACP, or TE mutants were able to carry out the synthesis at a reduced rate. These observations showed that the ACP and DH which are located far on the same subunit were able to interact which suggest that it

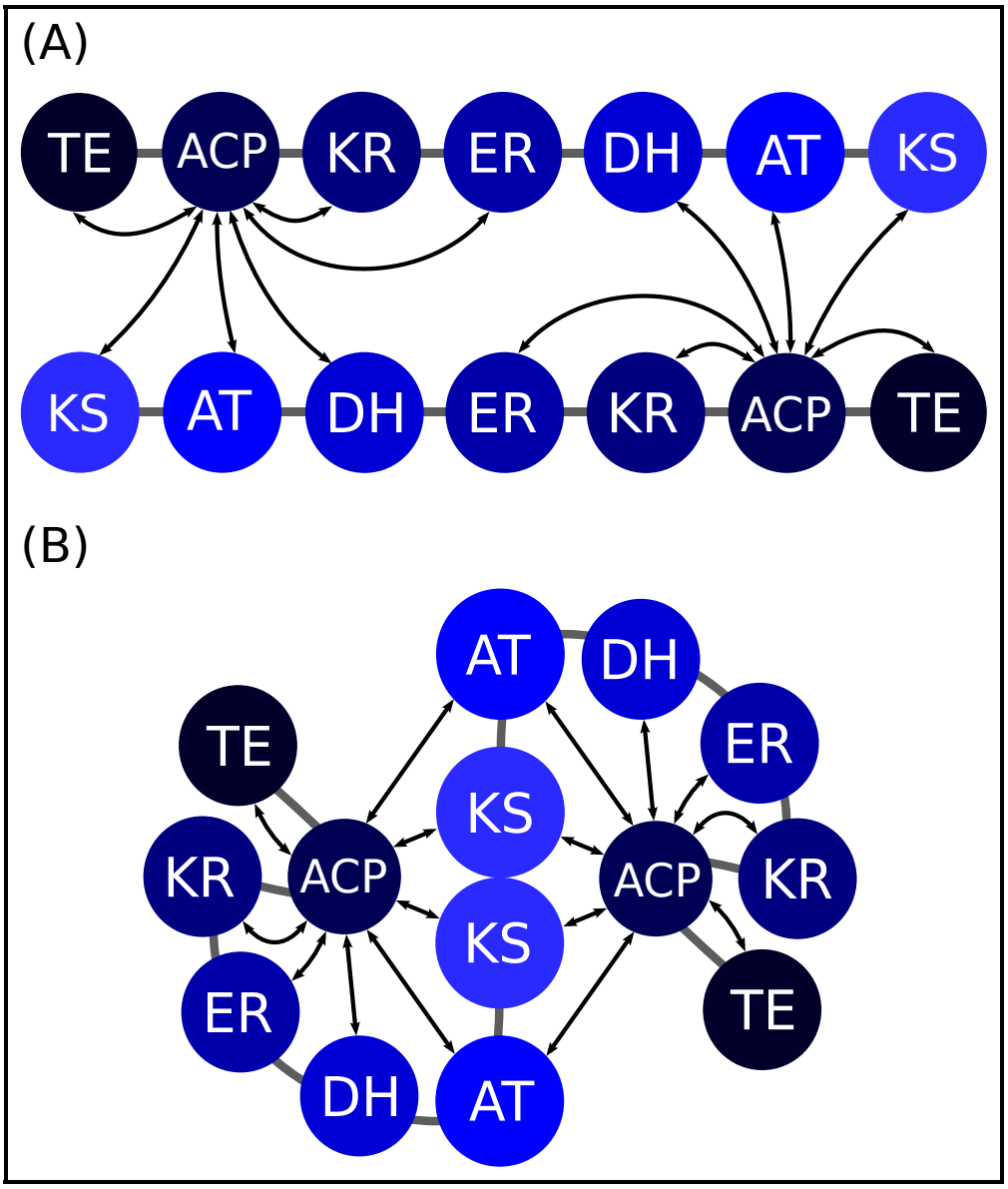

Figure 1.3: *Head to tail/head models for FAS. (A) The fully extended head to tail model (B) The coiled head to head model. Figure adapted from Smith and Tsai (2007).*

was not necessary for the FAS to be in fully extended head to tail conformation. On the other hand a coiled state would explain the ability of the far placed ACP and DH to interact (Joshi *et al.* 1997; Joshi *et al.* 1998).

These observations lead to questioning the interpretation of previously conducted crosslinking experiments, which had showed that 1,3-dibromopropanone can be used to cross link the active cite cysteine of ketosynthase to the phosphopantetheine arm of the ACP. The re-examination of the original cross linking experiment data revealed three bands, which would correspond to the double cross linked inter subunit species, single cross linked inter subunit species and single cross linked intra subunit species. If the interaction of the head to tail model, as in Figure 1.3 (A), held true then there would have been only two bands for the single and double inter cross linked species. The presence of the third band indicates the possibility of the intra subunit KS - ACP interaction.

To put the last nail in the coffin of the head to tail model, Joshi and co workers created an engineered mammalian FAS with only one functional subunit, all the domains from the other subunit were inactivated. This engineered FAS was able to catalyse all the biosynthetic steps in the fatty acid synthesis (Joshi *et al.* 2003). These observations were enough to refute the requirement of the fully extended state of the head to tail model but would rather support the requirement of a scaffold which allows the domains on the two subunits to be accessible to their companion domain (Figure 1.3 (B)).

This scaffold would support the newer head to head model in which the KS forms a dimer at the centre and the flanking domains coil around it, instead of KS being at the polar ends as in the head to tail model as shown in Figure 1.3 (A). This model also agrees with the KS in the FAS type II in which the KSs are known to exist as homo dimers and the active site is formed by the contribution of both the subunits (Moche *et al.* 1999; Olsen *et al.* 2001; Price *et al.* 2003). Experiments carried out by Witkowski and co-worker showed that a truncated N terminal FAS (i.e. with incomplete KS) fails to dimerise. They also showed that the two KS subunits can be cross linked via an engineered cysteine at the N terminal of the KS at a close proximity of 6 Å. This crosslinking failed if one of the subunits lacks an engineered cysteine.

They concluded their observation by performing mass spectrometry on the digested cross linked product (Witkowski *et al.* 2004b).

1.2.2.2 Models for modular PKS

In the 1990s groups from the USA and UK lead by Khosla and Cane, and Staunton and Leadlay respectively, proposed the likely models for modular PKSs. The independent studies carried out by both the groups on the DEBS system using different methods agreed on the presence of two catalytic chambers for the condensation reaction in the modular PKSs. The studies also confirmed the dimeric nature of the PKSs similar to FASs. However, the proposed models by the two groups were strikingly different, details of which are discussed below.

The UK group in a series of experiments carried out limited proteolysis of the DEBS subunits using four proteases. The different sized fragments produced due to proteolysis were identified using N-terminal sequencing. They also determined the oligomeric state of the fragments using gel filtration and analytic ultracentrifugation. All the large fragments carrying the complete module were found to be homodimeric. The loading domain and the AT-ACP domains of the module 1 were found to be monomeric along with all the KR and ER domains in module 1,3,5 and 6 and module 4 respectively. On the basis of the cross linking experiment carried on the FAS (as mentioned in the previous section), the UK group used 1,3-dibromopropanone to cross link the 4'-phosphopantetheine moiety to the cysteine in the KS active site. The cross linking experiment confirmed an interaction between the KS and ACP from the opposite subunits (Aparicio *et al.* 1994; Staunton *et al.* 1996).

These observations laid the basis for two models for the modular PKS, which were proposed to be parallel dimers. In the first model they proposed parallel dimers formed by the interaction of the KS and ACP at the interface and the other reductive domains protruding outwards from the axis, keeping all the domains at the reachable distance of the 4'-phosphopantetheine arm. In the second model (which alter came to be known as **Cambridge model**) the parallel dimers form a helical core utilizing KS, AT and ACP and the reductive domains again protrude away from the axis. Both the models showed the possibility of stacking multiple subunit on top of each other without affecting the functioning of the biosynthesis pathway. However, on the basis

of the ultra-centrifugation results, that the protein dimer does not dissociate even at very low concentrations, the helical model was predicted and favoured to be more stable structure. These observations also proposed the inherent ability of the PKS to accommodate additional PKS which can be stacked with the rest of multi enzymes. Since the reductive domains protrude outside the dimeric core there is also a possibility of adding diversity to the reductive steps (Aparicio *et al.* 1994; Staunton *et al.* 1996), as represented in Figure 3 A and B of Staunton *et al.* (1996).

At the same time Khosla and Cane worked on finding the evidence for the domains participating in polyketide biosynthesis by generating several active site mutant PKSs. The experiments were performed *in vitro* using DEBS 1 + TE, which consist of two modules along with an engineered TE domain. DEBS 1 + TE produces a cyclic triketide by utilizing propionyl-CoA and (2RS)-methylmalonly Co-A as starter and extender units respectively, and NADPH as the hydride donor. These experiments were carried out with the assumption that the modular PKS forms head to tail homodimers similar to the then prevalent idea for mammalian FASs. In one study they produced three different active site DEBS 1 + TE mutants in which either the KS from module 1 or 2 or the ACP from module 2 was inactivated, refered to as $KS1^0$, $KS2^0$ and $ACP2^0$. These three parental mutant homodimers were utilized to produce three mutant heterodimers in which heterodimer a) consist of $KS1^0$ and $KS2^0$, b) consist of $KS2^0$ and $ACP2^0$ and c) consist of $KS1^0$ and $ACP2^0$. The first two heterodimers were found to be active and were able to produce the triketide moiety however, the heterodimer c was unable to produce the triketide product. These observations gave substantial evidence that polyketide biosynthesis is based on the participation of two sets of active sites from opposite subunits and the cognate pair of KS and ACP can only be involved in the chain transfer (Kao *et al.* 1996). These results were in contrast to FAS biosynthesis which allows chain elongation through either the KS-ACP pair from the opposite subunit or from the same.

In another study, following a similar mutant complementation strategy, Khosla and coworkers created an AT null mutant. This AT mutation was carried out in the module 2 and was paired with either the $KS1^0$ or $KS2^0$ mutants, thus generating two pairs of heterodimers. The aim of

the experiment was to test whether the AT2 domain would be able to load the methylmalonyl extender unit to ACP2 from the same or the opposite subunit. If the loading of the extender unit is similar to the KS-ACP interaction from the opposite subunit then only the heterodimer carrying $KS1^0$ and $AT2^0$ should be active and not the heterodimer carrying $KS2^0$ and $AT2^0$. The complementation experiment showed that both the heterodimers were active and the AT domain were able to load the extender unit both inter and intra subunit. Kinetic studies showed no difference in the rate of extender unit loading between the intra or inter subunit transfer (Gokhale *et al.* 1998).

1.2.3 Types of polyketide synthases

As mentioned in the previous sections, PKSs can be classified into three types, I, II and III where type I and II PKS are similar to FAS type I and II. Apart from the three canonical PKS types various PKS systems exist as hybrid with one or more other types. Hybrids also exists between one of the PKS types and a closely related non PKS system called non ribosomal peptide synthases (NRPS). The following sections describe the different PKS types in detail with relevant examples.

1.2.3.1 Type I PKS

Type I PKS are large single polypeptide multi domain protein complexes which can be further sub classified into modular and iterative types. As the name suggests, modular type I PKSs are composed of multiple modules where each module consists of all the necessary domains required for a single round of polyketide chain elongation and associated β-carbon processing. Whereas type I iterative PKSs utilize a single set of covalently bound catalytic domains iteratively until the required length of the polyketide product with correct β-carbon processing is reached (Hertweck 2009). Type I modular PKSs are usually found in bacteria, for example the DEBS system in *Saccharopolyspora erythraea* (Figure 1.4), whereas type I iterative PKSs are found in fungi for example the lovastatin system in *Aspergillus terreus* (Figure 1.5). Type I modular PKSs can also be further classified as *cis* and *trans*-AT systems. *Cis*-AT systems (e.g. DEBS) consist of a covalently fused AT domain within each module, which is responsible for

loading the extender unit specific for the cognate module. Where as in *trans*-AT systems (e.g. mupirocin) the AT domain is not covalently fused within each module but it exists as a separate discrete domain. *Trans*-ATs are responsible for loading the starter unit at the beginning of the pathway as well as the extender units at each module throughout the pathway.

In type I modular PKSs since each module is responsible for a single step of chain extension and subsequent β-carbon processing, it is possible to deduce the product being produced just by studying the sequence in which the modules are arranged in the biosynthetic cluster. This collinearity rule has enabled the development of various computational methods to predict the probable metabolite being produced directly from the DNA or protein sequence. Section 1.5 gives a more detailed account of the computational methods developed till now to predict the polyketide as well as other secondary metabolites from their biosynthetic gene/protein cluster. Although this collinearity rule in type I modular PKSs is effective in predicting metabolite production straight from the sequence, it does not hold true in *trans* AT systems and in systems where a certain module is used iteratively or skipped altogether.

For many years DEBS served as the model system for studying type I modular PKS. DEBS is responsible for the production of 6-deoxyerythronolide B which acts as a precursor in the production of the antibiotic erythromycin. The DEBS system consists of 28 domains spread across 3 polypeptides (DEBS 1, 2 and 3) of 350 kDa each, with each polypeptide containing two modules (Figure 1.4). DEBS 1 contains an extra module which consists of an AT and ACP domain prior to the condensing module 1. This extra module is responsible for priming the KS in module 1 with a propionate unit. DEBS 3 contains a TE domain after module 6, for polyketide release. Each module in the DEBS system is responsible for a single round of Claisen condensation utilizing methyl malonate as the extender unit. Every module in the DEBS system also performs β-keto reduction of the condensed product and only module 4 further reduces the β-carbon with a dehydratase and an enoyl reductase (Figure 1.4) (Khosla *et al.* 2007). Due to the modular nature of the DEBS system and ease of cloning the DEBS genes into an appropriate host like *S. coelicolor* and *E. coli*, many research groups exploited the DEBS machinery to understand and elaborate the re-engineering capability of modular type I

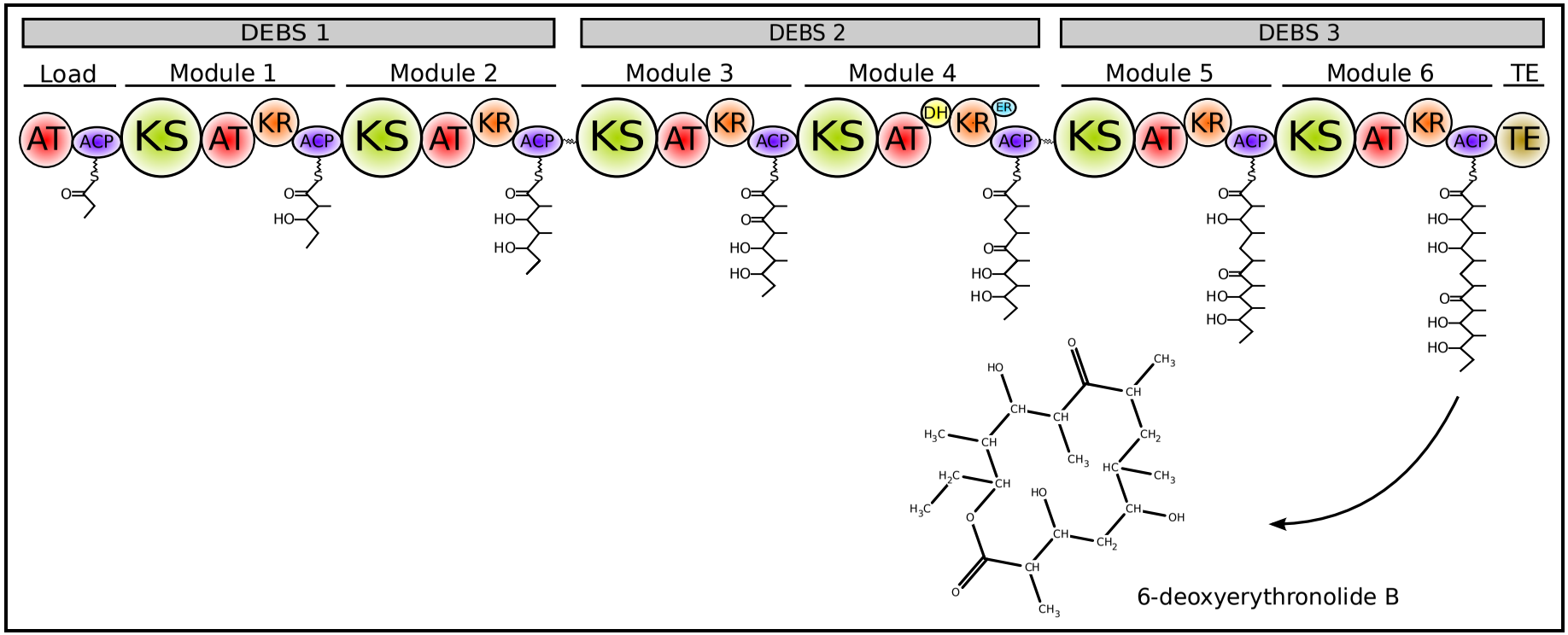

Figure 1.4: *Type I polyketide synthases (modular and cis AT) from 6-deoxyerythronolide B synthase (DEBS) system. The DEBS system consist of 28 domains spread across 3 polypeptides (DEBS 1, 2 and 3) of 350 kDa each, with each polypeptide containing two modules. Abbreviatiosn: KS, ketosynthase; AT, acyltransferase; ACP, acyl carrier protein; KR, ketoreductase; KR 0, inactive ketoreductase; DH, dehydratase; ER, enoylreductase. Figure adapted from (Khosla* et al. *2007)*

PKSs. Section 1.2.6 explains an example of successes and challenges in re-engineering PKSs.

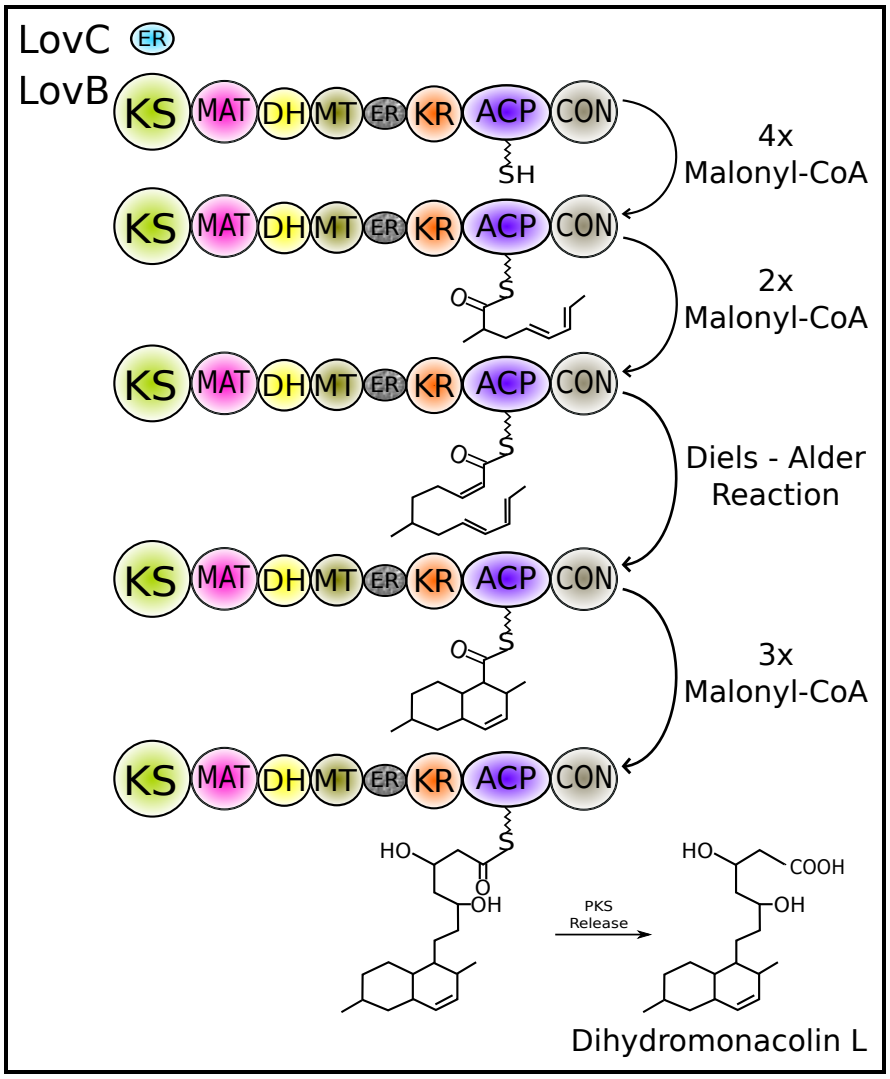

Figure 1.5: *A Type I polyketide synthase (iterative), the lovastatin system. The figure here only shows the PKS LovB out of the two PKSs LovB and LovF utilized in the production of lovastatin. LovB is responsible for the production of dihydromonacolin L moiety which is produced by the iterative condensation of 9 malonyl-CoA extender unit. LovB carries a non functioning ER domain for which the function is substituted by LovC. The other domains in LovB consist of a ketosynthase, a malonyl transferase (MAT), a methyl transferase (MT), a dehydratases (DH), an enoyl reductases (ER), a ketoreductases (KR), an acyl carrier protein (ACP) and a condensation domain (CON).*

Another popular and well studied example is the type I iterative PKS that produces lovastatin, which is produced by the fungi *Aspergillus terreus*. Lovastatin is a cholesterol lowering agent which acts as a precursor for the popular drug simvastatin. The lovastatin biosynthesis pathway consist of two PKSs LovB and LovF. LovB, which is responsible for the nonaketide moiety dihydromonacolin L, consists of a ketosynthase, a malonyl transferase (MAT), a methyl

transferase (MT), a dehydratases (DH), an enoyl reductases (ER), a ketoreductases (KR), an acyl carrier protein (ACP) and a condensation domain (CON) (Figure 1.5). The ER domain in LovB is non functional therefore LovB utilizes a free standing ER LovC. LovB utilizes 9 malonyl-CoA extender units, a methyl donated from a S-adenosyl-L-methionine (SAM) molecule and an NADPH to produce dihydromonacolin L. LovF consists of all the above mentioned domains but it has a functional ER. LovF catalyses the formation of the 2-methylbutyrate moiety by the condensation of two acetyl units. Other enzymes encoded within the same gene cluster includes LovA and LovD. LovA encodes a cytochrome P450 oxygenase which oxidises the dihydormonacolin L produced by the LovB, which is covalently joined to the 2-methylbutyrate moiety by the trans esterase activity of the LovD. Figure 1.5 shows the production of dihydromonacolin L by LovB and LovC in the lovastatin biosynthesis pathway, and includes the proposed stage of a Diels-Alder reaction for ring formation in the growing polyketide (Kennedy *et al.* 1999; Ma and Tang 2007; Campbell and Vederas 2010; Ames *et al.* 2012).

Apart from the minimal PKS domains and one or more required β-keto processing domains, several type I PKS gene clusters contain a number of tailoring enzymes which perform further different functions at different stages of polyketide synthases. These tailoring enzymes may act on the ACP bound polyketide intermediate or on the released product post production. One such tailoring enzyme is 3-hydroxy-3-methyl glutryl-CoA synthase cassette (HCS) which is responsible for the β-branching in many type I PKS systems for example mupirocin, thiomarinol and kalimantacin (Haines *et al.* 2013). The HCS cassette adds a methyl branch at the β-position of the growing polyketide chain by the coordinated action of five enzymes which include an ACP, an HMG-CoA synthase, two enzymes from the crotonase family and a decarboxylase. This β-methylation step is considered to be rate limiting and more complex as compared to the methylation of an α carbon by a methyl transferase. At the onset of β-methylation an ACP delivers a polyketide intermediate to the HMG-CoA domain in the HCS cassette. This delivery requires the ACP and the HMG-CoA from the HCS to interact and recognize each other at the appropriate stage of polyketide synthesis. At the time of starting my thesis, it was not very well understood what enables the HCS cassette to recognize the correct ACP for the β-branching.

One of the projects in the present study was to explore the specificity mechanism of ACP-HCS interaction in the mupriocin pathway, which is described in Chapter 3.

Another recently discovered tailoring enzyme of interest is halogenase. In the curacin system this halogenase domain works in conjunction with the HCS proteins and is responsible for the addition of a chloride on the γ carbon in the growing polyketide chain. In a study published by Busche *et al.* (2012), they found the recognition specificity between this halogenase domain and the ACP. Advances in PKS research are slowly unfolding the details of several different pathways, with a diversity of enzymes functioning in conjunction core PKS function. It is becoming increasingly interesting to understand structure function relationship of these auxiliary enzymes and to exploit them for synthetic biology purposes.

1.2.3.2 Type II PKS

Type II polyketide synthases produce aromatic compounds (polyphenols) in Gram-positive bacteria of the class actinomycetes, found in soil and marine environments. Some of the famous examples of type II polyketides are tetracyclines, which is a class of broad spectrum antibiotics, and doxorubicin, which is an anti cancer agent. In contrast to the type I PKSs, where all the catalytic domains are covalently bonded in a single polypeptide chain, the domains forming the type II PKS are encoded on separate genes. However, the genes encoding the type II domains are usually found to be clustered together. Type II PKSs utilize an acyl starter unit and a malonyl extender unit for Claisen condensation, using a single set of mono functional enzymes iteratively. A minimal set of a type II PKS domains comprises of two KSs, a KS_α and a KS_β domain, and an ACP domain. Other catalytic domains such as ketoreductases and aromatases amongst others perform β-keto processing. It is much more difficult to study type II PKS as compared to type I PKS due to the very short span of stability of the intermediates produced. With type II PKS it is also not possible to predict the metabolites produced straight from the sequence of the biosynthetic domains involved, which is much simpler in type I modular PKS (Hertweck *et al.* 2007).

The KS_α and KS_β domains form heterodimers, where KS_α is responsible for Claisen condensation and KS_β keeps a check on the synthesized chain length, also known as 'chain length

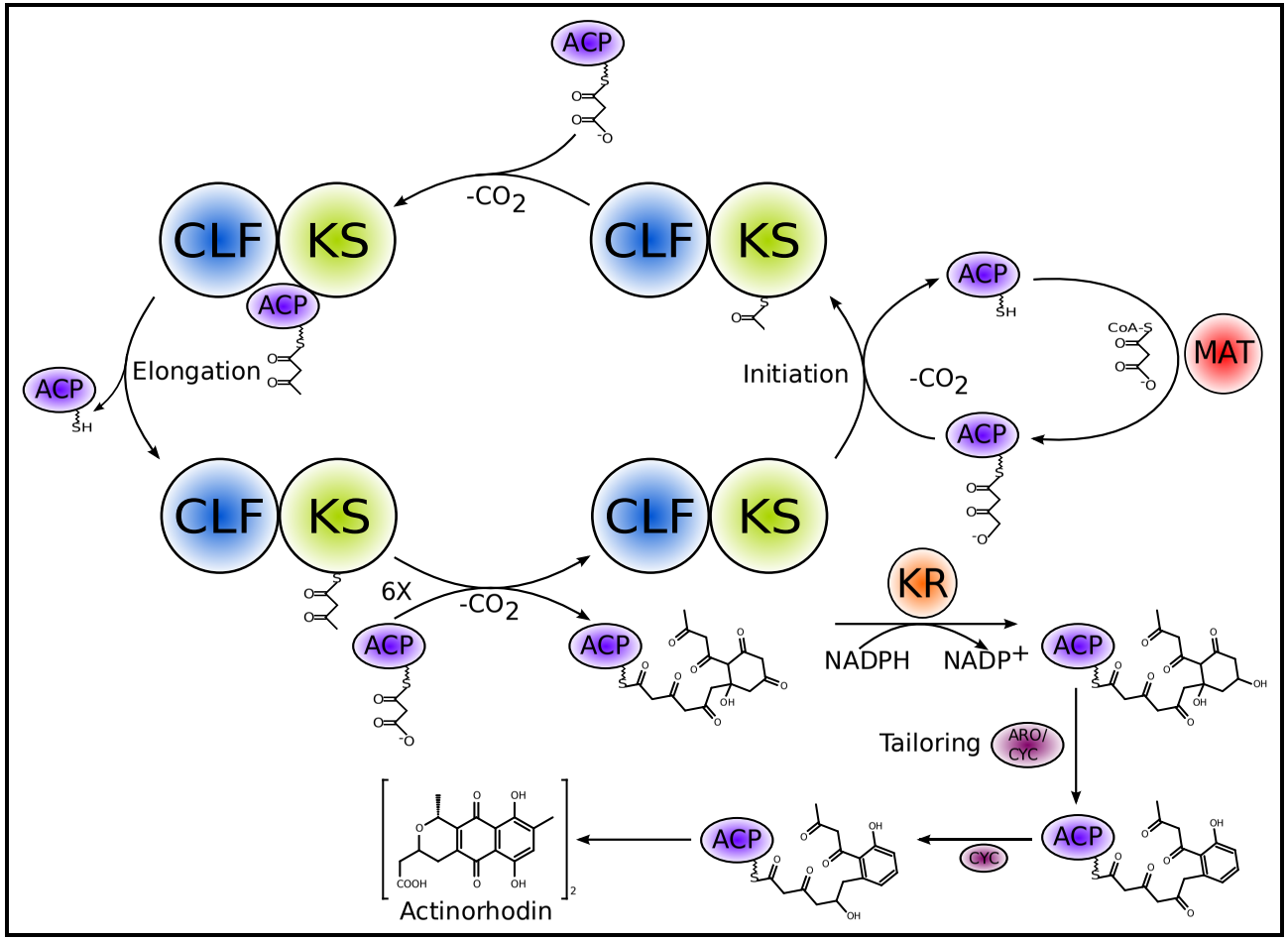

Figure 1.6: *Type II polyketide synthases from actinorhodin biosynthesis pathway. The chain length factor (CLF) and ketosynthase (KS) form a heterodimer. The malonyl-CoA:ACP transacylase (MAT) primes the ACP with the malonate unit. The biosynthesis begins with the decarboxylation of a malonyl unit on to the KS followed by the Claisen condensation with an activated malonyl unit. Overall 7 rounds of Claisen condensation takes place to reach the required length. Upon release the product undergoes post synthesis processing with aromatase and cyclase enzymes to produce actinorhodin molecule. Figure adapted from (Das and Khosla 2009).*

factor'(CLF). KS_{β} lacks a catalytic cysteine and therefore it can not perform Claisen condensation but upon mutating the position where a catalytic cysteine would be expected to a glutamine (KS_Q) it is capable of decarboxylating the malonyl unit into an acetate (Keatinge-Clay *et al.* 2004). This KS_Q functionality is similar to the KS_Qs found in the modular type I PKS where they are utilized to convert a malonyl unit into an acetate starter unit (Bisang *et al.* 1999). Swapping CLF from different type II PKS can result in the production of metabolites of different chain lengths.

It is still not very well understood what catalyses the loading of a malonyl unit to the ACP of a type II PKS. Two alternative hypothesis were proposed both with similar possibility and experimental evidence. One hypothesis favours the self malonylation of the ACP, which was experimentally verified *in vitro* in a purified sample of type II PKS (Arthur *et al.* 2005). The other hypothesis argues a potential involvement of a malonyl-CoA:ACP transacylase (MAT) but most of the type II PKS cluster lacks an MAT therefore it was hypothesized that this MAT is borrowed from a FAS in the cell (Dreier *et al.* 1999; Keatinge-Clay *et al.* 2003). Type II PKS are also found to utilize domains such as KR, DH and ER from the host FAS (Tang *et al.* 2006a).

Figure 1.6 summarizes the biosynthetic pathway of the type II antibiotic actinorhodin produced by *Streptomyces coelicolor* as an example of a type II PKS. Actinorhodin biosynthesis initiates by the priming of the malonate unit onto an ACP by the malonyl-CoA transacylase. The malonyl-CoA unit is decarboxylated and transferred to the cysteine thiol of the KS_{α} which acts as the starter unit. This step is followed by seven iterative Claisen condensation cycles. At the end of the condensation cycles the intermediate product undergoes keto reduction and successive aromatization and cyclization to produce the polycyclic end product actinorhodin. Actinorhodin biosynthesis is carried out purely by a minimal set of type II PKS followed by downstream processing with tailoring enzymes however, other secondary metabolites like R1128 and doxurubicin require an initiation module to produce a diketide starter unit prior to the minimal type II PKS activity. This initiation module consists of a KS homodimer and an ACP, the absence of the CLF suggests the non requirement of chain length control in this module. Further

details, including a cartoon representation of an initiation module, are given elsewhere (Das and Khosla 2009).

1.2.3.3 Type III PKS

Type III PKS commonly found in plants, but with recent discoveries in bacterial systems as well, belong to the superfamily of chalcone synthases (CHS) / stilbene synthases (STS) . CHS is the first step in the plant flavanoid biosynthesis and is responsible for a variety of plant metabolites, with functions including defence, pigmentation and fertility. CHSs utilizes a single homodimeric ketosynthase iteratively, catalyses Claisen condensation of a p-coumaroyl-CoA, as a starter unit, to the three acetate units derived from malonyl-CoA. The same catalytic centre is also further utilized for successive aromatization and cyclization to produce chalcone. Figure 1.7 explains the chalcone biosynthetic pathway proposed by Ferrer *et al.* (1999). PKS type III represents a divergent class of this CHS functionality with the diversity in the choice of starter units, extender units, number of chain extensions and subsequent cyclization. There were differences found in the bacterial and plant type III PKSs, for example bacterial type III PKS can utilize an acyl-ACP starter unit by involving an ACP from an FAS, whereas plants use acyl-CoA as the starter unit. Based on structural analysis of the type III PKS ketosynthase Qiu *et al.* (2001) have suggested the emergence of these enzymes from the fatty acid KAS III. FAS KAS III are responsible for the initiation of type II FAS biosynthesis. The type III PKS utilizes CoA thioesters instead of a phosphopantetheinylated ACP to tether the growing polyketide product (Austin and Noel 2003).

1.2.4 PKS domains

1.2.4.1 Acyl carrier protein (ACP)

Acyl carrier proteins (ACP) belong to a class of carrier proteins which are involved in channelling the substrates via a phosphopantetheine prosthetic group. These carrier proteins are known to be involved in fatty acid synthesis (FAS), polyketide synthesis (PKS) and non-ribosomal peptide synthesis (NRPS) . The carrier proteins involved in FAS and PKSs are called ACPs, but peptidyl carrier proteins (PCP) in NRPSs. In FAS the same ACP domain is utilized iteratively

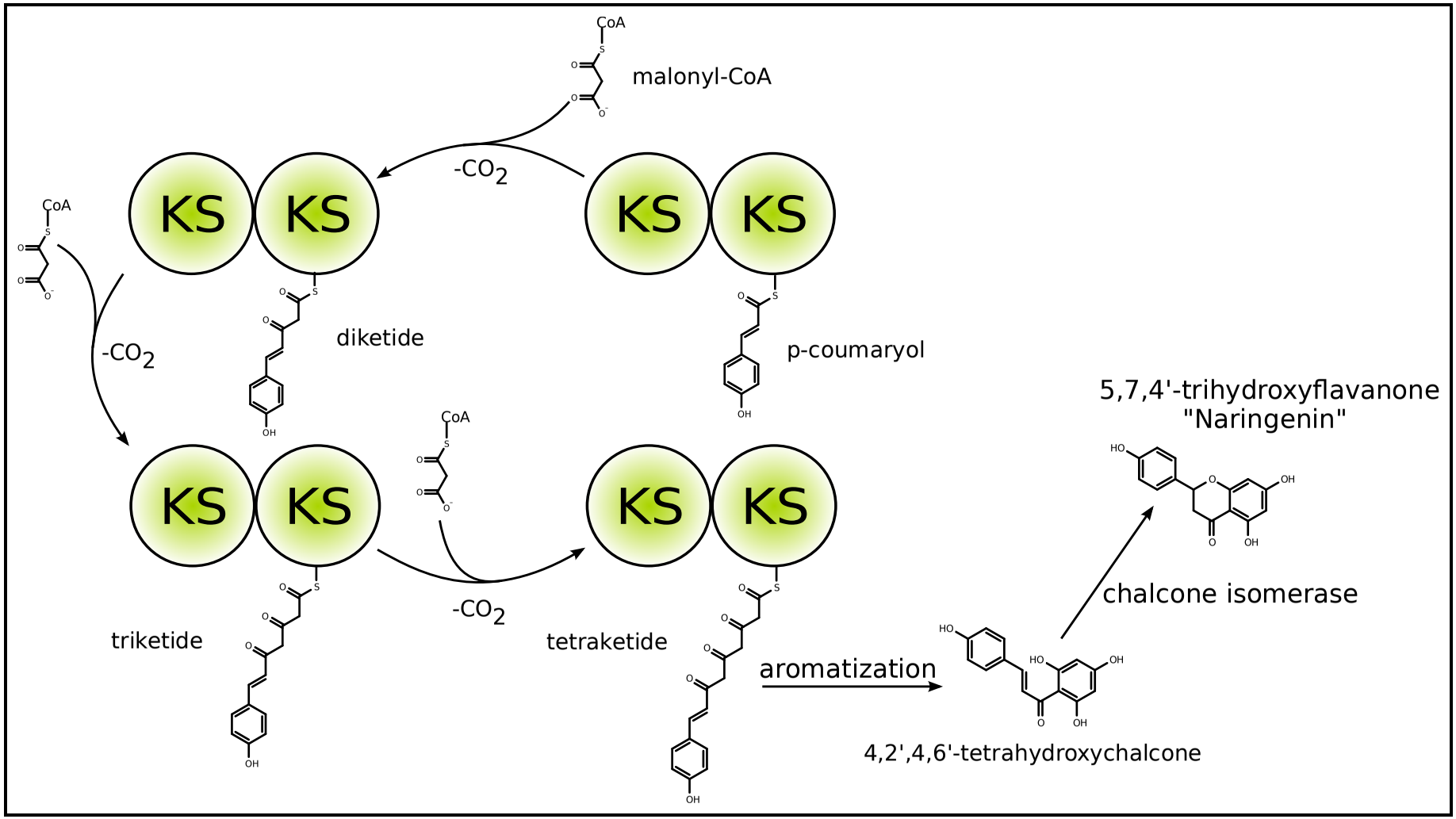

Figure 1.7: *Proposed reaction mechanism of Type III polyketide synthases/Chalcone synthase. The growing metabolite is shown attached to one of the KS of the homodimer. The p-coumarocyl-CoA starter unit is condensed iteratively with three acetate units derived from malonyl-CoA, followed by aromatization and chalcone isomerization to produce Naringenin. Figure adapted from (Ferrer* et al. *1999).*

to pass the substrate to the different domains of the module multiple times whereas in PKSs, especially in type I modular PKS, each module has its cognate ACP to perform the substrate channelling for that module. Although the carrier proteins share a wide range of sequence identity among various organisms, they still share a common structural fold. A sequence identity of >80% is reported between ACPs from *E. coli* and *V. harveyi*, whereas 21%-27% sequence identity between *E. coli* and rat ACPs (Byers and Gong 2007). A common carrier protein fold consists of a 4 helical bundle which ranges from 70 to 100 amino acids long. Among the three major α helices, helix I runs antiparallel to helix II and IV with a small helix III which runs almost perpendicular to the axis of the three major helices. Inspite of their small structure carrier proteins are found to exhibit great degree of plasticity in their backbone movement. FAS ACPs are found to sequester the growing fatty acid chain within the hydrophobic pocket in the centre of ACP (Chan *et al.* 2008). On the other hand NRPS PCPs have been shown to be involved in backbone rearrangement and helical movement during substrate chanelling (Koglin *et al.* 2006). However, no such intrinsic conformational motion has yet been reported in PKS ACPs, which is discussed in more detail in Chapter 5. These carrier proteins are translated as inactive apo proteins that are activated by addition of the 4'-phosphopantetheine moiety of a coenzyme A, through the action of 4'-phosphopanthetheine transferase (Mootz *et al.* 2001). This 4'-phosphopantetheine moiety is attached to a conserved serine at the end of the helix II via a phosphodiester bond (see Figure 1.8). Carrier proteins have also been found to be abundant in various organisms, with 80 reported in *Streptomyces avermitillis*.

In a typical PKS pathway an ACP has to interact with a large number of proteins which include *1*) an AT domain, either *cis* or *trans* *2*) an upstream KS domain *3*) a downstream KS domain *4*) other catalytic domains for example KR, DH, ER, and *5*) several other tailoring enzymes. Until recently it was not known how the relatively small structure of the carrier proteins interact with such a diverse set of interacting partners. There were several unanswered questions such as, how much of the carrier protein's surface interacts with the partner protein? Does each interaction occur at a different face of the protein? What drives the phosphopantetheine arm to the different catalytic domains? Some of these questions have been answered lately however,

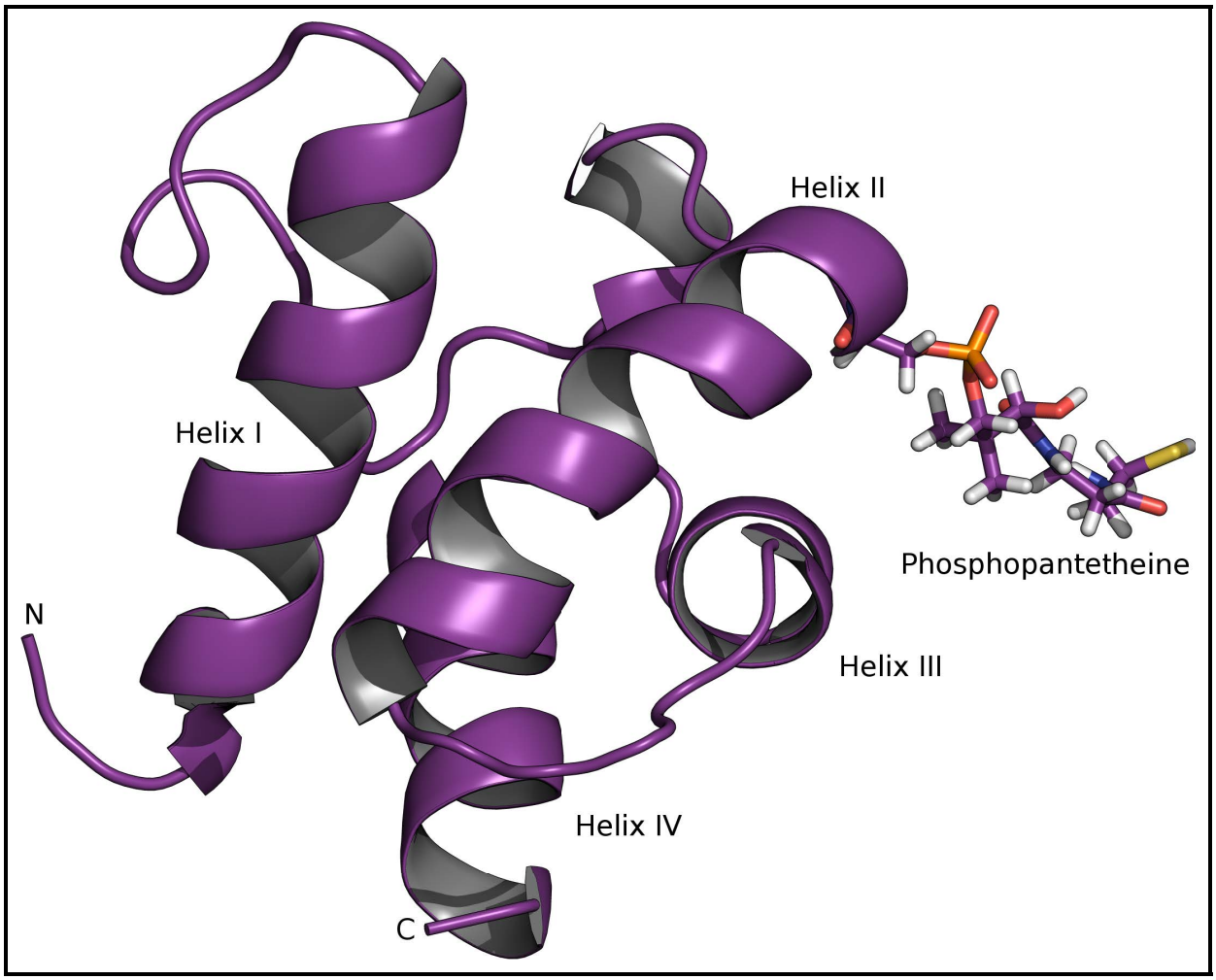

Figure 1.8: *Cartoon representation of an acyl carrier protein (ACP-A3a) from the mupirocin biosynthesis pathway (PDB ID 2L22). A phosphopantetheine arm is modelled attached to the catalytic serine, represented as sticks (Haines* et al. *2013).*

there is still a lot to be discovered.

In a study of FAS ACPs by Zhang *et al.* (2003) helix II was shown to interact with various interacting partners as the "recognition helix". Helix II being conserved and negatively charged interacts with the positively charged channels of various ACP dependent proteins (Zhang *et al.* 2001). However, studies by Khosla's group of PKSs have shown by structural modelling and mutagenesis work that loop I and helix I are the interacting interface for the chain elongation and translocation steps respectively. In another study on curacin biosynthesis helix III on the ACP was found to be interacting with and the halogenase enzyme (Busche *et al.* 2012). In the present work as well, helix III is shown to be one of the anchors that interacts with MupH, as described

in Chapter 3. Thus, it can be said that helix II, which was considered to be the "recognition helix", is not the only surface with which ACPs interact with other proteins. Further exploration may lead to finding some other novel features not yet discovered.

Till now there are only 6 entries in the PDB representing 4 different structures solved for the ACPs from PKS systems, 2 for each *cis* and *trans* AT systems. The first set of structures for the *cis* AT PKS were published in 2007 from module 2 of the DEBS system (PDB ID 2JU1, 2JU2) (Alekseyev *et al.* 2007), followed by the structure of ACP1 from the CurA module of curacin system (PDB ID 2LIU, 2LIQ) in 2012 (Busche *et al.* 2012). Structures from the trans AT systems came out very recently in 2013 for the didomain ACP in the MmpA subunit of the mupriocin system which are involved in β-branching mechanism (PDB ID 2L22) (Haines *et al.* 2013) followed by the solution structure of ACP 5 from the virginiamycin cluster (PDB ID 4CA3) in early 2014 (Davison *et al.* 2014).

1.2.4.2 Acyl transferases (AT)

In PKS and FAS systems acyl transferase domains are responsible for loading the starter and extender units onto the ACPs. In the type I modular PKS systems the AT domains can either be covalently bound in the same polypeptide chain, which are called *cis* AT e.g. as in erythromycin synthesis, or can exist in *trans* as discrete domains e.g as in mupirocin synthesis. In FAS systems, the acetyl starter and malonyl extender moiety are transferred on to the same AT domain, at different times, via acetyl/malonyl transferase from an acetyl-CoA/malonyl-CoA molecule. Thus, an FAS AT domain has dual specificity for the starter and extender units and there is a competition between the two substrates for the AT active site. This phenomenon also holds in many *trans* AT PKS systems, for example mupirocin, where a single set of AT domains is responsible for loading both the starter and the extender units. On the other hand, *cis* AT PKS systems have a separate module for loading the starter unit and the AT domain that is covalently bound to a particular module allows loading its cognate extender unit. Thus in principle AT domains can have specificity towards their own extender units allowing different extender units to be utilized at each elongation step (Khosla *et al.* 1999).

The AT domain catalyses the starter and extender unit's transfer from a CoA on to the phos-

phopantetheine of an ACP. This transfer is achieved by a ping-pong bi-bi reaction mechanism utilizing a conserved SER-HIS diad (e.g. S642 and H745 in DEBS AT5). The malonyl or methyl malonyl moiety was found to interact with the conserved active site ARG (e.g. R667 in DEBS AT5) at the beginning of the reaction (Tang *et al.* 2006b). The catalytic SER resides in a highly conserved motif, GHSXG, and is responsible for the nucleophilic attack on the carbonyl of the acyl moiety offered by the CoA. This nucleophilic attack leads to the formation of a SER-acyl tetrahedral geometry, which is hypothesized to be stabilized by an oxyanion hole formed by the backbone amides (e.g. Q9 and V98 in S. coelicolor MAT structure), and to the release of CoA (Keatinge-Clay *et al.* 2003). HIS in the diad helps to enhance the nucleophilicity of the SER. In the second step of the reaction the thiol of the incoming phosphopantetheine initiates a nucliophilic attack on the SER bound acyl carbonyl forming SER-acyl-ACP tetrahedral intermediate which is released by the protonation of the active site SER by the active site HIS (Figure 1.10) (Tsai and Ames 2009; Dunn *et al.* 2013).

Figure 1.9 shows the cartoon structure of an AT domain from the module 5 of the DEBS system and the proposed reaction mechanism is shown in Figure 1.10. AT domains follow an α/β-hydrolase fold as the core catalytic domain of approx 240 residues attached to a smaller ferredoxin like domain of approx 60 residues. The N-terminus of the AT domains in the *cis*-AT systems are found to contribute to the KS-AT linker (approx 140 residues) which is termed equivalent to the docking domain in the *trans*-AT systems (Tang *et al.* 2006b; Gay *et al.* 2014). However, the docking domain in the *trans*-AT systems does not carry any portion of the AT domain but is formed by the linker sequence in between the KS and the subsequence domain (Gay *et al.* 2014). This linker region in the *cis* system (DEBS) is found to be important for the interaction of the ACP during chain transfer and elongation (Kapur *et al.* 2010; Kapur *et al.* 2012). Although the crystal structure of a docking domain attached to the KS in the trans AT system was recently published the role of the docking domains are still not understood (Gay *et al.* 2014).

Although the ATs from type I and II PKS and FAS share the same catalytic reaction mechanism and protein fold exhibit a wide range of tolerance and specificity towards the starter and

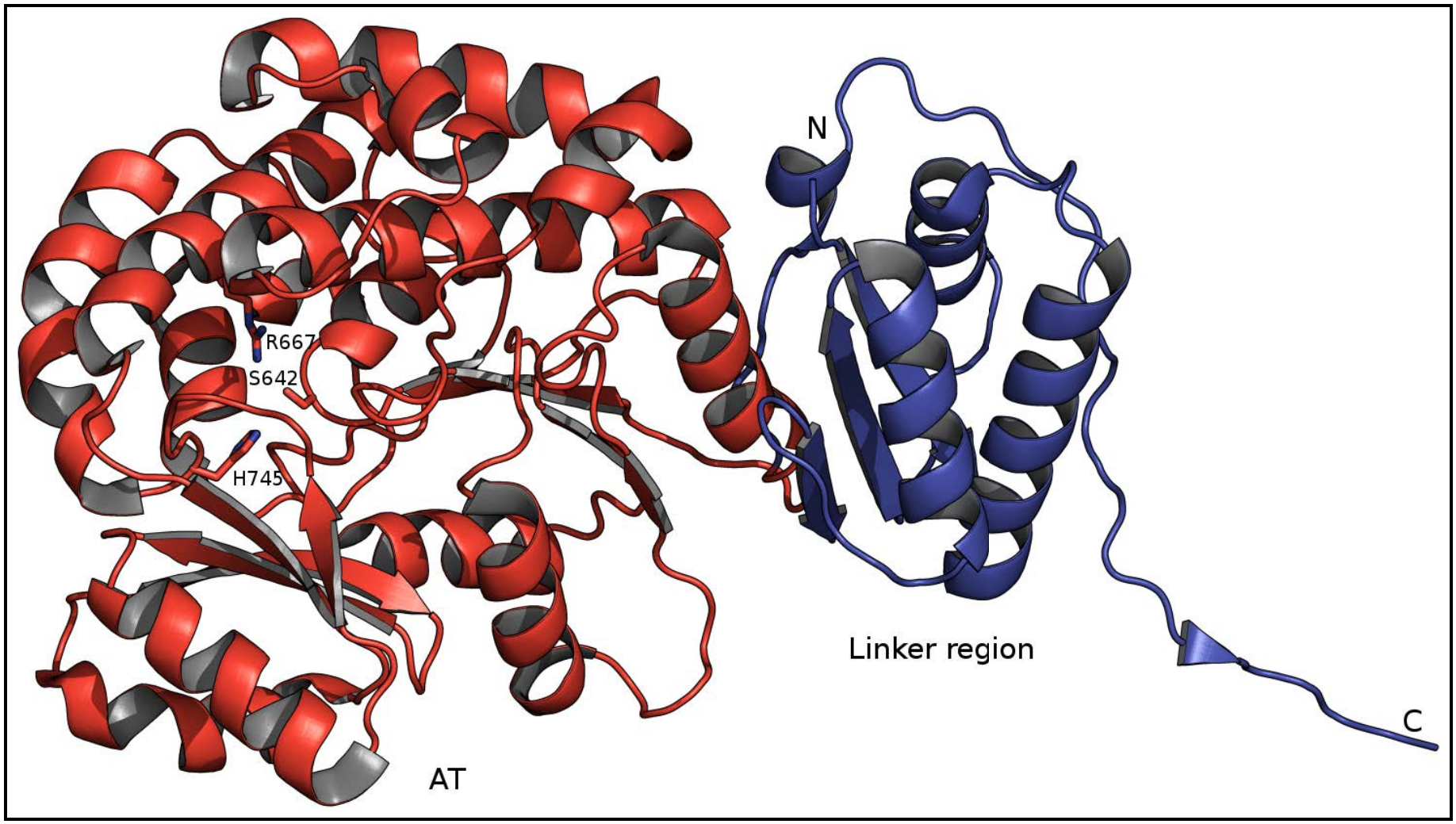

Figure 1.9: *Cartoon representation of an acyl transferase domain (red) from the module 5 of DEBS system along with the linker region (blue) between the KS and the AT domains (PDB ID 2HG4). The catalytic diad S642 and H745 are drawn as sticks along with the conserved active site R667 (Tang* et al. *2006b).*

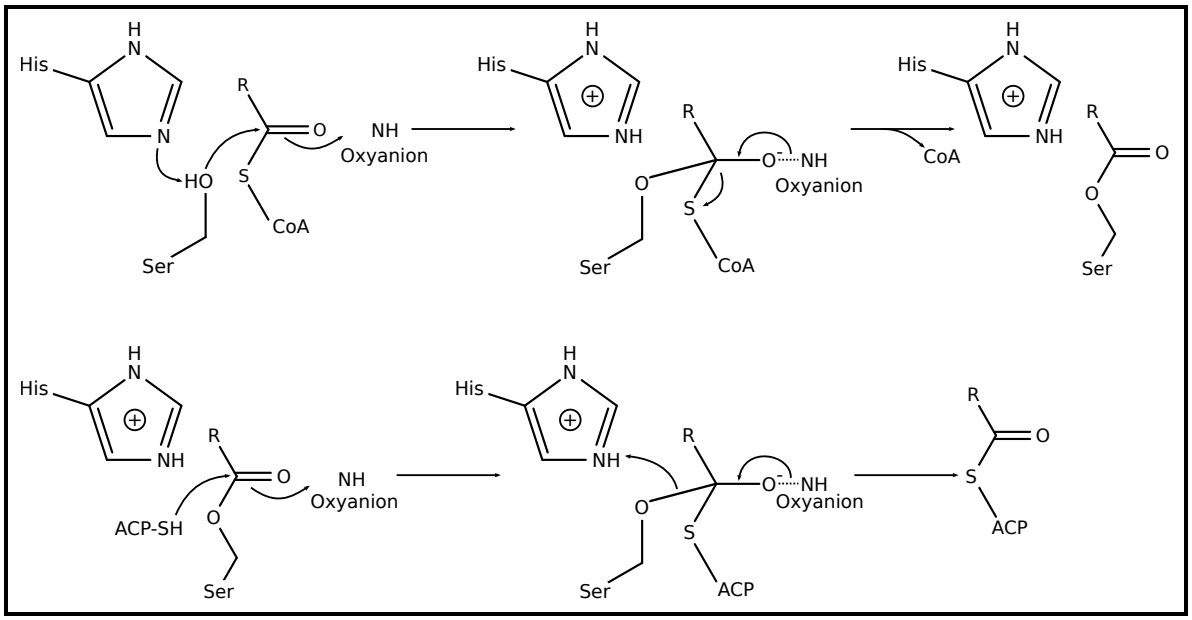

Figure 1.10: *Proposed reaction mechanism of AT domain. Figure adapted from (Smith and Tsai 2007).*

extender units they transfer on the ACP. *In silico* and mutagenesis experiments have identified conserved motifs within the vicinity of active site serine which determines the specificity towards the different substrates. The loading domains in the *cis*-AT systems were found to have more tolerance for a greater variety of starter units such as acetyl, propionyl, isopropionyl than the AT domains responsible for loading the extender units such as malonyl and methyl manlonyl (Lau *et al.* 1999; Liou *et al.* 2003).

The X in the GHSXG motif, which is usually a branched amino acid such as valine or isoleucine in the malonyl specific ATs, is either a glutamine or methionine in the methyl malonyl specific ATs (Haydock *et al.* 1995). On comparing AT structures from DEBS AT5 with a mammalian FAS AT, DEBS AT5 has a glutamine at the X position which makes it specific for methyl malonyl extender unit, whereas mammalian FAS AT has a valine or leucine at the X position which makes it specific for malonyl extender unit (Tang *et al.* 2006b).

In another study, Yadav *et al.* (2003a) studied 187 AT sequences from 19 type I modular PKS clusters and predicted motif QQ**GHS[QMI]G**RSHT[NS]V was responsible to confer specificity towards methyl malonyl substrate and QQ**GHS[LVIFAM]G**R[FP]H[ANTGEDS][NHQ]V motif was associated with malonyl specificity. They have also identified a position R117 in *E. coli* malonyl-CoA:acyl carrier protein transacylase crystal structure (PDB ID 1MLA) which

was conserved in all the malonyl and methyl malonyl specific ATs but changes to a non polar amino acid for the ATs specific for monocarboxylic substrates such as propionate. These motifs are embedded in their PKS domain detection program SEARCHPKS (Yadav *et al.* 2003b). Another motif is the YASH motif which contains the catalytic histidine 100 residues downstream of the catalytic serine. YASH motif was found to be specific in the methyl malonyl specific ATs whereas HAFH at the same position was specific for the malonyl substrates (Haydock *et al.* 1995).

1.2.4.3 Ketosynthases (KS)

Ketosynthases are responsible for catalysing Claisen condensation in PKS, FAS and NRPS systems. KSs belong to the thiolase family of proteins and in type I FAS and type I and III PKS they form homodimers (Austin and Noel 2003; Tang *et al.* 2006b). KSs in type II PKS also forms the dimer but only one subunit performs the Claisen condensation and the second subunit which is also known as chain length factor is a non functional KS considered to be responsible for keeping a check on the growing polyketide chain length (Tang *et al.* 2003; Szu *et al.* 2011). The homodimeric state in type I PKS and FAS is considered to be the primary factor for the dimerization of the two subunits. Although ER and DH domains are also found to form dimers, experiments have shown that upon KS deletion the FAS subunits fails to dimerize (Smith and Tsai 2007). The thiolase fold of the two dimer forming KSs follow the same alternating $\alpha/\beta/\alpha/\beta/\alpha$ architecture. The active site is formed at the dimer interface by the contribution of the residues from both the subunits. The active site can be viewed as two segments, one at the most buried segment is next to the dimer interface and bind the acyl intermediate, and an outer tunnel starts from the enzyme surface and binds the phosphopantetheine arm. Sequence analysis has shown diversity and lesser sequence similarity in the residues lining the acyl intermediate binding pocket as compared to the phosphopantetheine binding tunnel (Olsen *et al.* 2001). This is due to the ability of the KS across different systems to accept a variety of substrates which are however, attached to the same phosphopantetheine arm.

There are two main classes of Claisen condensations for carbon-carbon bond formation; the decarboxylating and the non-decarboxylating. The structures of the KSs used for both the

classes are low in sequence similarity but still have the same three dimensional thiolase fold (Heath and Rock 2002). The decarboxylating condensing enzymes can be further divided into two sub classes based on the primary sequence analysis 1) Initiation ketosynthase such as type I FAS, FabH and 2) Elongation ketosynthases such as type I FAS, FabB and FabF (Davies *et al.* 2000). The initiation enzymes utilize CoA substrate as the primer whereas the elongation enzymes utilize ACP thioesters however, the reaction mechanism followed by both the enzymes is same. The initiation ketosynthases as the name suggests are found at the beginning of the type I FAS and PKS usually in the loading module for example the niddamycin pathway (Kakavas *et al.* 1997).

After years of disagreement and research and with the availability of the high resolution crystal structures of KSs from both FAS and PKS systems researchers have come to the consensus that the Claisen condensation is a three step process involving a CYS-HIS-HIS catalytic triad (Figure 1.11). In the *E. coli* FAS (FabB/FabF) KS (PDB ID 1DD8) C163, H298 and, H333 forms the catalytic triad (Figure 1.12). The Claisen condensation initiates with the acyl chain transfer from an ACP on to the catalytic cysteine forming an ACP-acyl-KS thioester tetrahedral intermediate. In FAS (FabB/FabF) this tetrahedral geometry is stabilized by the backbone amides contributed by the catalytic cysteine and a glycine (G391). The pKa of a free cysteine (8.0 to 8.8) is not sufficient for the deprotonation and the nucleophilic attack on the acyl carbonyl. A pKa of about 7.0 or lower is required for the cysteine to act as a nucleophile under physiological conditions. Qiu *et al.* (1999) suggested H244 may be involved in assisting the deprotonation of the cysteine thiol. However the distance between the N ϵ ans S γ of the H244 and catalytic cysteine respectively were too far for this to happen. Furthermore replacement of H244 with an alanine resulted in a protein that was unable to catalyse the complete condensation reaction, but was 6 times faster in transacylation than the wild type. However, mutagenesis studies on H244 do show that it assists in the deprotonation of the thiol at low pH (Davies *et al.* 2000). The probable explanation for the deprotonation of the catalytic cysteine thiol under physiological conditions was given on the basis that the cysteine is present at the N-terminus of the alpha helix, the positive end of the dipole moment of the helix could stabilize the negative

charge, allowing transfer to the acyl chain to the cysteine (Davies *et al.* 2000).

The second step involves the decarboxylation of the malonyl extender unit attached to the ACP. The two histidines are considered to be involved in the decarboxylation of the malonyl unit. A water molecule in the active site is thought to be activated by one of the histidines which in turn attacks the C3 carbonyl of the malonyl resulting in the formation of an enol intermediate, stabilized by the histidines. This decarboxylation step also happens in the absence of the acyl-cysteine thio ester and that is why the above mentioned initiation keosynthases that lacks catalytic cysteine are able to decarboxylate a malonyl or methylmalonyl units into an acetyl or propionyl units respectively. Such initiation ketosynthases carry a glutamine in place of a cysteine (also known as KS^{Q}). In the final step the enol intermediate forms a carbanion which in turn attacks the acyl-cysteine thio ester resulting in a tetrahedral intermediate which is again stabilized by the oxyanion hole followed by the release of the acyl-ACP (Heath and Rock 2002).

In modular *cis*-AT PKS, Ketosynthases have shown to be tolerant in terms of substrate specificity and can readily accept non native substrates. Studying the structure of the type I and type II KS, type I KSs have a loop at the interface in the homodimers which is replaced by a helix in the type II KSs. This loop region is hypothesized to be responsible for providing additional space to accommodate non native substrate which otherwise is restricted by the helix in the type II KSs (Pan *et al.* 2002). KSs were able to catalyse chain elongation with inactivated KR5 or ER4 domain in the DEBS system. KS2 in the DEBS system was also found to be tolerant towards several substrates (Khosla *et al.* 1999). Using a KS1 knockout strain Khosla *et al.* (1999) showed that (2S,3R)-diketide and analogous substrates attached to a NAC molecule tolerated and processed by the KS2 domain. Not only different variants of diketide moeities they also tested an anhydro-triketide moiety which was able to be processed by the KS2, these substrate variants produced analogues of 6-DEB molecule. As well as KS2 being tolerant towards (2S,3R)-diketide moiety KS3, KS5 and KS6 were also able to accept and produce triketide products (Khosla *et al.* 1999).

Recent phylogenetic studies have shown that the *cis* and *trans* PKS evolved independently

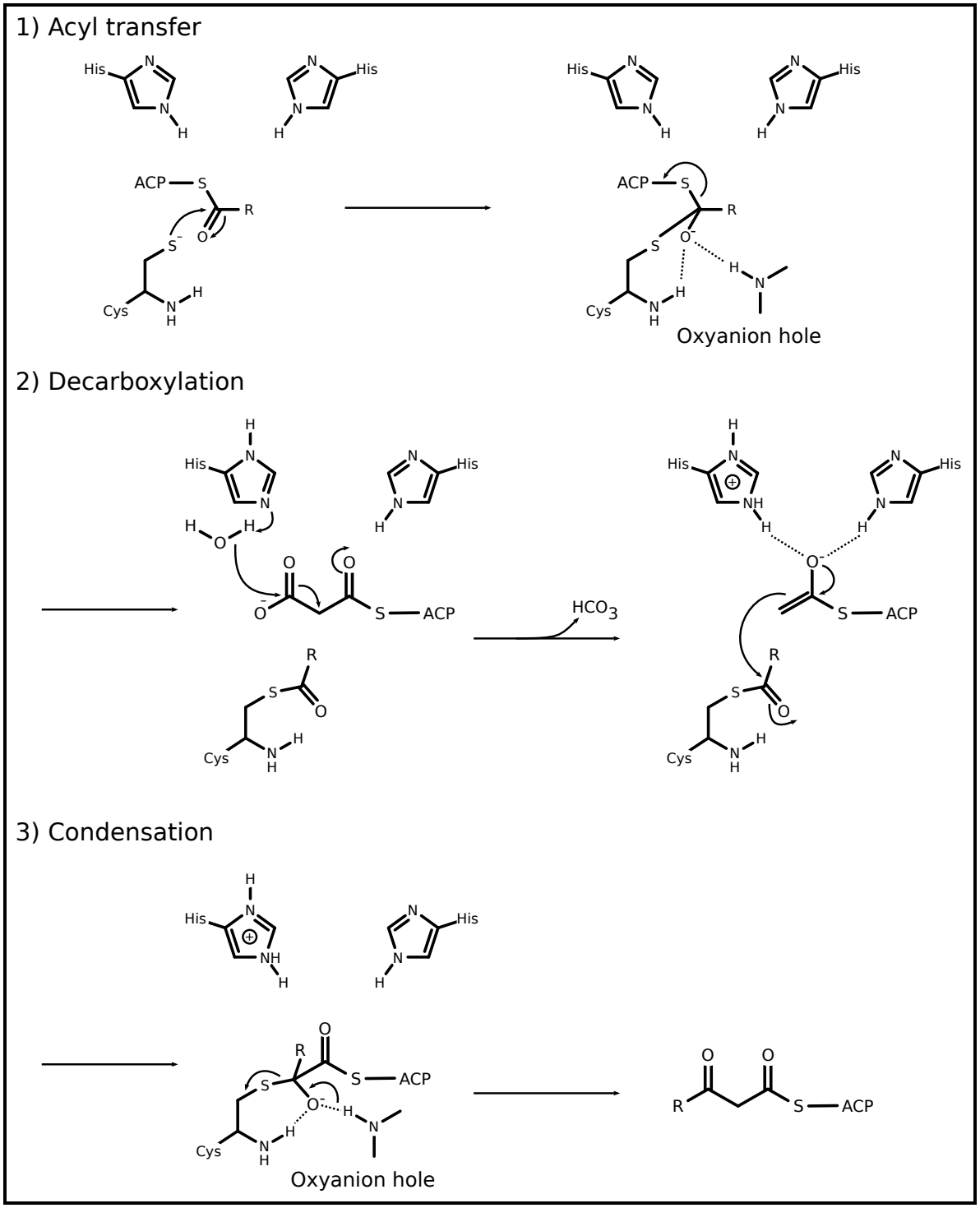

Figure 1.11: *Proposed reaction mechanism for Claisen condensation. Figure adapted from (Smith and Tsai 2007).*

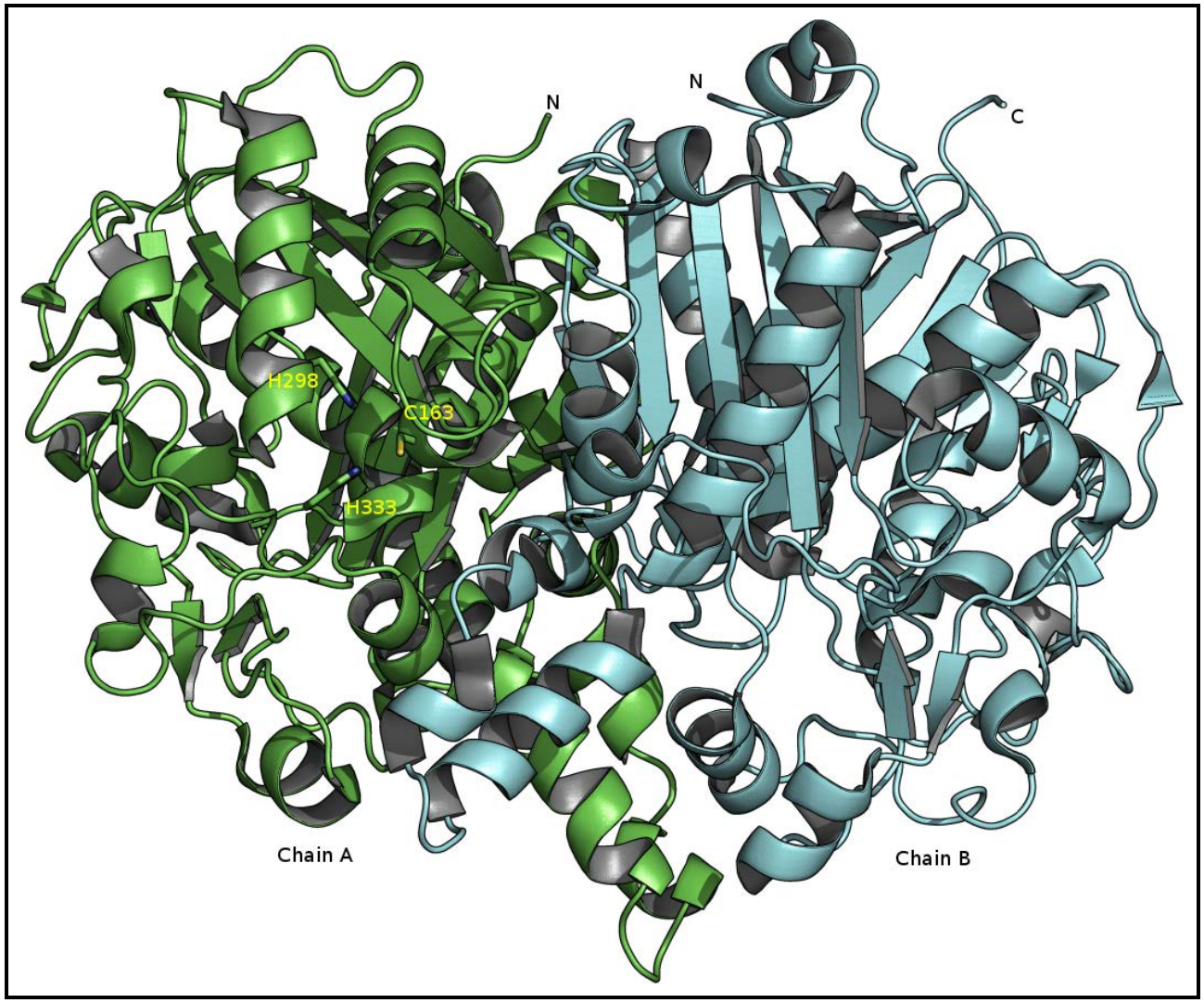

Figure 1.12: *Cartoon representation of a ketosynthase homo dimer from FAS in E. coli (PDB ID 1DD8). The catalytic triad C163, H298 and H333 are drawn as sticks (Olsen* et al. *1999).*

through different routes. The KS from the *trans*-AT systems were likely to be incorporated via horizontal gene transfer as opposed to gene duplication events in the *cis*-AT systems. This difference may explain how *trans* AT systems came to be more specific towards their substrates as compared to their *cis*-AT counterparts (Nguyen *et al.* 2008). These hypothesis were also experimentally supported by the two recent collaborative papers by Piel and Oldham group. In the first paper by Jenner *et al.* (2013) utilizing a newly developed mass spectrometry (MS) based method to study ketosynthases specificity analysed substrate specificity of KS5 from bacillaene and KS1, KS2 and KS3^0 from psymberin biosynthesis pathways. KS5 in the BaeL subunit resides right before the key β-branching step which naturally makes it tolerant for non branched substrates. Similarly KS1 and KS2 from PsyA processes non branched substrate and KS3^0 from PsyD is a non condensing keto synthase but is capable of acyl transfer. Jenner *et al.* (2013) incubated these ketosynthases with different SNAC analogues containing branched and unbranched substrates and tested their ability to produce product with MS. They found that the KS5 from BaeL and KS3^0 from PsyD do not tolerate branched substrate however were able to successfully process unbranched substrates. On the other hand KS1 and KS2 from PsyA were able to process branched substrate with lowered specificity. Upon sequence analysis, homology modelling and substrate docking of KS5 of BaeL, they found that the residue right before the catalytic cysteine (X-cys) seems to interact with the β-carbon. This position X is usually a bulky residue like methionine in the non branch accepting ketosynthases and a smaller residue like glycine or alanine in the branched chain accepting ketosynthases. KS5 from BaeL and KS3^0 from PsyD have methionine on that position whereas KS1 and KS2 form PsyA have an alanine, which explains the more tolerant behaviour of the branched substrate by KS1 and KS2. To test this prediction from the computational analysis Jenner *et al.* (2013) created a KS5 M237A mutant and incubated it with branched and unbranched SNAC analogues. The unbranched analogues were successfully transferred onto the catalytic cysteine of the KS5 mutant as the more space created inside the active site would have not affected the binding of the unbranched substrate. However, the KS3 mutant also showed increased levels of branched substrate specificity which suggests the necessity of the space created by replacing a bulky residue with a small residue

immediately preceding the catalytic cysteine. These experiments laid down the new rules for predicting the specificity of KS domains towards branched and unbranched substrates. Although this rule is only restricted to the β- position in the substrates authors do not deny the possibility of discrimination based on the interactions at other parts of the substrate.

In the second paper, Kohlhaas *et al.* (2013) tested the specificity of the bacillaene system KS1 domain from BaeJ for amino acid derived intermediates using the same protocols as mentioned above. KS1 in BaeJ accepts a glycine derived substrate from an upstream NPRS module, to test the tolerance of KS1 for different amino acid based substrates Kohlhaas *et al.* (2013) created several different full length acyl-SNAC analogues utilizing glycine, alanine and valine to generate 2-amido variants. Upon incubating KS1 with the three 2-amido variants, the glycine derived substrate was able to readily transfer to the KS1 whereas the alanine based substrate did so with lower efficiency. The substrate based on valine failed to acylate KS1. In order to accommodate the bulkier valine derived substrate Kohlhaas *et al.* (2013) identified M268 and L450 as the likely residues that offer steric hindrance in the entry of the substrate. They created two mutants M268A and L450A and tested all the three amino acid derived substrates. Glycine and alanine variants were still able to acylate however, the valine variant couldn't. These differences in the acylation efficiency suggested the intrinsic incompatibility of the groups at the α position. They also created SNAC analogues to test any long range interaction beyond the β-position in the intermediates and the affect of not incorporating amide from an α amino acid. There were no acylation observed in the substrate created to test the long range interaction as well as for the amide from the non α amino acid. Two other acyl-SNAC analogues featuring 2-amino and 4-keto groups were also incubated to test for acylation efficiency. 2-amino SNAC was able to acylate efficiently however, no observable acylation occurred with a 4-keto SNAC substrate. These observations suggested the need for an NH at the 2nd position, which would act as an efficient hydrogen bond donor, and that there is no requirement of the 4th carbonyl for the successful acylation.

Kohlhaas *et al.* (2013) also tested KS1 for the effect of different residue type N-terminal to the catalytic cysteine on KS1's ability to accept different substrates. Through sequence analy-

sis they found that most of the amino acid accepting KSs posses an asparagine residue at the position before the catalytic cysteine along with a few occurrence of alanine. On comparing with other ketosynthases in either *cis* or *trans*-AT systems they could never find an asparagine at that position. The dominating presence of an asparagine residue suggested its likely involvement in the acylation of the 2-amidoacetly substrate to KS1. To test this hypothesis they created an N206A mutant and incubated with 2-amidoacetly-SNAC. The mutant strain showed much lowered acylation efficiency.

1.2.4.4 Ketoreductases (KR)

Ketoreductases are enzymes responsible for reducing the β-keto group to hydroxyl in FAS and PKS systems with the involvement of NADPH as the reducing agent. Unlike KSs which need to exist as a homodimers to function, KRs exist as functional monomers in the PKS dimer. KRs are thought to be responsible for the formation of most of the chiral centres in Polyketide products.

The stereo specificity of the products of KR domains was verified through domain swapping experiments with the KR2 from the DEBS system and the KR2 from the rapamycin system. The KR2 in the DEBS system produces an L-β-hydroxyl group whereas the KR2 from the rapamycin system is responsible for a D-β-hydroxyl group. In a model system using DEBS1 (Figure 1.4) and the DEBS thioesterase, domain swaps from rapamycin KR2 to DEBS KR2 produced a triketide with a D-β-hydroxyl group, confirming the stereo specificity of the KS product (Cortes *et al.* 1995). It should also be noted that the mammalian FAS are known to produce D-β-hydroxyl groups whereas DEBS KRs produce an L-β-hydroxyl group, this observation also suggests that the KRs are not only responsible for the β-keto reduction but also for assigning the correct stereo chemistry to the resultant hydroxyl group (Kao *et al.* 1998). KRs from the DEBS systems were also found to be tolerant of different substrates. In a domain swap experiment, a system with KR5 from the DEBS system replacing KR2 successfully produced the DEBS1 triketide product, suggesting a wide tolerance for KR substrate specificity (Khosla *et al.* 1999).

Ketoreductases can be classified into three types, type A which is responsible for L-β-hydroxyl group, type B for D-β-hydroxyl group and type C which are incapable of β-keto

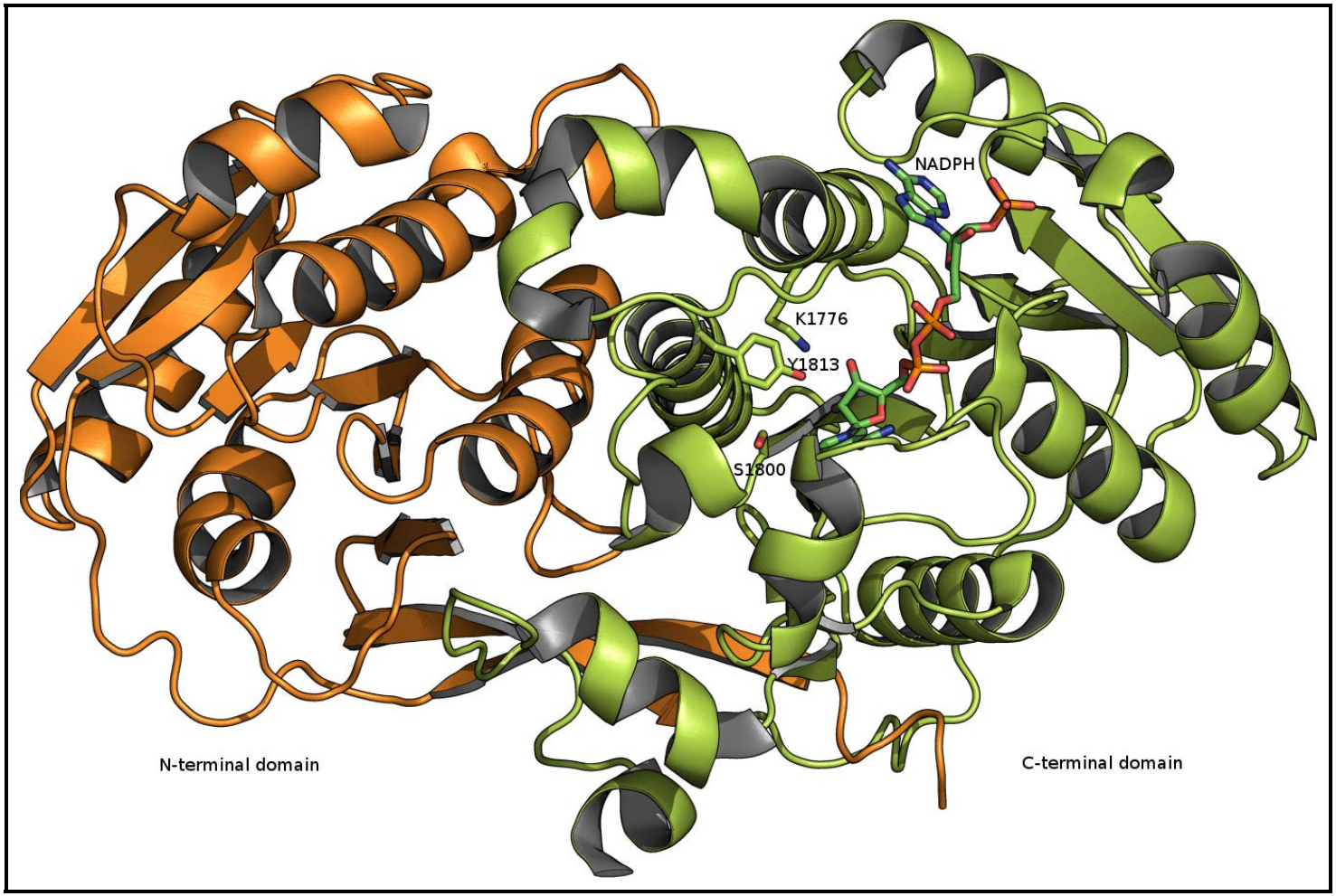

Figure 1.13: *Cartoon representation of a keto reductases domain from the first module of DEBS biosynthesis pathway (PDB ID 2Fr0). The N-terminal structural domain (orange) and the C-terminal catalytic domain (lime) are highlighted along with the catalytic triad Y1813, S1800 and K1776 drawn as sticks. The bound NADPH molecule is shown as sticks (Keatinge-Clay and Stroud 2006).*

reduction. Type A and B can be further sub classified as 0, and 1 and 2 depending upon whether they reduce an un-α-substituted substrate or an α-substituted substrate. A0 and B0 will be type A and B which cater an un-α-substituted substrate whereas A1, B1 accepts D-α-substituted substrates and A2, B2 accepts L-α-substituted substrates (i.e. [LD]-α-alkyl-[LD]-β-hydroxyl). The type C KRs can also be sub classified as C1, which are non functional, and C2, which do not perform β-keto reduction but are capable of epimerase activity for the α-substituted substrates (Keatinge-Clay 2007). These classifications were made on the basis of the conserved motifs found through sequence analysis and subsequent mutagenesis and domain swap experiments (Caffrey 2003; Zheng and Keatinge-Clay 2011).

Ketoreductases belong to the superfamily of short chain dehydrogenases/reductases (SDR) and their structures is made up of two Rossmann like fold domains for ≈480 residues. The N-terminal structural domain (KR_s) does not catalyse the reduction but is thought to be important for providing correct orientation to the C-terminal catalytic domain (KR_c). The junction of the (KR_s) and (KR_c) is marked by the N-terminal β (β1) strand of (KR_s) which is also thought to mark the boundary of the KR, forms a continuous β sheet with the C-terminal β (β7) strand leading to the (KR_c) domain. Similar to the SDRs, KRs in modular polyketide synthases were hypothesized to involve the conserved catalytic triad of tyrosine, serine and lysine to carry out β-keto reduction (Keatinge-Clay and Stroud 2006; Zheng and Keatinge-Clay 2011). Homology modelling of KR6 from the DEBS3 and subsequent mutagenesis study on the catalytic residues proved the importance of these residues in the β-keto processing. Reid *et al.* (2003) found that in the complete DEBS system a KR6 Y2699F mutation completely abolishes the pathway whereas KR6 K2664Q and KR6 S2686A significantly reduce polyketide production. However, in a truncated DEBS Module6+TE system all the three KR6 mutations completely abolish the triketide production. Figure 1.13 shows the equivalent positions of the catalytic residues in the KR crystal structure from the first module of the DEBS biosynthesis pathway. In the proposed reaction mechanism for the PKS KR, similar to SDRs the catalytic tyrosine was responsible for donating the proton to the carbonyl oxygen of the substrate. Serine provided stability to the ligand through hydrogen bonds and lysine was hypothesized to lower the P_ka of the catalytic

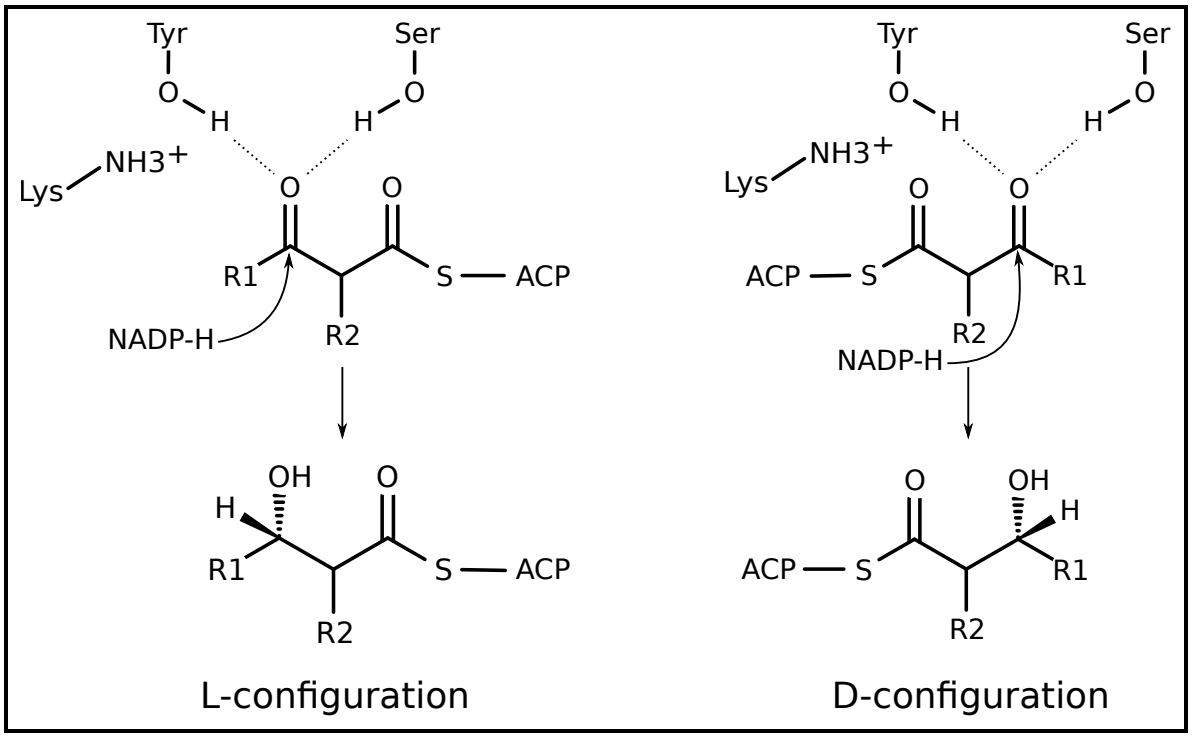

Figure 1.14: *Proposed reaction mechanism for β-keto reduction. Figure adapted from (Reid* et al. *2003).*

tyrosine to assist proton transfer. All the KRs from the DEBS system were also shown to utilize pro-S hydride of NADPH. For the L and D configuration of the resultant β-hydroxyl group it was proposed that keeping the position of the catalytic residues and the bound cofactor (NADPH) fixed there could be two ways in which the substrate presents the β-carbonyl to the active sites (Figure 1.14).

1.2.4.5 Dehydratases (DH)

DH domains catalyse the reversible dehydration of β-hydroxy acyl-ACP to α,β-unsaturated acyl-ACPs in *cis* or *trans* configuration. The dehydratase domains in the bacterial FAS forms a dimer in a structural fold called a "hot dog"(Figure 1.15). This hot dog fold comprises a central helix enveloped by the seven β-strands. In this hot dog dimeric structure there were two active sites formed at the interface in which an active site dyad histidine is contributed from one subunit and an aspartate from the other subunit. On the contrary the DHs in both the mammalian FAS and dimeric PKS (for example curacin DHs) are formed of dimers of double hotdog folds in which the histidine and aspartate are provided by the N and C-terminal hotdog respectively

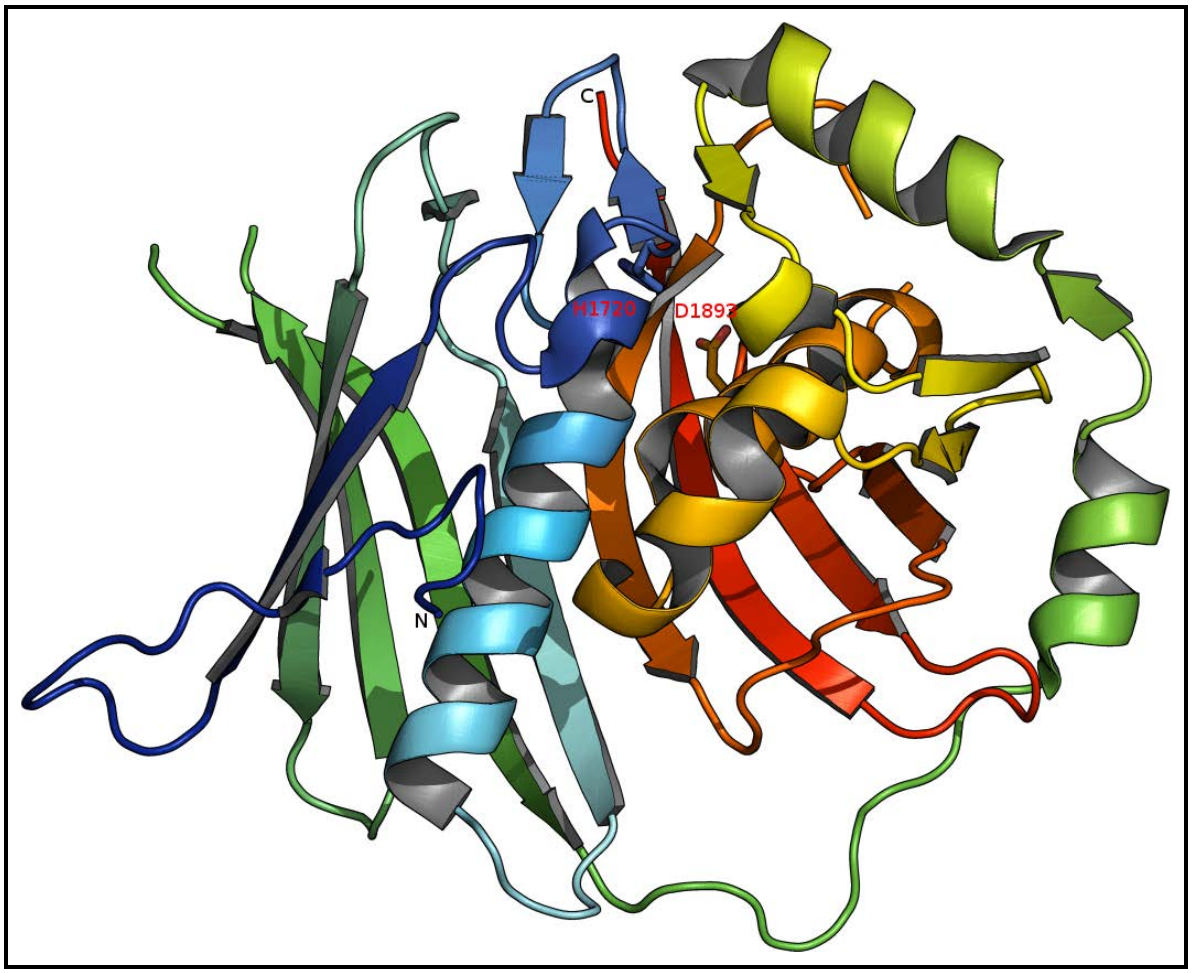

Figure 1.15: *Cartoon representation of the dehydratases monomer domain (CurF PDB ID 3KG6 chain A) from curacin biosynthesis pathway. The double hot dog structure is coloured in rainbow colours with blue at the N-terminal and red at the C-terminal. The catalytic diad H1720 and D1893 are drawn in sticks and labelled (Akey* et al. *2010).*

(Figure 1.15). The orientation of the dimers in the mammalian FAS and dimeric PKS were different with respect to each other (Akey *et al.* 2010). In the proposed reaction mechanism the catalytic histidine which acts as a general base helps in the removal of the proton from the α position and aspartate helps in the removal of β-hydroxyl (Figure 1.16). The stereochemistry of the dehydrated product is thought to be dependent on the configuration of the β-hydroxyl group (Keatinge-Clay and Stroud 2006; Maier *et al.* 2008; Akey *et al.* 2010).

1.2.4.6 Enoyl reductases (ER)

The enoyl reductases (ER) in the PKS systems belong to the superfamily of MDR (medium chain dehydrogenase reductases) proteins and are responsible for the reduction of the enoyl-

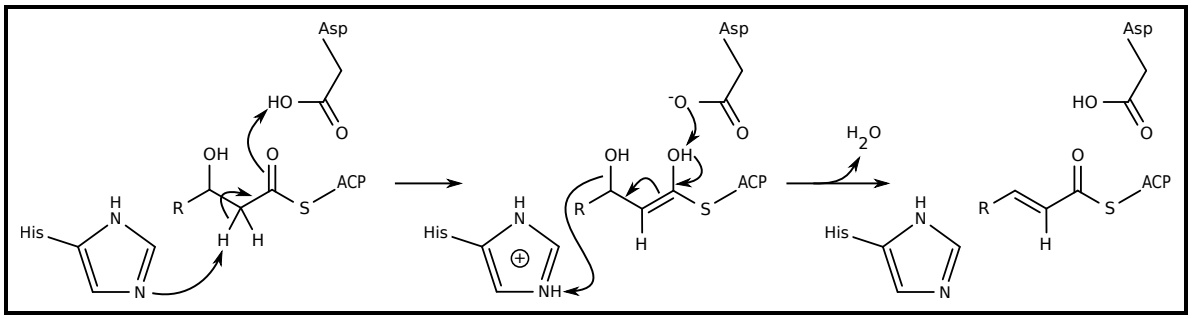

Figure 1.16: *Proposed reaction mechanism for β-hydroxy dehydration.*

ACP to a fully saturated acyl-ACP on the growing polyketide molecule. These domains usually serve as the last domain in fatty acid and polyketide chain extension and modification cycle before the next round of extension begins. Most of the members of MDR superfamily including ERs exist as homo dimers or tetrameters in FASs and PKSs with the exception of the type I iterative PKS, lovastatin, which contains a free standing monomeric ER (Figure 1.17).

The lovastatin ER lovC is composed of two domains the catalytic domain and the cofactor binding domain (Rossmann fold), these structural domains were also found in the other members of the MDR superfamily. The crystal structure solved by Ames *et al.* (2012) and subsequent mutagenesis studies have shown key residues which are responsible for LovC catalysis. Figure 1.17 highlights the active site residues S51, K54, T68, N263 and G282 in the LovC crystal structure responsible for the oxyanion hole formation and catalysis. In the proposed reaction mechanism (Figure 1.18) K54 and G282 (backbone amide) participate in the oxyanion hole formation to stabilize the enolate oxide formed due to the pro-R hydride transfer from an NADPH. In the next step the enolate oxide accepts a proton most likely from a water molecule and produces a fully saturated α,β-acyl-ACP. However, the ER in the type I fungal FAS is an exception to this commonly followed reaction mechanism in which the proton is not transferred by an NADPH but instead it utilizes flavin mononucleotide (Ha *et al.* 2006).

The co-factor (NADPH) binding domain can be characterised by the presence of GXGXXG / AXXXG / A motif and in the DEBS ER4, mutating the NADPH binding motif HAAAGGVGMA to HAAASPVGMA completely abolishes the ER activity. Experiments have shown that ERs have broad substrate specificity and can easily be swapped across different pathways

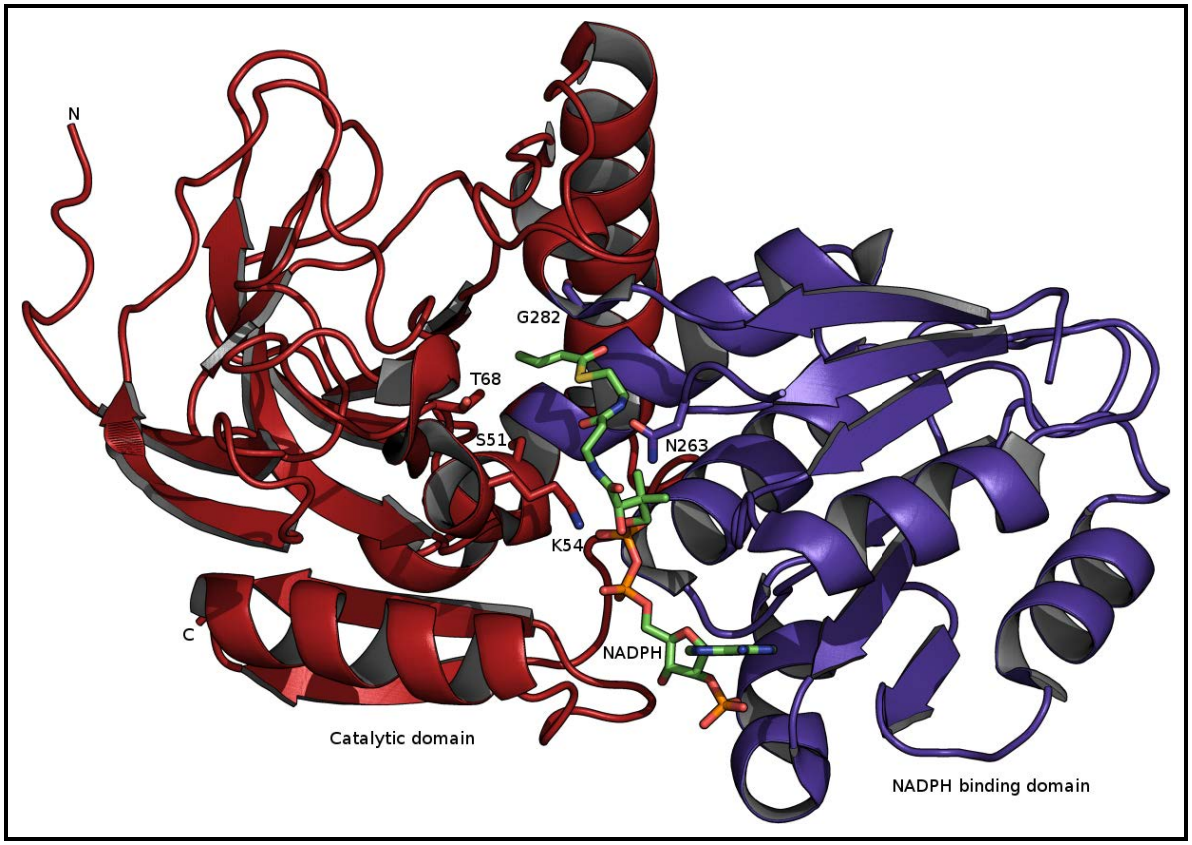

Figure 1.17: *Cartoon representation of the enoyl reductase domain from the lovastatin biosynthesis pathway (LovC PDB ID 3B6Z). The catalytic domain (fire brick red) and the NADPH binding domain (purple blue) are separately highlighted. The key active site resides S51, K54, T68, N263 and G282 and NADPH are drawn as sticks and labelled (Ames* et al. *2012).*

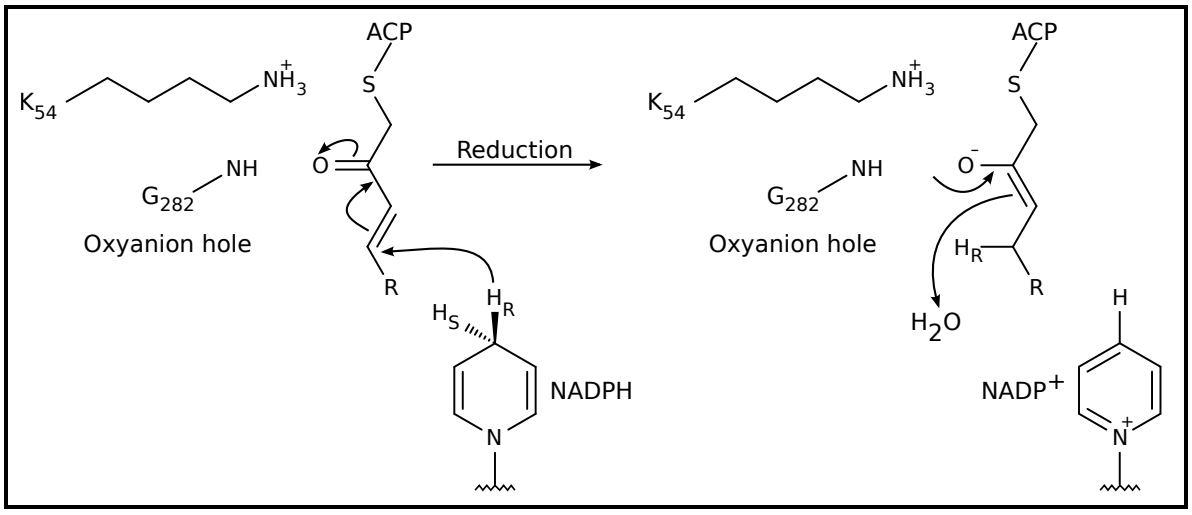

Figure 1.18: *Proposed reaction mechanism for enoyl reduction in lovastatin biosynthesis pathway (LovC). Figure adapted from (Ames* et al. *2012).*

in both FAS and PKS systems (Khosla *et al.* 1999; Witkowski *et al.* 2004a; Khosla *et al.* 2007).

1.2.4.7 Thioesterases (TE)

Thioesterases (TE) are the domains responsible for the hydrolytic release and sometimes macrocyclization of the polyketide chain attached to an ACP. In modular PKSs, TEs are usually the terminal domains however, experiments from Khosla's group have shown that TE domains are also capable of performing well when attached to a bi modular DEBS1 subunit (Gokhale *et al.* 1999). Thus, TE domains are capable of exhibiting a broad range of substrate tolerance. TE domains have a typical α/β hydrolase fold and a S, H and A catalytic triad, both in the type I modular PKSs and type II fungal iterative PKSs. In the DEBS system the hydrolytic release of the polyketide chain has been proposed as a two step process, in the first step the catalytic S142 initiates a nucleophilic attack on the incoming polyketide intermediate resulting in the transacylation of the acyl chain to the catalytic S142 forming an acyl-enzyme complex. In the second step H259 helps to deprotonate the C-13 hydroxyl on the TE bound polyketide intermediate which results in the release and macro cyclization of the 6-deoxyerythronolide B. In the DEBS TE the 14 membered ring is also supported by seven hydrogen bonds inside the active site, which ensures the correct orientation of the C-13 hydroxyl group close to the Ser-O-acyl linkage (Gokhale *et al.* 1999; Sharma and Boddy 2007; Khosla *et al.* 2007). Figure

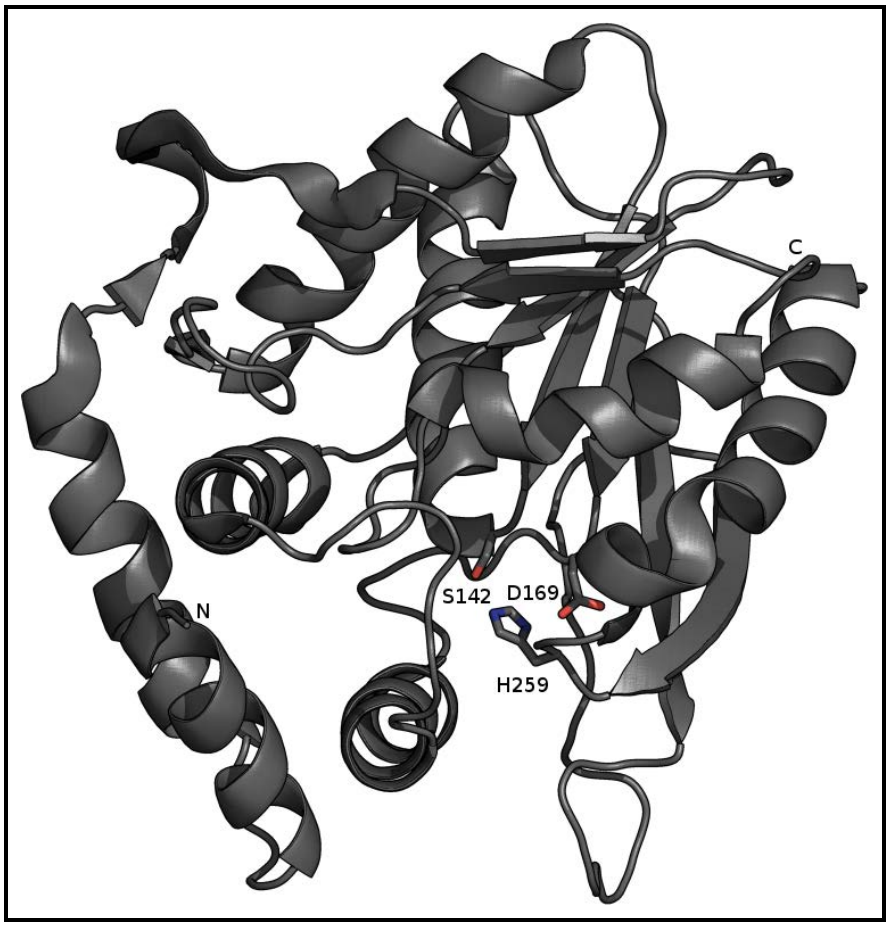

Figure 1.19: *Cartoon representation of the thioesterases domain monomer from DEBS biosynthesis pathway (PDB ID 1KEZ). The catalytic triad S142, H259 and D169 are drawn as sticks and labelled (Tsai* et al. *2001).*

1.19 and 1.20 show the crystal structure of the TE monomer from the DEBS system with the highlighted active site triad and the proposed reaction mechanism respectively (Tsai *et al.* 2001).

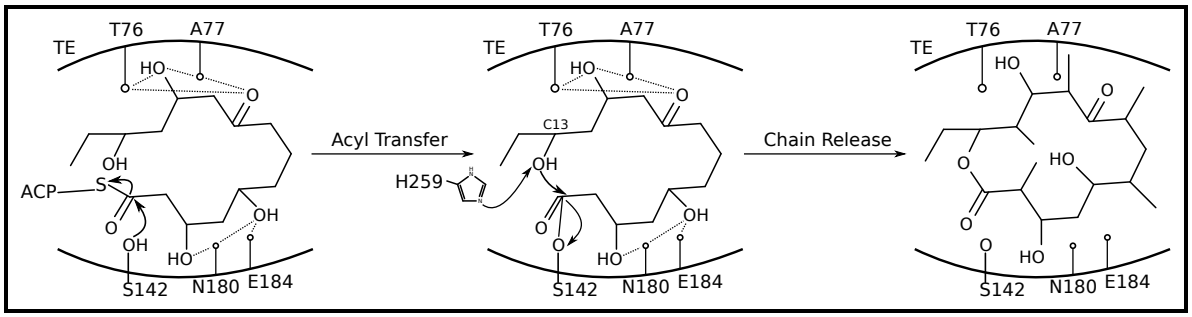

Figure 1.20: *Proposed reaction mechanism for thioesterase in DEBS biosynthesis pathway. Hydrogen bonds are drawn as dotted lines. Figure adapted from (Tsai* et al. *2001).*

1.2.5 Complete modular structures of FAS/PKS

The earliest attempts made to experimentally determine the complete modular structure of an FAS were carried out in the late 1980s using small angle neutron scattering and electron microscopy using samples from chicken, pig and human FAS (Stoops *et al.* 1987; Kitamoto *et al.* 1988; Brink *et al.* 2002). Owing to the poor quality of the electron density measured those models were only used to support the feasibility of the previously proposed head-to-head model over the fully extended head-to-tail model. Those models were also not good enough to correctly determine whether the two chains of the FAS were arranged in a back to back manner or were crossed over.

In the mid 2000s Nenad Ban and his colleagues were able to successfully determine the structure of a complete FAS module from mammalian and fungal species through X-ray crystallography at various resolutions (Maier *et al.* 2006; Jenni *et al.* 2007; Maier *et al.* 2008). And later on in 2013 they produced a cryo-EM structure of the Mycobacterial FAS (Boehringer *et al.* 2013). Similar to the previous attempts Ban's group also struggled in correctly determining the structure of the highly flexible regions such as the terminal ACP and the TE domains and it was only the yeast FAS in which they could successfully locate the position of the ACP. All the other structures were determined without the ACP and the TE domain.

1.2.5.1 Mammalian FAS

The first crystal structure determined by the Ban group was the 4.5 Å resolution mammalian FAS an α_2-homodimeric protein with all the seven FAS domains on each of the two subunits.

In spite of being successful in obtaining crystals of the porcine FAS the resolution was not good enough to correctly determine the position of the side chains in the molecule however, the secondary structural elements were clearly detectable. Therefore, they utilized the electron density map to fit the previously determined crystal structures of individual domains from type II FAS to obtain a domain architecture of the mammalian FAS (PDB ID 2CF2) (Maier *et al.* 2006).

The overall shape of the mammalian FAS was like an X, imagining this as a person, the KS homodimers form the lower and DH domain pseudo dimers form the upper portion of the torso of the body, the MAT domains forming the legs attached to the KS, the ER domain homo dimer the head and the KR domains forming the arms (Figure 1.21). This arrangement of the domains in the quaternary structure was in contrast to the linear arrangement of the domains in the primary structure and has dimensions of 210 Å x 180 Å x 90 Å. The architecture also revealed two main dimerization interfaces, one at the KS and the other at the ER domains covering an area of ≈5000 $Å^2$. The KS dimer was found to be similar to bacterial KS I from FabB and was in agreement with cross linking experiments conducted in the mammalian FAS to cross link the N-terminus of a KS domain with the engineered active site cysteine of the other (Witkowski *et al.* 2004b). The ER domain homodimers form a continuous 12 stranded beta sheet at the interface joining the two nucleotide binding domains, six strands coming from each monomer. Apart from the dimer interfaces between the KS and the ER domains a small portion of the lower end of the DH domains also contribute to the overall dimer formation. The DH domains sit on top of the KS domains with the lower portion of the DH domains making a contact with the upper portion of the KS, domains which forms the waist like region. Linker regions were found in between the KS and the MAT domain and also between the MAT and DH domain. The KR domains are separate single domains hanging out as arms with no contacts between the two which was contrary to observation of tetrameric bacterial KRs (Maier *et al.* 2006; Boehringer *et al.* 2013).

Maier *et al.* (2006) made an interesting observation that the X shape of the mammalian FAS is asymmetrical. The conformation in which the X-ray structure was obtained had different

sized reaction centres in the KS domain. The distance between the KRs and the MAT domains on the same sides were not similar and they were measured to be 72 Å on one side and 87 Å on the other. It was speculated that this difference in the size of the reaction centres may be caused by the mechanism of substrate binding to the KS and the product release after elongation, mediated by a hinge in the "waist" region. In support of this hypothesis, the bacterial KSs in FabB have shown non symmetrical mode of substrate binding. They also proposed that the KS/MAT linker region might also allow the MAT domains to move in the "up and down" direction. The proposed cumulative effect of these motions were to accommodate the ACPs close to the enzymatic domains. As the observed length of the phosphopantetheine arm was only long enough for the bound substrate to reach to the centre of the active sites, implies if the ACPs were in close contact with the individual catalytic domains then there should be enough space in the quaternary arrangement of the FAS domains to accommodate the ACP next to the active site entrance of each domain. This hypothesis also strengthened by work by Zhang *et al.* (2003), which identified the key residues responsible for the interaction between the ACPs and the KR domains in the bacterial type II FAS, thus proving the existence of domain-domain interaction between the ACPs and the FAS catalytic domains.

Following the determination of the mammalian FAS domain architecture based on the 4.5 Å resolution X-ray crystallographic map, Ban's group successfully determined the 3.2 Å resolution crystal structure of FAS in its $NADP^+$ free and bound states (PDB ID 2VZ8). The resolution for this crystal structure was good enough to correctly determine the side chains along with the secondary structure (Figure 1.21). However, even in this attempt they couldn't determine the structure for the ACP and the TE domains. The 3.2 Å resolution structure also agreed with the previously observed X shaped domain organization of the mammalian FAS. However, they identified two additional structural domains. One of the additional domains resembled a methyl transferase but is apparently catalytically inactive, and thus referred to as pseudo methyl transferase, and the other domain was similar to a ketoreductase which was thought to provide structural scaffold for the correct orientation of the catalytically active KR. This structure had similar dimerisation interfaces between the KS and the ER domains to those observed in the 4.5

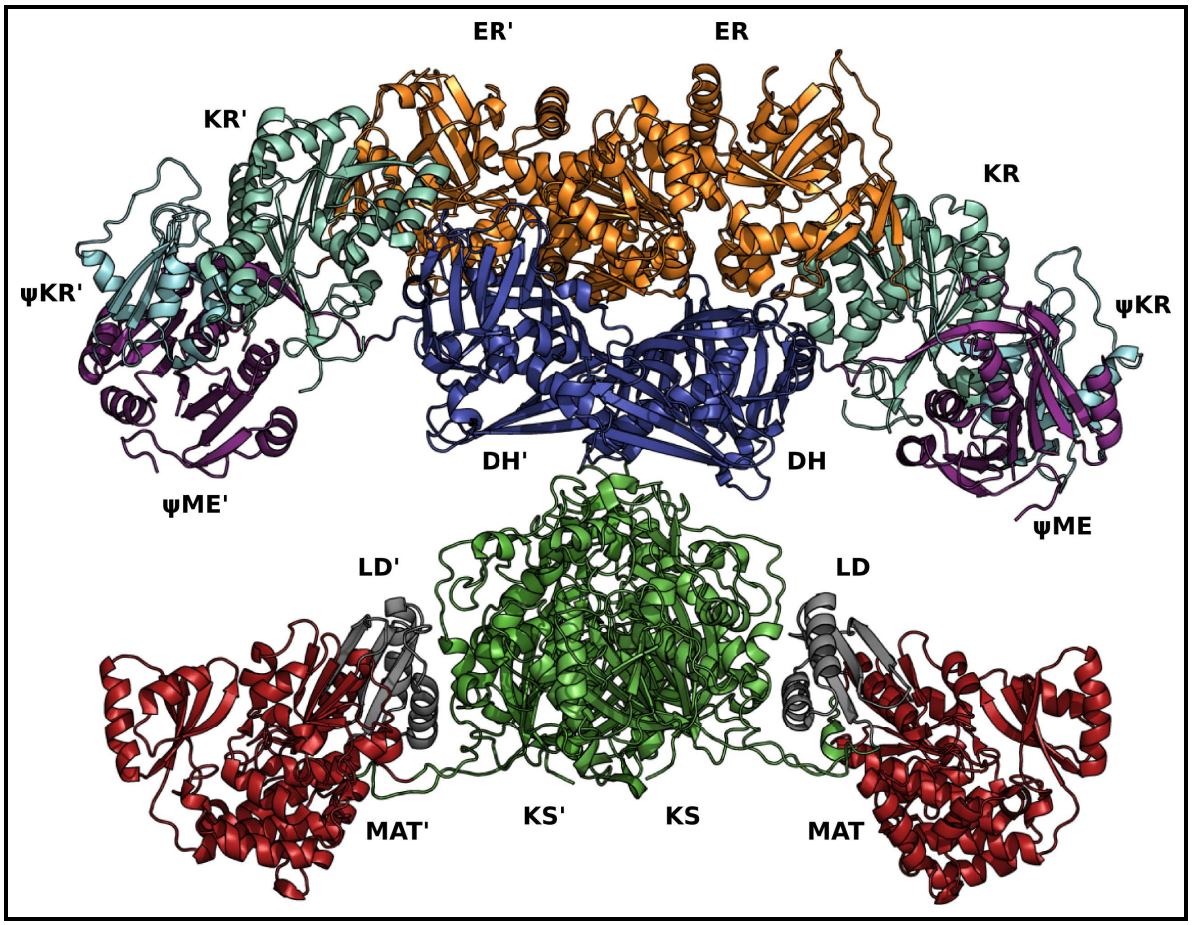

Figure 1.21: *X-ray structure of the mammalian FAS (PDB ID 2VZ8) rendered as cartoon. The catalytic (KS and MAT) and the modifying domains (KR, DH and ER) are coloured and labelled as indicated. The linker domains (LD) are coloured as grey and the pseudo domains (ME and KR) are labelled with the symbol ψ.*

Å resolution model. However, the size measured was according to the electron density observed for the mammalian FAS side chains rather than the fit from the type II FAS discrete domains. The overall contact area in the dimerisation interface of the 3.2 Å was measured to be 5400 $Å^2$ which included 150 amino acids with the contributions of ≈2600 and 1600 $Å^2$ from KS and ER domains respectively. The two DH domains also contributed 800 $Å^2$ to the dimerization interface along with the linker region between MAT, DH and the KS domains. This crystal structure also revealed the linker regions between the KS and the MAT domains, consisting of two helices and three β-strands forming an anti parallel β-sheet. This linker, which was thought to prevent any direct linkage between the KS and MAT domains, was also observed in the KS-AT crystal structure in the type I modular DEBS PKS determined at the same time by Khosla's group (Tang *et al.* 2006b; Tang *et al.* 2007). On the basis of the sequence context this region was also speculated to be present in the trans AT systems and a very recent crystal structure of the KS homodimers from a trans AT system showed the presence of a similar linker region attached to the C-terminal of the KS domains (Gay *et al.* 2014). The precise role of this linker region is still unknown however experiments from Khosla's group have shown that they might be necessary for the ACP docking during the elongation and acyl-transfer stages (Kapur *et al.* 2010).

1.2.5.2 Fungal FAS

In 2007 Ban group solved the crystal structure of 2.6 mDa $\alpha_6\beta_6$ -dodecameric fungal FAS from *Thermomyces lanuginosus* (PDB ID 4V58) and that of *Saccharomyces cerevisiae* (PDB ID 2UV8), both at 3.1 Å resolution (Jenni *et al.* 2007). The structure determined from *S. cerevisiae* also revealed the position of the ACP attached to the KS active site. Although both the mammalian and fungal FAS are type I FAS where the domains are covalently attached in a single polypeptide their quaternary structures are very different from each other. The fungal FAS structure comprised of a central wheel sandwiched between two dome like structures (Figure 1.22) providing two distinct reaction centres with five entrances through the walls and top of the domes. The two reaction centres were connected by six passages through the central wheel. The central wheel was composed of 6 α subunits whereas the domes were composed of 3 β

subunits each. One α and one β subunit join to form a single non-redundant FAS unit. Thus in a fungal FAS complex there are six functional units of FAS formed by 12 polypeptide chains. The α chains consists of KS and the KR domains and the β chains consist of AT, MPT (malonyl/palmitoyl transferase) , DH and ER domains with numerous inter-chain contacts between the domains. All the active sites were found to point towards the inner side of the catalytic centres in the two domes. The fungal FAS also had phosphopantetheinyl transferase (PPT) domains attached to the C-terminal hanging outside the reaction centres. PPT are responsible for attaching the phosphopantetheine arm to the ACP.

The all helical acyl carrier proteins in the fungal FAS were twice the size of their bacterial counterpart with the first four helices overlapping very well with their homologues. The catalytic Serine was found to be on a loop between helices 7 and 8 whereas helix 8 was considered to be the recognition helix when making a contact with the KS. There were three ACPs each in both the reaction centres and these are doubly tethered, to the reaction centre wall at the N-terminus and to the central wheel at the C-terminus. The AT and the MPT domains had similar folds and were thought to be involved in charging ACPs with the acetyl or malonyl, and palmitoyl moieties respectively. The fungal FAS KS domains exhibit broad substrate specificity as compared to their bacterial homologues. A single FAS KS is capable of iteratively catalyse fatty acid chain elongation from C_2 to C_{16} whereas three different KS domains are required in the bacterial FAS. All the KS domains were found to be tightly embedded in the central wheel of the fungal FAS with the active sites pointing towards opposite reaction centres. The ketoreductases found in fungal FAS followed the classical Rossmann fold however they were dimeric as compared to the monomeric mammalian and tetrameteric bacterial KRs. The DH domains in the fungal FAS formed a triple hot dog fold as compared to the mammalian double hot dog fold. The ER domains in the fungal FAS were also different from their mammalian and bacterial counterparts since the fungal ER utilizes flavin mononucleotide (FMN) instead of an NADPH and forms a TIM barrel as opposed to Rossmann fold (Jenni *et al.* 2007; Leibundgut *et al.* 2007).

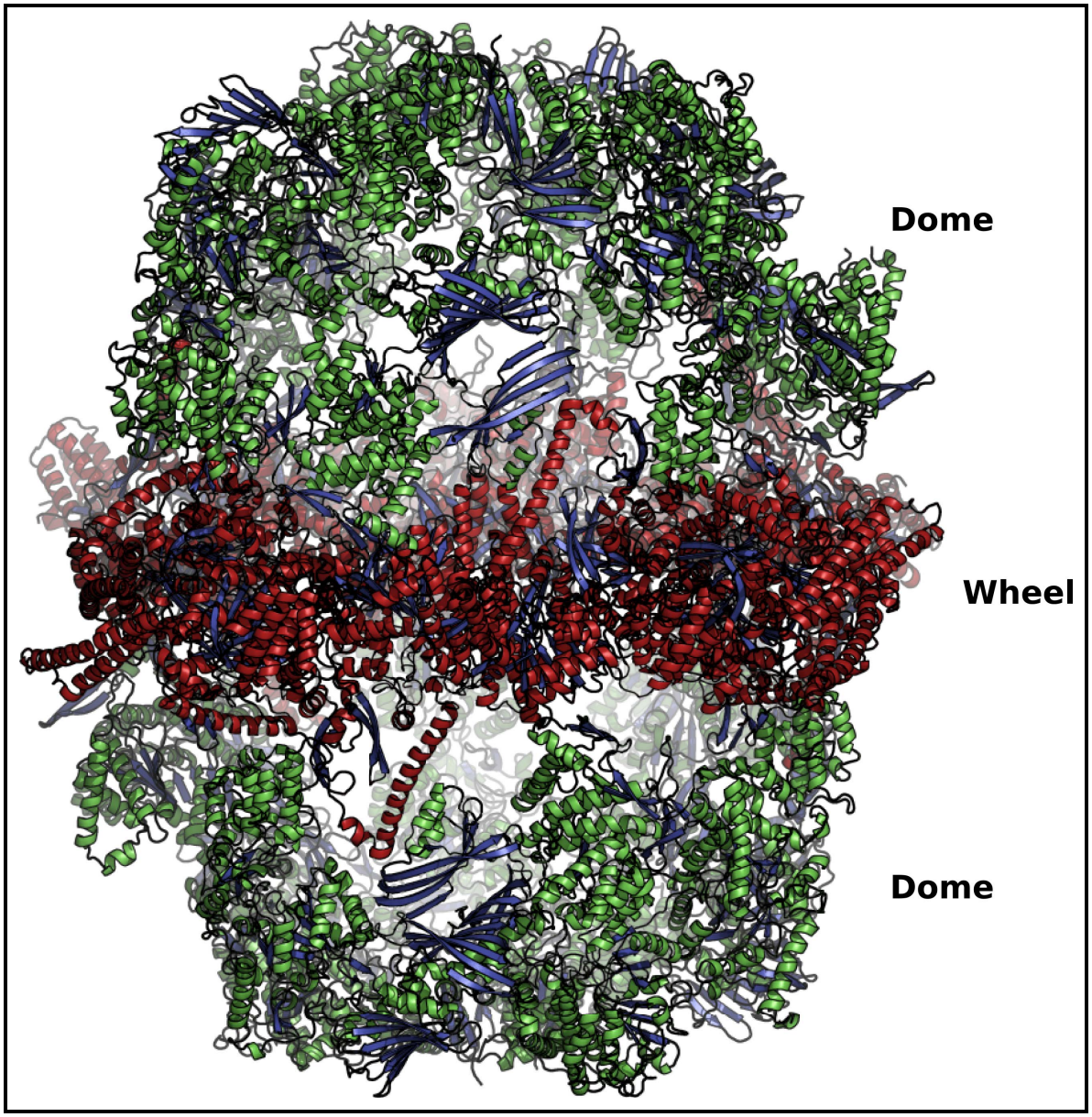

Figure 1.22: *X-ray structure of the fungal FAS (PDB ID 4V58) rendered as cartoon. Helices in the domes and the middle wheel are coloured in green and red respectively. Beta sheets are coloured in blue.*

1.2.5.3 Bacterial type I FAS

In the year 2013 Ban's group published the cryo-EM reconstruction of a Mycobacterial FAS structure at 7.5 Å resolution from the *Mycobacterium smegmatis* species (PDB ID 4V8L, Boehringer *et al.* (2013)). The Mycobacterial FAS showed a barrel shaped domain architecture similar to the fungal FAS (Figure 1.23). However, the structure was more compact with wider openings for the external entrance into the structures. Boehringer *et al.* (2013) used the yeast FAS crystal structure to fit the Mycobacterial EM density map. The 7.5 Å resolution was good enough to detect the individual helices however, the β strands were not clearly visible. Although the flattened shape of the β sheets were detectable.

The *Mycobacteium smegmatis* FAS is a 2.0 mDa α_6 homohexameric structure coded as a single polypeptide containing all the catalytic domains as compared to the domains spread across two polypeptides in the fungal FAS. However, it lacked the phosphopantetheinyl transferase domain attached to the C-terminus. PPTs, called ACP synthases in *Mycobacteria*, are standalone homotrimers encoded by a separate gene.This explains the presence of wider openings in the dome walls to let the ACP synthases reach the ACPs inside the reaction centres within the domes.

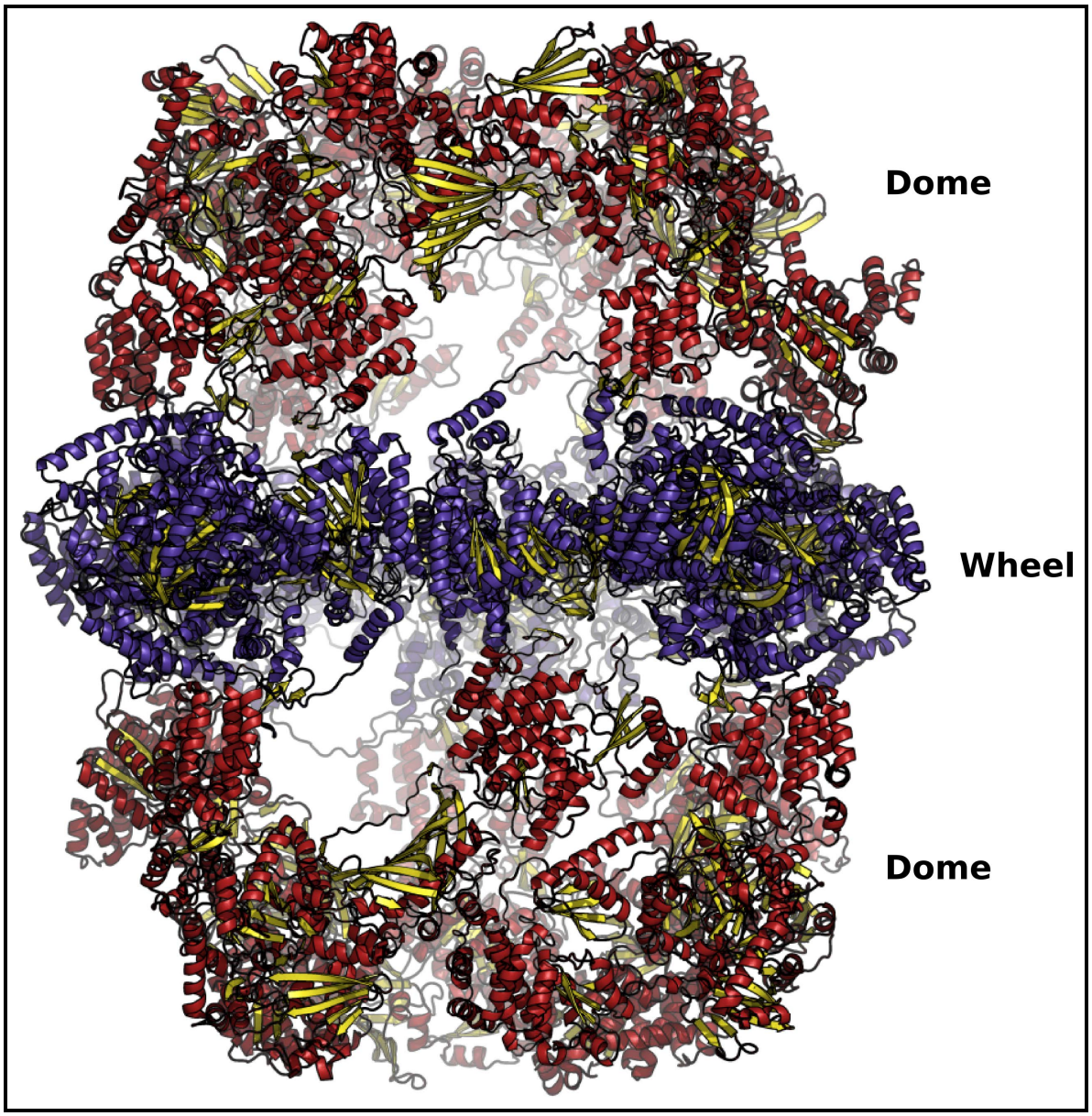

Figure 1.23: *X-ray structure of the mycobacterial FAS (PDB ID 4V8L) rendered as cartoon. Helices in the domes and the middle wheel are coloured in red and blue respectively. Beta sheets are coloured in yellow.*

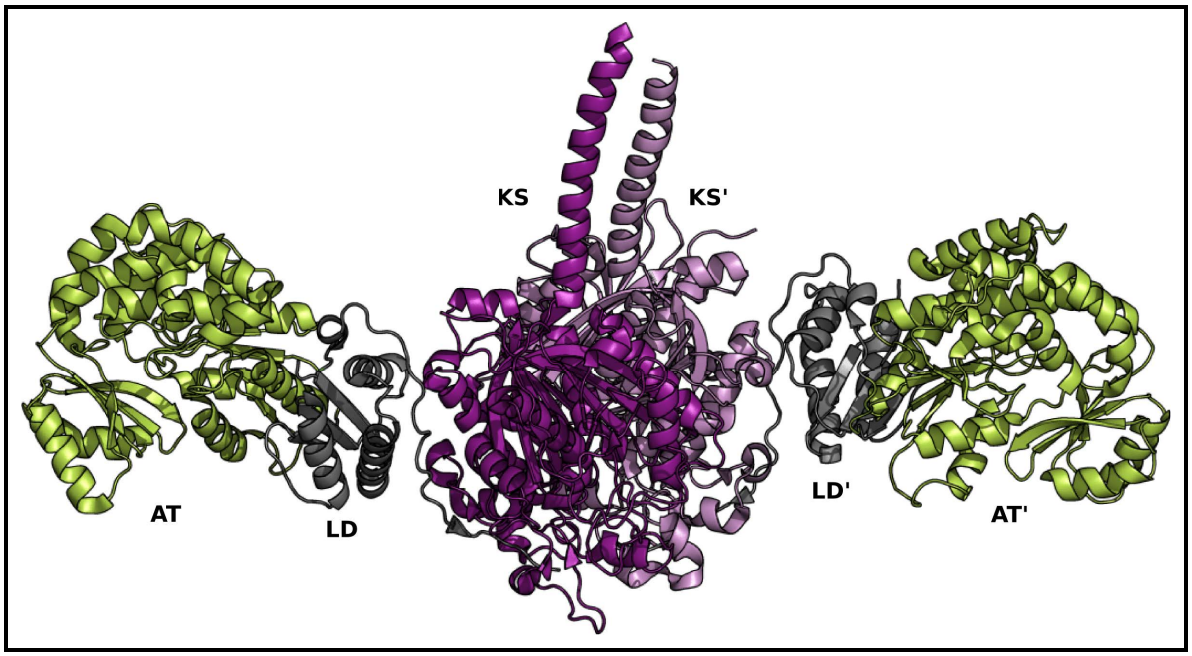

Figure 1.24: *X-ray structure of the KS-AT homodimer from the DEBS (PDB ID 2HG4) system rendered as cartoon. Ketosynthase (KS) dimers are coloured in two different shades of purple, linker domains (LD) are coloured in grey and acyltransferase (AT) domains are coloured in lemon yellow.*

1.2.5.4 *Cis* and *trans* AT PKS

Soon after the FAS structural elucidation by Ban's group the PKS community also released atomic resolution structural models for PKS systems. Khosla and colleagues were only able to crystallize the KS-AT homodimeric structures from DEBS 3^{rd} and 5^{th} modules (Tang *et al.* 2006b; Tang *et al.* 2007), and until very recently those were the only structures available for the PKS systems, which only represent part of the module. The first structure which came out was the 2.7 Å resolution crystal structure of the 194 kD KS-AT homodimer from the module 5 of DEBS system (PDB ID 2HG4). This structure agreed very well with the then resolved 4.5 Å resolution domain architecture structure of the mammalian FAS and was similar to the leg region of the mammalian FAS structure (Figure 1.24).

Apart from the two full length domain structures of the KS and AT domain the solved structure also had three linker regions. The N-terminal helical linker region, the linker region/domain between the KS and AT domains and the C-terminal linker which was thought to connect the AT and the KR domains. The KS domains form the dimer at the two fold axis and cover an area of 2,828 $Å^2$ at the interface. The KS domain fold was similar to their bacterial type II FAS

counterpart and superimposed well with the *E. coli* KS I domain. The N-terminal linker forms a helical dimer which protrudes outwards towards the solvent and does not make any contact with the rest of the protein complex. These linkers were thought to be involved in the inter modular interaction between the N-terminal linker of the KS domains with the C-terminal linker of the ACP from the previous module. The two helices make hydrophobic contacts and a salt bridge between them. The well defined KS to AT linker region was formed by three β-strands sandwiched between three α-helices from one side and two alpha helices from the other side contributed by the AT domains and the AT-KR linker forming an $\alpha\beta\alpha$ fold. The C-terminus of the AT domain folds back towards the KS-AT linker and extends a further 30 aa extend across the surface of the KS, which presumably forms the linker between the AT and the KR, making 8 hydrogen bonds with the KS domain and 5 hydrogen bonds with KS-AT linker. Such a fold was not previously reported in the PDB.

Similar to the KS domains, AT domains also had the similar fold to that of the bacterial type II AT domains and superimposed well with the *E. coli* AT domain, with the RMS deviation of 1.59 Å. This structure also revealed an ≈80 Å distance between the active site C199 and the S642 in the KS and AT domains respectively which suggested an extensive movement of the ACP domains in substrate channelling that would otherwise not be possible just by the phosphopantetheine arm of ≈20 Å.

The active site of the KS domains also matched well with their bacterial type II homologues where it centres around the active site Cysteine reaching towards the dimer interface. The active site cleft was formed by the contribution of the residues from both the subunits and was connected via a tunnel to the outer opening at the surface of the KS domain. The residues lining the tunnel were highly conserved among the KS's homologue with the primary role of supporting the phosphopantetheine arm. Another interesting observation made through the KS-AT homodimer structure was the flexible loop region (residue 153-161) at the dimer interface which forms the part of the active site. This loop region is a helix in the type II KSs. It was hypothesized that this loop might be responsible for the KS specificity and helps in accommodating substrates of different size. This observation agreed well with the previous observations that the KS 5 can

accommodate substrates of larger size than the native ligand (Tang *et al.* 2006b).

The second structure solved by Khosla and colleagues was the 190kD KS-AT homodimer from the DEBS module 3 (PDB ID 2QO3, Tang *et al.* (2007)). The overall fold of this strcutre was similar to the previously determined KS 5 structure however, it lacked the N-terminal linker region. The KS3 domain overlapped with the KS5 domain with an RMS deviation of 0.85 Å for 386 backbone C_α atoms. One obvious difference was observed between the two KSs in the region from residue 71 to 91 (number according to KS3). This loop region on the surface in KS3 was longer as compared to the equivalent region in KS5. Sequence alignment with the other KSs from the DEBS system revealed that all the other KSs in the DEBS system were longer than KS5 by 12 residues in this region.

The KS-AT linker/domain in the KS3-AT3 structure also superimposed well with the KS5-AT5 linker region with an RMSD of 1.54 Å. Tang *et al.* (2007) also observed the difference in the active site lining residues in the KS3 and KS5 active sites. KS5 active site had fewer bulkier residues than KS3 active site, presumably to accommodate a larger substrate. They also found a conformational difference between a loop at the dimer interface from the residue 153 to 158 in the KS3 structure and residue 149 to 154 in the KS5 structure, the region that is a helix in type II KSs as explained above. Thus this loop region shows the importance not only in differentiating between the KS type I and type II but also exhibit differences between the same type from one KS to another.

In 2014, for the first time, a complete modular structure for a PKS module was determined by Dutta *et al.* (2014) for module 5 (PikAIII) in the pikromycin cluster from *Streptomyces venezuelae*. The PikAIII consists of KS, AT, KR and ACP domains, the structure was determined by single particle electron microscopy (cryo-EM) at 7.5 to 9.5 Å resolution. The cryo-EM maps were able to identify the secondary structures, which were used for the rigid body fitting of X-ray structures of the KS, AT, KR, and ACP domains from the DEBS system (Figure 1.25).

The EM structure revealed a striking difference between the PikAIII PKS module and the mammalian FAS structure which was till now considered to be the scaffold for the PKS modules

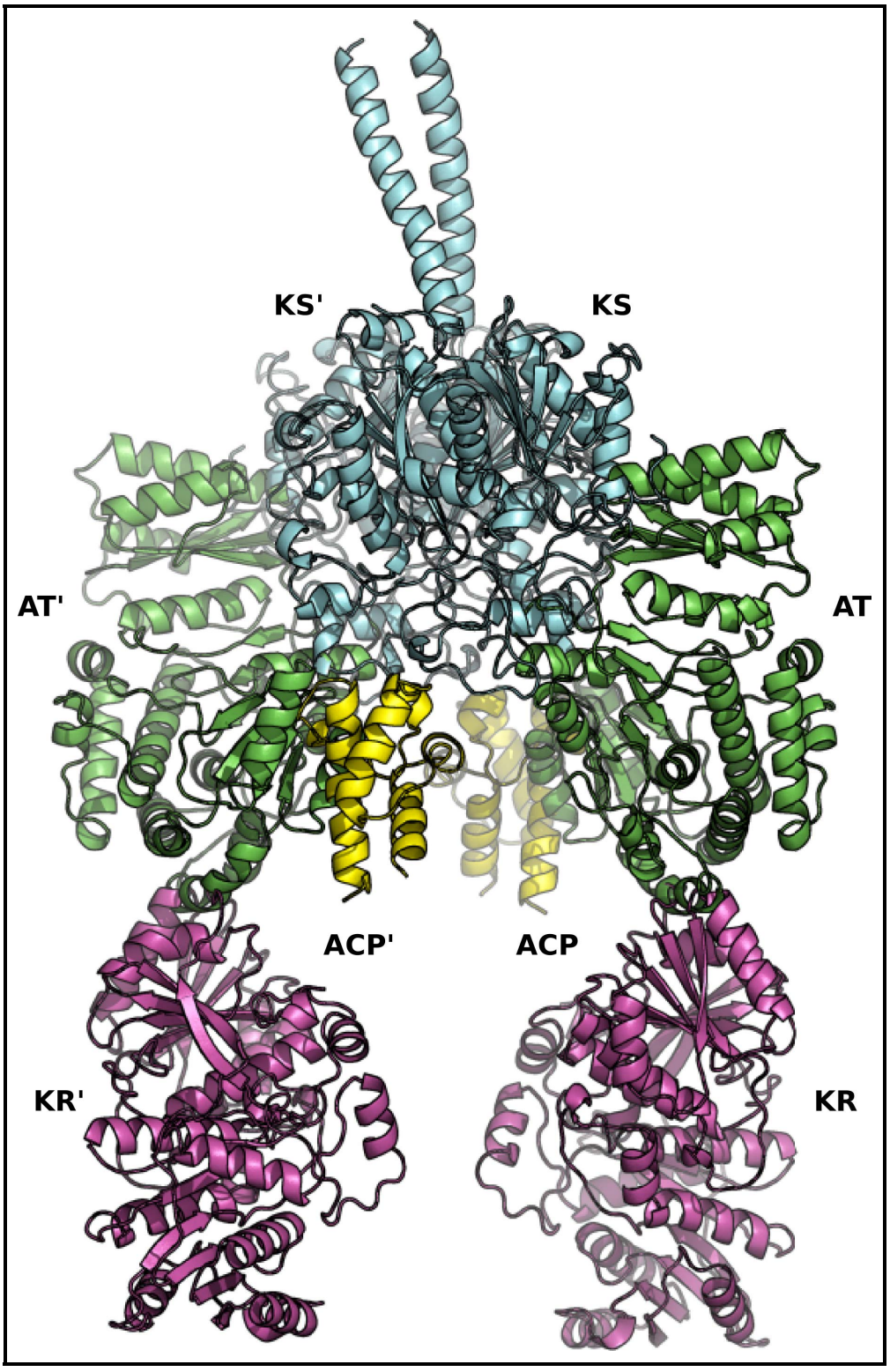

Figure 1.25: *EM structure of module 5 (PikAIII) from pikromycin cluster. Ketosynthase (KS) dimers are coloured in cyan, acyl transferase (AT) domains are coloured in green, ketoreductase (KR) domains are coloured in magenta and acyl carrier protein (ACP) domains are coloured in yellow (Dutta* et al. *2014).*

as well. The PikAIII structure revealed a single reaction chamber utilized by that module's ACP to reach all the catalytic domains within the module, whereas the ACP from the previous module delivers the polyketide intermediated through a separate entrance outside the reaction chamber. The PikAIII symmetrical dimer folds into an arch shaped structure with KSs forming the dimer interface at the dome with the ATs hanging on either side of the KS dimer rotated at ≈120°from their relative positions in mammalian FAS and also DEBS KS-AT dimer structures. The AT domains were also found to form an extensive interface with the KSs as compared to their counterparts in the mammalian FAS and DEBS KS-AT structures. The KR domains are lowered towards the base of the arch attached to the AT domains with the active sites of the AT and KR domains facing towards the reaction chamber. The ACPs were found either at the top, just below the KSs in between the two ATs, or at the base of the arch sandwiched between the two KRs. No evidence was detected for the two ACPs to be in different positions at the same time for example one near the ATs and the other near the KRs.

The cryo-EM maps also revealed that the linkers between the KRs and ACPs are long enough to reach both the KSs in the dimer. In another work from the same group, Whicher *et al.* (2014) have shown a structural rearrangement of the KS-AT domains upon substrate binding in the PikAIII. In the parent paper, Dutta *et al.* (2014) have also shown that dimer formation of the PikAIII is not completely dependent on KS dimerization but is partly contributed by the post ACP dimerization helices. The presence or absence of the ACP domains was also found to influence the orientation of the KR domains. In a PikAIII Δ ACP5 strain, the KR domains were found to be rotated at ≈165°about the axis of the arch legs. Upon expressing ACP4 from the previous module into the PikAIII Δ ACP5 strain the ACP4 was found to be interacting at the N-terminal docking domain of the module 5 KSs, completely away from the reaction chamber where ACP5 would be found. This evidence established the involvement of a separate entrance for the acyl chain transfer on the KSs from an upstream module than utilizing the same reaction chamber involved in the chain elongation and β-carbon processing. In another unexpected observation KSs were found to have a second entrance to the active site for the entry of the extender units during elongation reaction. This second entrance was previously not reported

either in the FAS or PKS structures.

1.2.6 An example of re-engineering PKSs

A number of groups have made contributions to our current understanding of PKS pathways and how to re-engineer them (Weissman and Leadlay 2005; Challis 2008; Piel 2010; Kwon *et al.* 2012), with the macrolide systems being particularly popular systems for study (Park *et al.* 2010). Much work has been focused on the type I PKSs. For example, Khosla's group have extensively manipulated and modelled the DEBS system (Khosla *et al.* 2007). In their work using a number of chimeric constructs of ACPs, crystal structure determination and computational protein-protein docking revealed how an ACP has specificity for the chain elongation in the DEBS pathway (Kapur *et al.* 2010). How the same ACP specifically passes the processed substrate onto the next module (Kapur *et al.* 2010; Kapur *et al.* 2012), and why a PKS dimer might be required for function.

Chimeras of ACP3 and ACP6 from deoxyerythronolide B synthase, the ACPs from modules 3 and 6 of the synthetic pathway, indicate that the loop between helix I and helix II were critical during the elongation process of synthesis. Computer docking of the ACP onto the crystal structure of the KS5-AT5 homodimer indicated two residues in the loop that appeared critical for electrostatic complementarity (D44 and R45). Further modelling showed electrostatic complementarity between the KS-AT didomain in each module and the equivalent residues in their cognate ACP, ACP3 (R44, R45), ACP4 (R44, K45), ACP5 (D44, R45) and ACP6 (D44, Q45). Mutations R44A/R45A in ACP3 confirmed the importance of the residues at these positions. The authors noted that these key residues of the ACP interact with the linker (docking domain) between the KS and AT modules, as well as the AT module, and thus this mechanism cannot be the same for PKSs that have the AT acting in trans unless it docks in a similar orientation to the cis AT.

In contrast to the elongation mechanism, for the transfer of the substrate from an ACP to the KS on the next module they found a different mechanism, one that relied on the 1^{st} ten residues in helix I of the ACP. They also concluded that the PKS domains are made up of subdomains and

they have a distinct role in mediating interdomain interaction. These experiments also support the idea of homodimeric architecture of multimodular PKSs. This work emphasizes the benefit of computational modelling working together with well-designed experiments to elucidate the engineering principles underlying the PKS.

From understanding the principle behind the ACP-KS recognition in the polyketide chain elongation and translocation steps, an obvious question arises. Is there a way that this unidirectional ratchet flow of the pathway can be modified? Khosla's group identified a charge complementarity mechanism during chain elongation at the position 23 on the ACP at the KS-AT interaction interface. They found that the opposite charge on KS-AT linker of the following module attracts the preceding ACP, whereas the similar charge on the KS-AT linker of the parent module repels it, thus allowing the reaction to always move forward. Exploiting this mechanism they re-programmed module 3 of the DEBs system to carry out an iterative step of chain elongation. In an experiment they swapped the helix I of ACP3 with the helix I of ACP2, as the helix I is responsible for the chain elongation step. This swapping did not allow the ACP3 to transfer the chain to the module 4 instead it falls back to the KS3 in the module 3 for one more round of chain elongation. The reaction stopped after one round of iterative chain elongation for reasons unknown. However, it could be a limitation of the KS active site to accommodate the larger substrate.

1.3 NRPS

Nonribosomal peptide syntetases (NRPSs) are also highly prevalent in secondary metabolite biosynthesis and often synthesize compounds in conjunction with PKSs. The NRPS and PKS systems have similar modular structures and thus NRPSs warrant some discussion here. Nonribosomal peptides are the product of the sequential addition of amino acid monomers catalyzed by NRPSs, involving the domains similar in function to that of PKS system. The amino acid monomers are selected and activated by an adenylation (A) domain which transfers to a thiolation domain or a peptide carrier protein (T or PCP), peptide bond formation is catalyzed by a condensation (C) domain. The PCP has a phosphopantetheine prosthetic group that facili-

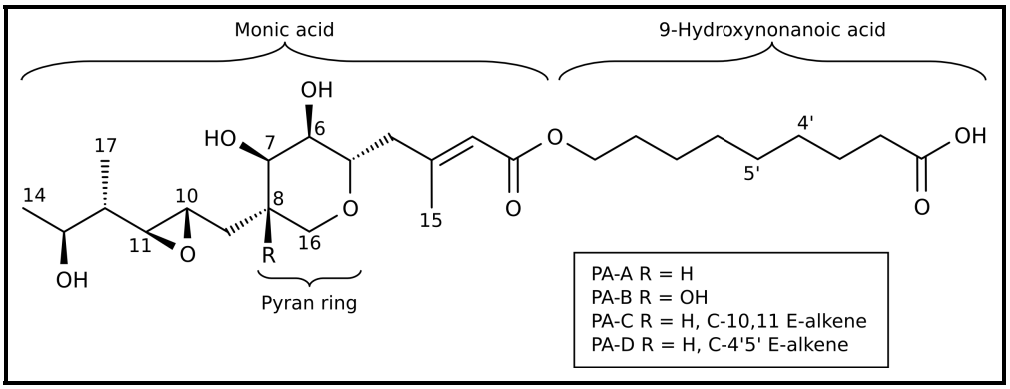

Figure 1.26: *Mupirocin structure*

tates the transfer of growing peptide chain/monomers to the various active sites and function analogously to an ACP with a similar four helix bundle structure. The adenylation (A) domain, condensation (C) domain and thiolation or peptide carrier protein (PCP) form the minimal set of domains required to carry out the NRP biosynthesis. In both, the modular PKS and NRPS system the individual domains are linked together by a polypeptide linker region which has also been found to be responsible for functional communication within the domains.

1.4 Mupirocin

Mupirocin is a polyketide antibiotic used as a topical drug against MRSA and various other Gram positive bacteria, to treat bacterial skin and hospital acquired infections (Fuller *et al.* 1971). Mupirocin is produced by *Pseudomonas fluorescens* NCIMB 10586 and consists of four different pseudomonic acids (PAs) A,B,C and D. PA A, B, C, D contributes 90%, 8%, <2% and <2% respectively to the mixture. PAs are made up of monic acid esterified to 9-hydroxynonanoic acid (Figure 1.26). Mupirocin acts by inhibiting bacterial isoleucine tRNA synthetase thus inhibiting protein production resulting in the cell death (Hughes and Mellows 1978).

1.4.1 Mupirocin Drawbacks

Mupirocin has proved to be more potent on bacterial IleRS than on the corresponding eukaryotic enzyme, which is highly desirable for any pharmaceutical product. Weak affinity for the eukaryotic IleRS minimises the chances of eukaryotic toxicity and side effects of the drug.

Thus, the pseudomonic acids have excellent structures and biological activity as pharmaceutical compounds. However, mupirocin can only be used as a tropical drug as it gets disintegrated inside the body by the hydrolysis of the ester bond between monic and 9-hydoxynonanoic acid (Figure 1.26). It also loses its activity at higher pH, thus preventing it from being used as a systemic drug.

Recent studies (Patel *et al.* 2009; Thomas *et al.* 2010) have also shown increasing occurrences of mupirocin resistance in *S. aureus*. This resistance can be of low level as well as high level. In low level Mupirocin resistance a single amino acid mutation has been observed in the Rossmann fold region of IleRS, which normally binds to the ATP or else to the 9-hydroxynonanoic acid of Mupirocin. Hence the single amino acid change prevents Mupirocin from inhibiting its target. High level resistance usually results from the presence of eukaryotic like IleRS. This can be acquired either through horizontal gene transfer or through a plasmid containing the mupA gene. The *P. fluorescens* strain, which produces mupirocin, has two different IleRS producing genes of which one is similar to eukaryotic IleRS. This eukaryotic similar IleRS enables *P. fluorescens* to keep on synthesising proteins even at high concentration of Mupirocin. Low-level resistance is not of present concern but the high-level resistance in clinical strains of *S. aureus* acquired through the expression of the mupA gene is alarming. Many of the plasmids containing these genes can conjugate which may cause clonal expansion of the resistant strains.

Therefore with increasing levels of Mupirocin resistance and its limitation for systemic use, there is a need to develop new analogues which may overcome the limitations of Mupirocin. However, before we reach up to the level of producing new analogues we need to understand better the underlying pathway.

1.4.2 Mupirocin Biosynthesis

The mupirocin biosynthetic cluster (*mup*) consists of a 75kb region encoding 35 ORFs. Figure 1.5 explains the mup cluster and the probable pathway as proposed by Thomas and co- workers (Thomas *et al.* 2010; Gurney and Thomas 2011). The *mup* cluster encodes six designated MMPs

(Mupirocin multifunctional proteins; *mmpA* to *mmpF*) out of which five encode for polyketide synthases. The non-PKS domain is MmpC, encoding two acyltransferase domains, which is a characteristic of a trans AT system. The proposed biosynthesic pathway initiates in a typical Type I PKS manner where MmpD holds the first starter unit and produces a C_{12} unit, with MmpA extending this to C_{17}, with further tailoring enzymes producing monic acid. MmpD and MmpA consist of four and three modules respectively each containing a KS and ACP domain. MmpD also consists of KR, DH and MT domains while MmpA only has one KR domain. The first module in MmpD has a non-functional DH and the first module in MmpA is proposed to be involved with the transfer of the growing polyketide chain from MmpD to MmpA. Mupirocin biosynthesis starts with an activated acetyl as a starter unit and uses malonyl-CoA as an extender unit. C_{16} and C_{17} carbons in the monic acid unit are derived from S-adenosyl methionine (SAM).

The C_{15} carbon is hypothesised to be incorporated by the HMG-CoA synthases activity of MupH in the HCS cassette (MupG, MupH, MupJ and MupK). Knocking out any of the HCS genes stalls the pathway at the end of MmpA and results in the accumulation of intermediate mupirocin H. (Wu *et al.* 2007). The HCS cassette has at least four enzymatic functions: an ACP, a hydroxymethyl glutaryl-CoA synthase, a decarboxylase and one or more dehydratases from the crotonase superfamily, in some cases supplemented with other functions. HCS cassettes variously introduce e.g. β-methyl, cyclopropane and vinyl chloride moieties, depending on the system they are found in and the exact nature of the cassette. HCS cassettes are found to act on the type I modules, which lack KR, DH or ER domains but which tend to have tandemly repeated ACPs. Given the variety of functions that the HCS cassette can perform, artificially including them in a PKS cluster provides a powerful tool for the synthesis of novel compounds. To achieve this we need to understand what allows the enzymes of the cassette to know at what point in the pathway they are supposed to work, in particular how a system such as myxovirescin uses two HCS cassettes to produce two specific β modifications in the same system.

MupC, mupF, mupO, mupU, mupV and *macp*E provides the correct oxidation state around the pyran ring and mutating any of these genes would result in the accumulation of pseudomonic acid B. Whereas MupT and MupW are hypothesised to be involved in the activation of methyl

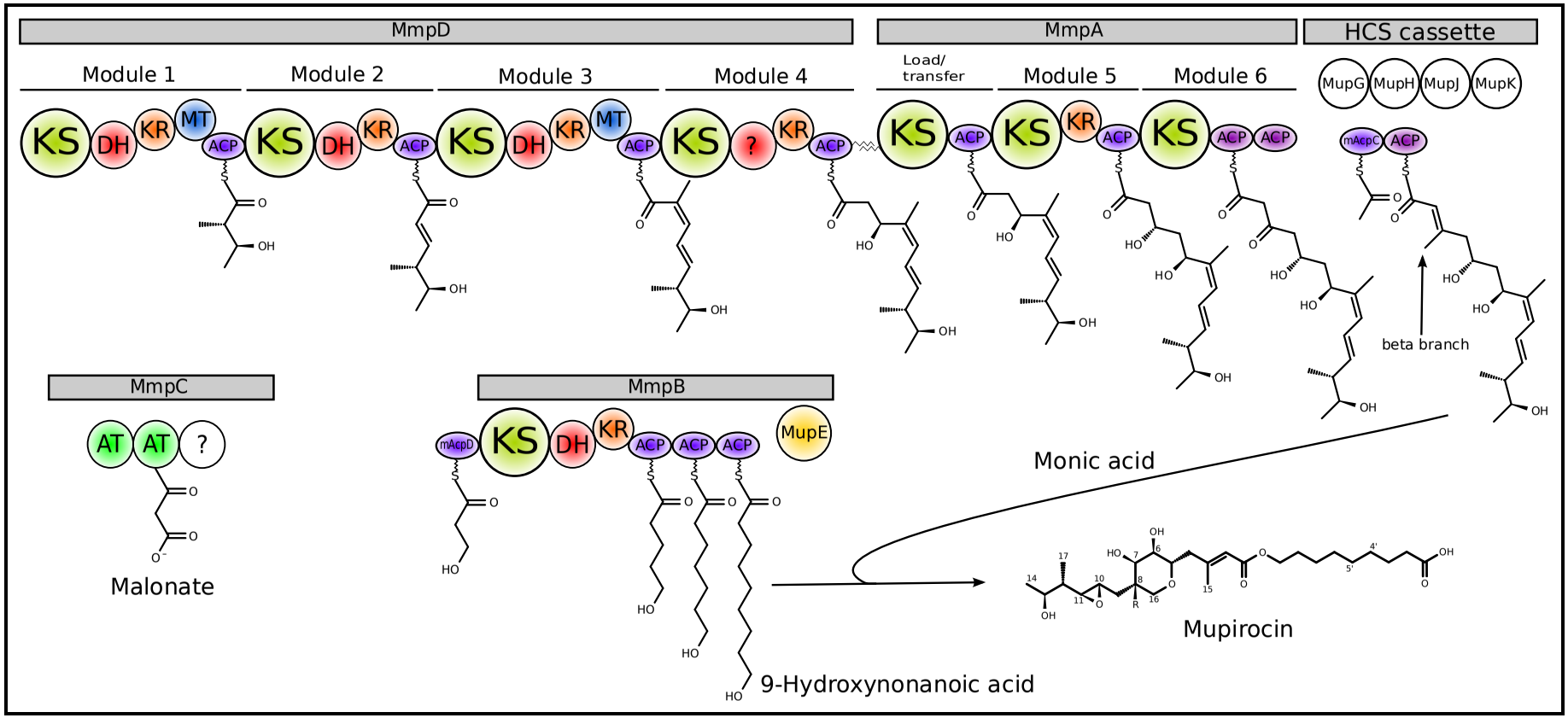

Figure 1.27: *Mupirocin biosynthesis pathway. Figure adapted from (Thomas* et al. *2010).*

group C_8 in order to create the pyran ring.

Finally, MmpB is proposed to synthesise 9-hydroxynonanoic acid starting from 3-hydroxy-propionate via three condensations with malonate iterated three times in the same PKS. 3-hydroxypropionate starter unit is hypothesized to be produced jointly by mAcpD, MupS and MupQ. It is also not understood whether 3-hydroxy-propionate is first attached to mAcpD and then elongated to produce 9-hydroxynonanoic acid or is first esterified to monic acid and then elongated further.

The mupirocin biosynthesis pathway has an ACP doublet (ACPmupA3a/b in MmpA) and an ACP triplet (ACP5, 6 and 7 in MmpB) which is thought to hold multiple substrates for high throughput in rate limiting steps. Apart from the six orfs, which code for the Mmps, the rest of the orfs encode for discrete polypeptides.

1.5 Bioinformatics approaches in PKS research

Research into PKS pathways has been carried out in two major areas. Firstly to identify and experimentally characterize new polyketide natural products and secondly, to develop a synthetic biology tool box for the design and synthesis of novel natural products by the re-engineering of naturally occurring polyketide biosynthetic machinery, both of which might provide the basis for novel drugs. Bioinformatics analysis of PKS clusters has played a major role in guiding recent experiments via a variety of tools, examples are PKSDB (Yadav *et al.* 2003a) a database of PKS domains developed a decade ago, which was one of the first tools available, SBSPKS (Anand *et al.* 2010) and antiSMASH (Medema *et al.* 2011), which are recent advanced sequence analysis tools guided by protein structure analysis.

Most work in the field has focused on the core PKS functions, particularly type I PKSs, thus there is still little known about the mechanisms of the tailoring enzymes acting in trans. Such tailoring enzymes are an essential part of any PKS system since they provide additional chemical diversity to the polyketide chain. An ability to predict the function of tailoring enzymes and their compatibility with other PKS modules would provide much additional functionality to the synthetic biology tool box. *Trans*-AT systems are less well studied and since software

tools are primarily targeted at *cis*-AT systems they tend to completely fail or have limited capabilities in predicting *trans*-AT systems. PKSDB/SEARCHPKS and NRPS-PKS were the first available webservers for identifying PKS/NRPS domains in an unknown sequence as well as relating PKS/NRPS sequences to their corresponding secondary metabolites; recent reviews (Jenke-Kodama and Dittmann 2009; Bachmann and Ravel 2009) give a detailed discussion on the utility of the PKSDB and the NRPS-PKS databases. Following similar lines, resources like ASMPKS (Tae *et al.* 2007), ClustScan (Starcevic *et al.* 2008), CLUSEAN (Weber *et al.* 2009), NP.searcher (Li *et al.* 2009), NRPSpredictor (Röttig *et al.* 2011), NRPSsp (Prieto *et al.* 2012) and antiSMASH (Medema *et al.* 2011) have been developed for the discovery of secondary metabolites through genome analysis. All these servers primarily utilize sequence information either for domain identification or to correlate the various PKS domains to their corresponding metabolic products. However, many of them also utilize structural information for predicting the most likely starter and extender units picked by the acyl transferase domains, and SBSPKS models the 3D structure of PKS modules. Table 1.1 gives a summary of the resources available for PKS/NRPS pathway analysis with a detailed description in Section D.1 of Appendix IV.

Table 1.1: *Resources available for secondary metabolite prediction.*

Resources	**Clusters/ types**	**Prediction tools**	**Domain for which specifity is predicted**	**Backend or Training data source**	**Hyperlink**
SEARCHPKS, PKSDB (database)	PKS	BLAST	AT	PKSDB	http://www.nii.res.in/searchpks.html
NRPS-PKS (database)	NRPS, PKS	BLAST	AT, A	PKSDB, NRPSDB, ITERDB, CHSDB	http://www.nii.res.in/nrps-pks.html
ASMPKS	PKS	GLIMMER, BLAST	AT	PKSDB, More	http://gate.smallsoft.co.kr:8008/~hstae/asmpks/genome.pl
NRPSpredictor, NRPSpredictor2	NRPS	SVM, TSVM	A	Training data amalgamated from various sources	http://nrps.informatik.uni-tuebingen.de/Controller?cmd=SubmitJob
CLUSTSCAN, CompGen (homologous recombination module)	PKS, NRPS, PKS-NRPS hybrid	HMM	KR, AT	Pfam, Specialized	http://bioserv.pbf.hr/cms/, http://bioserv.pbf.hr/cms/index.php?page=compgen
SBSPKS	NRPS, PKS	BLAST, 3D structure modelling	AT, A	PKSDB, NRPSDB, ITERDB, CHSDB	http://www.nii.ac.in/~pksdb/sbspks/master.html
NORINE (database)	NRPS products, monomers				http://bioinfo.lifl.fr/norine/
CLUSEAN (Perl module framework)	NRPS, PKS	BLAST, HMM	A	NCBI NR, Pfam, Specialized	http://redmine.secondarymetabolites.org/projects/clusean

antiSMASH (metaserver & standalone)	NRPS, PKS, terpenes, aminoglycosides, aminocoumarins, indolocarbazoles, lantibiotics, bacteriocins, nucleosides, -lactams, butyrolactones, siderophores, melanins and others	NCBI BLAST+, HMMer, Muscle, Glimmer, FastTree, TreeGraph	AT, A, KR	Amalgamated from various previously published works	http://antismash.secondarymetabolites.org/

1.6 Research objectives and thesis outline

The mupirocin biosynthesis pathway provides a model system for understanding *trans*-AT PKSs, and contributing to our understanding of the way proteins and substrate recognition events are regulated. In the present work the analysis was mainly focused on the structural modelling of the protein complexes involved in the mupirocin synthesis via the tools of structural bioinformatics supplemented with information from experiments. The findings in this thesis will be a prototype for the modelling of large macromolecular complexes in prokaryotic systems and the structural models produced will enhance our understanding of the synthetic pathways allowing us to re-engineer them with greater success.

The overall thesis is divided into four results chapters Chapter 3 to 6, which discuss the results of three major projects in their respective chapters and two minor projects compiled in one chapter. Chapter 2 describes the methods used in the different projects in detail. The last chapter of the thesis summarizes and discusses the entire thesis.

The project discussed in Chapter 3, aimed to elucidate the mode of interaction between ACP-mupA3ab:MupH complex, which are involved in beta branching, proposing key interacting residues for mutagenesis experiments. Structural models and properties of MupH the HMG-CoA synthase homologue in HCS cassette were predicted using various bioinformatics

tools. ACP sequences from characterised PKS pathways were also classified into beta and non-beta branching type using hidden Markov model analysis. Mutagenesis experiments carried out by Prof. Thomas' group on the predicted ACP:MupH interface supports the predicted complex and residues responsible for the specificity of the interaction. Some of the findings of this project are already published in Haines *et al.* (2013).

Based on the hypothesis generated in the first project, further lab experiments were carried out as described in Chapter 4. The aim was to complement the β-branching-ACPs in the mupirocin cluster with the β-branching-ACP(s) from the kalimantacin cluster. It was hypothesized that the β-branching ACP(s) from the kalimantacin cluster would not work with MupH but will work with BatC (MupH equivalent form kalimantacin cluster). The aim was also to identify the key residues at the interface of the β-branching ACPs from the *mup* cluster and BatC which can be modified to work efficiently with BatC.

As the experiments and molecular dynamics (MD) simulations carried out on FAS ACPs (Chan *et al.* 2008) have shown the formation of a hydrophobic tunnel which sequesters the fatty acid chain, the third project was to see if a similar mechanism might exist in the PKS ACPs. Molecular dynamics simulations of the ACPs from module 2 and 3 of MmpA in the mupirocin cluster were carried out in explicit solvent with a time scale of 50 ns to 1 μs. Several macroscopic properties were calculated by analysing the MD simulation trajectories, a sequestering of the substrate to the ACP surface was seen but the atomic details are different from what was seen in the FAS system.

Chapter 6 presents the results of two independent projects, the first project aimed to find out the means of recognition specificity of the KS-mupA2 towards the predicted α-OH substituted substrate and the results suggest a specific recognition motif that may stall substrate progression until it has been hydroxylated. The second project aimed to investigate the significance of movements of loops on the MupH surface that were seen in the simulations presented in Chapter 4. Simulations suggested the loops may leave easy access to the active site until the substrate is bound.

Chapter 2

Materials and Methods

2.1 Databases

In the present study various databases have been used to retrieve sequence (DNA and Protein) and protein structure information. For the sequence analysis in Section 3.2.1 RefSeq microbial and UniProtKB/TrEMBL (with 6408654 seq and 20127441 seq respectively, as on 9^{th} March, 2012) were used to search for ACP homologues using the hidden Markov models. RefSeq is a non-redundant database of DNA, RNA and protein sequences curated by the sequences from International Nucleotide Sequence Database Collaboration (INSDC) hosted at the National Centre for Biotechnology Information (NCBI) . RefSeq (`http://www.ncbi.nlm.nih.gov/refseq/`) provides the reference sequence for a species, each entry being annotated and cross-linked to other databases by collaborating groups and NCBI staff. RefSeq can be accessed via the Entrez search engine, through a sequence similarity search using BLAST or by the RefSeq FTP server. The RefSeq microbial dataset was used, which consists of sequences from microbial genomes/proteomes.

UniProtKB/TrEMBL (`http://www.uniprot.org`) is a collection of computationally annotated protein sequences. The sequences in UniProtKB/TrEMBL are derived from the computational translation of the coding sequences (CDS) submitted in INSDC databases. UniProtKB/TrEMBL can be accessed through keyword or uniprot identifier searches on the UniProtKB website or through a sequence similarity search using BLAST. The purpose of selecting two

databases was not only to retrieve the highest quality sequence information from a curated database but also to have maximum coverage of sequences available through an automated database. The sequence data from both the databases were obtained from their FTP websites in the FASTA format.

For molecular modelling, as explained in Section 2.3.1, protein structure coordinate information was taken from the Protein Data Bank (PDB) (`http://www.rcsb.org`). The PDB is the single archive for macromolecule (DNA and Protein) 3D coordinate information. It consists of structural information determined through experimental methods such as NMR spectroscopy and X-ray crystallography. The PDB database can be searched by keyword or PDB identifier, or by running a BLAST similarity search either from within the PDB website or through NCBI BLAST.

2.2 Sequence analysis

In molecular biology, routine sequence analysis is carried out to infer the biological function of newly sequenced DNA or protein sequences. For the purpose of inferring biological function similarity searches against annotated sequence databases using pairwise alignment methods such as BLAST and FASTA are in common use. Pairwise sequence alignment methods give equal weighting to all aligned positions. However, in a multiple sequence alignment it can be clearly seen that several positions within a family of proteins are under more evolutionary pressure to be conserved than others. Thus pairwise sequence alignment methods are good for fast searches of closely related sequences but are poor at detecting remote homology. In contrast, methods based on position specific scoring and probability such as Profiles and hidden Markov models are much more sensitive in detecting remote homology.

2.2.1 PSI-BLAST and multiple sequence alignment

In the present study position specific iterative BLAST (PSI-BLAST) has been used to identify template structures for homology modelling (see Section 2.3.1). PSI-BLAST is an extension of the normal protein-protein BLAST but it uses a position specific scoring matrix (PSSM) or profile. This PSSM is generated from the multiple sequence alignment of the sequences found

in a normal protein BLAST, which is the first iteration of PSI-BLAST. The PSSM is used and refined in further iterations of PSI-BLAST, rather than a simple sequence. PSI-BLAST is more sensitive in detecting distantly related sequences as compared to the normal protein-protein BLAST.

In Section 3.2.1 multiple sequence alignments (MSAs) were used to identify conserved residues which differentiate branching-ACPs from non-branching-ACPs. An MSA is created with three or more sequences with the aim of identifying regions of sequence homology and conserved positions, which may be used to characterize a protein family. MSAs are also used for phylogenetic reconstructions as differences between sequences, can be used to estimate how recently two proteins from a given family a common ancestor in a family of protein/DNA sequences. As explained in Section 2.3.3.1 MSAs were also used for evolutionary trace analysis to predict positions under evolutionary pressure within a clade as compared to between the clades. The ClustalW and Muscle programs were used for generating MSAs.

2.2.2 Hidden Markov models (HMMs)

HMMs are models which represent a probability distribution of the occurrence of residues for each position in a multiple sequence alignment allowing for the residue type to occur through conservation, insertion, deletion, or some combination of the later two with the aim of encapsulating the common features of a protein family that are necessary to recognize other members of the family. HMMs are built on a set of aligned sequences belonging to a family of proteins. Each column in an MSA is considered as a “state”which “emits”symbols (residues) according to the emission probabilities. Each state is interconnected to other states called transitions with state transition probabilities. The model produces the sequence of symbols based on the emission and transition probabilities and the probability of a given sequence being produced is the product of all the emission and transition probabilities. The hidden part in a hidden Markov model is the sequence of the states taken to produce the sequence of symbols. Its only the emitted symbols which are visible, not the underlying state. In practice HMMs are generated (or trained) using an MSA. To align a sequence to an HMM model the HMM can be seen as travers-

ing all the possible sequences of states through the model to generate the target sequence. This will assign the probability values to the sequences produced of which the best ones are selected through a dynamic programming based algorithm such as the Viterbi algorithm.

2.2.2.1 HMM analysis of β-branching and standard ACPs

HMMs were used to classify a subclass of acyl carrier proteins (ACPs) called the branching ACPs (for results see Section 3.2.1). The two HMM models were built using the HMMER program from the 15 clusters with known pathways, with 38 and 178 sequences (sequences provided by Dr. Anthony Haines) for the branching and the non-branching-ACPs respectively. The training and the test set were scored with each HMM models and the scores were plotted on a scatter graph to highlight two distinct clusters between branching and non-branching ACPs. The non-branching HMM model was also searched against TrEMBL and RefSeq databases which fetched 16,490 unique sequences with the length greater or equal to 60 amino acids. These sequences were also scored with both the models and plotted in a separate scatter plot along with the training set. The models were deposited with SMART since no domain prediction software (SMART, Pfam etc.) had the capacity to distinguish these two types of ACP.

2.3 Molecular Modelling

The present study aims to model the polyketide synthase complexes which are large macromolecular complexes. Determining the structure of large protein complexes through experimental methods such as X-ray crystallography and NMR spectroscopy is possible however at times can be quite difficult and time consuming. On the other hand with the increase in computing capacity and the large amount of molecular biology data available it is increasingly possible and efficient to use computational methods to predict structures based on the existing structural data. Molecular modelling can be used for tasks such as protein structure prediction through homology modelling, understanding protein dynamics through molecular dynamics simulations, protein protein interaction interface prediction and molecular docking to generate atomic resolution three dimensional protein complex structures.

2.3.1 Homology modelling

Homology modelling or comparative modelling is a computational method to generate three dimensional protein structures of atomic resolution. As the name suggests homology modelling is based on the fact that the proteins which perform similar functions tend to have similar sequences with regions important for the function being conserved. Furthermore, the protein structures of even divergent protein sequences, which share a common ancestry, tend to fold into a similar three-dimensional structure. Thus a protein structure can be used to predict the atomic resolution model of a homologous target sequence (Chothia and Lesk 1986). The theoretical models generated through homology modelling primarily rely on the quality of the sequence alignment between the target sequence and the homologous sequence and the quality of the homologous structure. Over the past few years several state of art computational tools have emerged with increasing level of accuracy, which are continually being monitored by the biannual CASP competitions (`http://predictioncenter.org/`).

In the present work, different versions of the Modeller (Sali and Blundell 1993) program have been used to carry out the homology modelling for all the predicted structures. Modeller generates the three dimensional models of the target protein sequence by satisfying spatial restraints obtained from the sequence alignment of the target and the homologous protein templates. This method in principle is quite similar to employing the distance geometry approach used in NMR experiments. The spatial restraints extracted from the sequence alignment are expressed as probability density functions (PDFs) which are used to restrain Cα-Cα bond distance, N-O distance, main chain and side chain dihedral angles. Apart from the bond distance restraints derived from the sequence-template alignment, stereochemical restraints are added using a molecular mechanics force field. In the final model building step the model is generated by minimizing the violation of all the restraints (Martí-Renom *et al.* 2000; Eswar *et al.* 2006).

This method of model prediction has an advantage over many other tools for the same purpose as it allows several restraints to be added by the user derived from various resources. It is thus possible to apply restraints derived from experiments such as NMR spectroscopy, fluorescence resonance energy transfer (FRET), crosslinking experiments as well as restraints

derived from Bioinformatics based experiments. The model building step can also be coupled with model refinement steps using simulated annealing and molecular dynamics simulations. However, in the present work many of the models were generated without any model refinement step which allows a direct comparison with the template structure as the algorithm places the side chains at the similar positions to the template.

2.3.1.1 Modelling of MupH + ligand complex

MupH was modelled using Modeller version 9.8 (Eswar *et al.* 2006). The MupH sequence was used to search the PDB for the solved homologous structures, using PSI-BLAST. 10 structures were selected ranging from 27% to 32% sequence identity to generate the initial alignment using ClustalW (Larkin *et al.* 2007). The structures chosen in similarity with the MupH sequence were from the species *Staphylococcus auereus, Enterococcus faecalis, Brassica juncea, Streptococcus mutans* and *Homo sapiens* thus providing a wide range of organisms for comparisons. All the homologues found were HMG-CoA synthases. The secondary structures of the templates, determined by DSSP (Joosten *et al.* 2011), guided manual refinement of the final alignment using Seaview (Galtier *et al.* 1996). The HMG-CoA synthase structure from *Enterococcus faecalis* (PDB ID 1X9E), having 87% query coverage with 32% sequence identity, was selected as the template. Modeller produced five structures which were further tested for stereochemical quality using the PROCHECK (Laskowski *et al.* 1993) program (`http://nihserver.mbi.ucla.edu/SAVES/`). The model with the best PROCHECK score was selected for further analysis.

In order to dock the polyketide intermediate ligand in the MupH active site the modelled MupH structure was superimposed on the 1X9E crystal structure and the X-ray determined coordinates for the phosphopantetheine moiety from the bound ligand in 1X9E were copied. The rest of the polyketide intermediate was built manually inside the MupH active site using the PyMol program (`http://www.pymol.org`). To remove the steric clashes energy minimization was carried out using Chimera (Pettersen *et al.* 2004). The AMBER99SB-ILDN (Hornak *et al.* 2006) force field and Antechamber program were used to assign the force field parameters to the protein and ligand respectively.

2.3.1.2 Modelling of ACP-mupA2

ACP-mupA2 (the ACP from the second module of MmpA in the mupirocin cluster) was modelled using Modeller version 9.10. The ACP-mupA2 sequence (GenPept Acc No. AAM12909 range 1822 to 1916) was used to search against the PDB database using BLAST to identify the template. The NMR structure of the Holo-Acpi Domain from the CurA module from *Lyngbya Majuscula* (PDB ID 2LIU) was selected as the best template with 80% query coverage and 29% sequence identity. To generate the initial alignment for modelling, a multiple sequence alignment was carried out using two template sequences, 2LIU, 2L22 (ACP-mupA3ab NMR structure from the third module of MmpA in the mupirocin cluster) and other ACP sequences from the mupirocin cluster. Sequence alignment was manually refined using secondary structure information as explained in Section 2.3.1.1 and the homology modelling was carried out using 2LIU as the template, generating five structures. The structures were tested for stereochemical quality using PROCHECK and the structure with the best PROCHECK score was selected for further analysis.

2.3.1.3 Modelling of the KS-mupA2 dimer

KS-mupA2 (KS from the second module of MmpA in the mupirocin cluster) was modelled using Modeller version 9.11. KS-mupA2 was modelled as a dimer along with the KS docking domain also called as the linker region between the KS and AT domain in the *cis* systems. KS-mupA2 sequences (GenPept Acc No. AAM12909.2 range 715 to 715 to 1288) was searched against the PDB database using BLAST to identify the template. Two templates from the DEBS system, the KS-AT dimer from module three and five (PDB ID 2QO3 and 2HG4) along with the KS sequences from the mupirocin system, were used to generate the initial alignment which was manually refined using secondary structure information and visual inspection of the alignment. The DEBS structure 2QO3 with 88% query coverage and 39% sequence identity was used as the template to generate five homology models. The structures were tested for stereochemical quality using PROCHECK and the structure with the best PROCHECK score was selected for further analysis.

2.3.2 Molecular dynamic simulation

Molecular dynamic (MD) simulation is a computational method to simulate the time dependent behaviour of a physical system. In molecular modelling, MD simulations of proteins and DNA macromolecules are used to observe the changes in forces and structural dynamics with respect to time. In a typical MD simulation of a physical system of N interacting atoms, such as a protein immersed in a box of solvent, the atoms are allowed to move following Newton's laws of motion. Equation 2.1 represents the differential form of Newton's second law of motion i.e. F=ma, describing the motion of a particle i of mass m_i, in the direction x_i with the force F_{x_i} acting on the particle in the x_i direction.

$$\frac{\delta^2 x_i}{\delta t^2} = \frac{F_{x_i}}{m_i}, i = 1...N \tag{2.1}$$

The calculation of the forces highly depends on the parameters of the molecular mechanics force field equation used. The equation below is the functional form of the AMBER force field equation.

$$\begin{aligned} V(r^N) = &\sum k_b(l - l_0)^2 + \sum k_a(\theta - \theta_0)^2 + \sum \frac{1}{2} V_n[1 + \cos(nw - \gamma)] \\ &+ \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ \left[\epsilon_{ij} \left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \end{aligned} \tag{2.2}$$

The first term in the equation represents the energy of the covalent bonds. The second term represents the energy of the bond angles. The third term represents the energy due to torsional angles and the last term represents the energy due to non-bonded interaction between all the atom pairs, which include van der Waals, and electrostatic energies.

In molecular dynamics simulation the equation of Newton's second law of motion is solved for the desired length of time. The change in the coordinates of the atoms due to the motion is recorded at regular intervals. The coordinates with respect to time represents the trajectory of the system. Once the system has reached the equilibrium state many macroscopic properties can be studied by averaging over the equilibrium trajectory. MD simulations are widely used

in molecular modelling to study folding and unfolding of the proteins, protein structure stability, protein-protein interactions, free energy binding of drug-protein complexes, the effects of mutations on the structural dynamics of the proteins etc.

In the present study the GROMACS 4 (Pronk *et al.* 2013) molecular dynamics program has been used to study the protein dynamics in explicit solvent (water) and the protocol can be summarized in three major steps.

Structure preparation

In the first step, the physical system to be simulated is represented in the computer. The protein structure from X-ray or NMR experiments or through homology modelling contains the coordinate information only. However, for running the MD simulation this coordinate information needs to be expressed in terms of atomic coordinates, atom types, bonds, electric charge distribution etc. This information is stored in a topology file along with the necessary force field parameters from the selected force field and the water model to be used for the explicit solvent simulation. There are several popular force fields supported by GROMACS, which includes GROMOS (Oostenbrink *et al.* 2004), AMBER (Lindorff-Larsen *et al.* 2010), OPLSA (Jorgensen *et al.* 1996), and the CHARMM27 (MacKerell *et al.* 1998) force fields. During the structure preparation step hydrogen atoms (if any are present) are usually removed from the coordinate file and added back according to the specifications of the force field used. This rebuilding of hydrogen atoms may cause some steric clashes which need to be relaxed through energy minimization.

For simulations in explicit solvent usually a simulation box is created with periodic boundary conditions (PBC) which is filled with a suitable water model for example SPC, TIP3P (Spoel *et al.* 1998). There are a few limited number of shapes in which the simulation box can be created with a cube being the largest in terms of the volume and truncated dodecahedron being the smallest. A truncated dodecahedron surrounds spherical globular proteins better because it does not accumulate huge amount of solvent at the corners therefore it is computationally less expensive than e.g. a cube. The PBC means a single unit is repeated infinitely in order to overcome the edge effects caused by the walls of

the simulation box. For the Amber force field used in the present work the TIP3P water model is considered a compatible model and was used here. At this point once the solvation box is created and filled with the suitable solvent the net charge of the system is neutralized by adding counter ions. The energy of the system is minimized either using a steepest descent algorithm, conjugate gradient algorithm or both to remove any steric clashes (Hess *et al.* 2008).

Equilibration and production run

In the second step, the system is equilibrated by running short MD simulations in which the protein is restrained to the reference position while keeping the water molecules flexible. This allows the water molecule to relax around the protein. It is a general practice to first run a short simulation with only temperature coupling to equilibrate the system at the desired temperature for example at 300K. Once the system is equilibrated at the desired temperature another short simulation can be run with the pressure coupling as well to equilibrate the system under both the temperature and pressure conditions. These equilibration steps are usually performed for few hundred pico seconds depending upon the size of the system.

In the final production run, MD simulation is carried out without any restrains on the protein under the influence of temperature and pressure, which can be of several nanoseconds long depending upon the macroscopic property under study and the computational power available (Hess *et al.* 2008).

2.3.2.1 Parameter determination

As mentioned above, the forces in the simulation depend on the parameters used. These parameters include atomic properties such as mass, partial charge and atom type, bond properties such as equilibrium bond length, equilibrium bond angle and dihedral angle preferences and parameters representing the partial charges and van der Waals interaction. Since, forcefields in GROMACS are optimized for protein and nucleic acid simulation, parameters are given only for the standard amino acids, nucleotides and a few counter ions. However, GROMACS allows

the addition of extra parameters to represent systems which consist of atoms not represented by the standard set of parameters. In order to include a molecule which is not a part of the standard set, different methods based on either experimental data or quantum mechanics calculations can be used to determine parameters. While working with the AMBER force fields small molecules can be simulated using parameters from the general amber force field (GAFF). GAFF was created to be compatible with the AMBER force fields (Wang *et al.* 2004) and is part of the AMBER molecular dynamics package which is not included in GROMACS by default.

In the present study at several instances a ligand molecule was covalently attached to a residue in the protein, for example a phosphopantetheine arm attached to the active site serine in an ACP. To carry out molecular dynamics simulation with the phosphopantetheine and other molecules attached to the serine on ACP a new residue type was created for each serine ligated with ligand. Since, these new residues were not a part of the standard amino acid set, GROMACS did not have the parameters to carry out the simulation. For this reason parameters from the GAFF were added to the AMBER 99SB-ILDN (Lindorff-Larsen *et al.* 2010) force field. The GAFF parameters were obtained from the AMBER molecular dynamics package and the values which differed in units from GROMACS format where converted using several in house perl scripts (see Appendix I A.4).

While using GAFF in AMBER molecular dynamics package, partial atomic charges (AM1-BCC) are assigned using antechamber software included in the package. However, in GROMACS suite there was no such software available which can assign partial atomic charges to small molecule. Therefore, in order to derive partial atomic charges for the new residues, AMBER force field compatible RESP charges were calculated using RED server (`http://q4md-forcefieldtools.org/REDS/` (Bayly *et al.* 1993; Dupradeau *et al.* 2010; Vanquelef *et al.* 2011)). For the calculation of RESP charges the RED server requires the structure file in a P2N format. The program Ante RED, available in the RED server, converts a PDB file into the P2N format. The RED IV server was used to calculate RESP-A1A charges for all the ligands, using the Gaussian 2009 D.01 quantum mechanics program. Fully automated mode 1 was chosen while running RED IV which performs geometry optimisation as well as charge

fitting. Once the charges were calculated a new entry was created for each new residue type in the *aminoacids.rtp* database of GROMACS. The new residue name was also updated in the *residuetype.dat* database. For each new residue the corresponding hydrogen atoms were also created in the *aminoacids.hdb* database. While determining charges for the new moiety attached to an existing amino acid, the charges for the serine side chain were varied but the backbone atoms of the parent residue were kept the same as that of the original forcefield. Section B.1 in Appendix II lists the ligands with the charges which were added to the AMBER99SB-ILDN *aminoacids.rpt* database.

2.3.2.2 Molecular dynamics simulation of W44L mutant and wild type ACP-mupA3a

A) The mutation from W to L at the 44^{th} position in the ACP-mupA3a NMR structure was carried out using PyMol. The two constructs (wild type and the mutant) were subjected to 20 independent simulations for 10 ns each. The GROMACS 4 (Hess *et al.* 2008) molecular dynamics engine with AMBER99SB-ILDN (Lindorff-Larsen *et al.* 2010) force field and TIP3P water model was used to carry out the simulation with a time step of 2fs and in a water box extending 5 Å beyond the surface of the protein. The simulation systems were each neutralized with 2 Na^+ counter ions, minimized for 1000 steps of conjugate gradient and 1000 steps of steepest descent energy minimisation and were equilibrated for 100 ps. The final simulation run was performed for 10 ns at 300K temperature and 1 atm pressure with other parameters set as the GROMACS 4 defaults.

B) Another set of two independent simulations were setup for the W44L mutant and wild type ACP-mupA3a for 1 microsecond each. The simulation systems were neutralized by adding 2 Na^+ counter ions and the protocol and the setup was same as described in Section 2.3.2.3.

2.3.2.3 Molecular dynamics simulation of ACP-mupA3a with covalently bound phosphopantetheine

The ACP-mupA3a NMR structure was simulated with the phosphopantetheine covalently attached to S38 of both the wild type ACP-mupA3a and W44L ACP-mupA3a mutant. In order to simulate phosphopantetheine covalently attached to S38 a new residue named SPT was intro-

duced in the GROMACS *aminoacids.rtp* database (Appendix II Section B.1.1). The structure for SPT was drawn using Pymol by extending S38 coordinates from the NMR structure (PDB ID 2L22). Parameters for the serine backbone atoms were kept the same as that of the original AMBER99SB-ILDN forcefield and for the rest of the ligand (from phosphate onwards) were derived from GAFF. Charges for the phosphopantetheine moiety were determined using the RED server as described in Section 2.3.2.1. The structure file used for charge calculation consisted of a phosphopantetheine moiety attached to the serine side chain oxygen with all the atoms of the serine present and the backbone carbonyl carbon and amide nitrogen blocked with methyl groups. After charge fitting with RED server, the charges on the serine backbone were set to the same values ass the AMBER99SB-ILDN force field, the blocking methyl groups were removed, and the remaining charges were adjusted in the third decimal places to keep the overall charge of the molecule as -1 (Appendix II Section B.1.1).

The two constructs (wild type and the mutant) were subjected to 5 independent simulations for 50 ns each. Simulations were carried out using TIP3P water with a time step of 2fs and placed in an octahedron water box extending 10 Å beyond the surface of the protein. Each simulation system was neutralized with 3 Na^+ counter ions and minimized for 1000 steps of conjugate gradient and 1000 steps of steepest descent energy minimisation. Following the energy minimisation each system was equilibrated with 100 ps of NVT and 100 ps of NPT simulations keeping the position of the protein atoms restrained. The final simulation was performed for 50 ns at 300K temperature and 1 atm pressure using V-rescale thermostat (Berendsen *et al.* 1984) and Parrinello-Rahman pressure coupling (Parrinello *et al.* 1980) respectively. Out of the five simulations per construct one was extended for to 200 ns for both wild type and mutant ACP-mupA3a.

2.3.2.4 Molecular dynamics simulation of ACP-mupA3a with its substrate

ACP-mupA3a wild type and W44L ACP-mupA3a mutant were simulated with their substrate prior to β-branching (Figure 1.27). The ACP-mupA3a substrate was built directly onto the phosphopantetheine of Section 2.3.2.3 by using PyMol. All the bond parameters and charges for the phosphopantetheine moiety were kept the same as the previous calculations. Param-

eters for the ACP-mupA3a substrate were obtained from GAFF and charges were calculated using the RED server. The charge calculations were performed on the ACP-mupA3a substrate attached to a sulphur and peptide moiety, representing the terminal atoms of the phosphopantetheine arm, the peptide being blocked by the addition of a methyl group. Including the entire phosphopantetheine in charge calculations was not preferred as it showed to bias the geometry optimization in a U shaped conformation rather than an extended one and these interactions would polarize the charges. HG-63G* charges already over polarize the charges compared to gas phase, to mimic the effect of intermolecular interactions (Winn *et al.* 1997) Therefore, the charges were calculated separately and then merged with the previously calculated charges for the phosphopantetheine, keeping the overall charge on the molecule as -1. The new residue, SPM, was added into the AMBER99SB-ILDN *amoniacids.rtp* database in GROMACS (Appendix II Section B.1.2). The simulation system was neutralized by adding 3 Na^+ counter ions and five independent simulations were carried out for 50 ns each following the same protocol as described in Section 2.3.2.3. One of the simulations was extended to 1 μs for the wild type ACP-mupA3a and upto 200 ns for the W44L ACP-mupA3a mutant.

2.3.2.5 Molecular dynamics simulation of ACP-mupA3a with a covalently bound saturated carbon chain

ACP-mupA3a wild type was simulated with a covalently bound 14C saturated chain to the phosphopantetheine. The carbon chain was of the same length as that of the ACP-mupA3a substrate's backbone. The parameters and the charges for the phosphopantetheine moiety were the same as in the previous calculations. The parameters for the 14C saturated carbon chain was obtained from GAFF and the charges were calculated using RED server. The new residue SPD was added into the AMBER99SB-ILDN *amoniacids.rtp* database in GROMACS (Appendix II Section B.1.3). The simulation system was neutralized by adding 3 Na^+ counter ions and three independent simulations were carried out for 50 ns each following the same protocol as described in Section 2.3.2.3. One of the simulations was extended upto 200 ns.

2.3.2.6 Molecular dynamics simulation of ACP-mupA2a with its substrate

ACP-mupA2a wild type was simulated with its substrate (Figure 1.27). All the parameters for the phosphopantetheine moiety were the same as in the previous calculations. Parameters for the ACP-mupA2 substrate were obtained from GAFF with charges calculated using the RED server. The new residue SPB was added into the AMBER99SB-ILDN *amoniacids.rtp* database in GROMACS (Appendix II Section B.1.4). The molecular dynamics simulation procedure was as in the Section 2.3.2.3. The simulation system was neutralized by adding 5 Na^+ counter ions and three independent simulations were carried out for 50 ns each following the same protocol as described in Section 2.3.2.3. One of the simulations was extended to 200 ns.

2.3.2.7 Molecular dynamics simulations of ACP-mupA3a:MupH complex

The starting configuration for MD simulation of the ACP-mupA3a:MupH complex was a representative complex from the *in silico* docking described in Section 2.3.4.1 and Chapter 3 Section 3.2.6. ACP-mupA3a had its substrate attached to the phosphopantetheine. An acetyl molecule was covalently attached to the C115 in the MupH active site. The force field parameters for the phosphopantetheine and ACP-mupA3a substrate were taken from the previous calculations as described in section 2.3.2.4. Parameters for the acetyl molecule attached to the C115 were obtained from GAFF with the charges calculated by using the RED server. A new residue type CYA for C115Acetylated-cysteine was introduced into the AMBER99SB-ILDN *amoniacids.rtp* database in GROMACS (Appendix II B.1.5). The simulation system was neutralized by adding 16 Na^+ counter ions and three independent simulations were carried out for 50 ns each following the same protocol as described in Section 2.3.2.3. One of the simulations was extended upto 100 ns.

Another set of simulations were carried out for ACP-mupA3a:MupH complex but with MupH as a dimer. All the parameters were kept the same as that for the ACP-mupA3a:MupH monomer complex. However, new topologies were generated including the MupH dimer. The ACP-mupA3a:MupH dimer system was neutralize by adding 29 Na^+ counter ions. Three independent simulations were carried out for 50 ns each following the same protocol as described

in Section 2.3.2.3.

2.3.2.8 Molecular dynamics simulations of MupH and MupH acetylated at C115 in isolation

Three independent simulations for MupH and MupH acetylated at C115 were carried out for 50 ns each. The force field parameters for the acetyl molecule attached to the C115 were from previous calculations (as described in Section 2.3.2.7). The simulation systems for both the structures were neutralized with 13 Na^+ counter ions and were carried out following the same protocol as described in Section 2.3.2.3.

2.3.2.9 Calculating the RMSD and RMSF of ACP-mupA3a and ACP-mupA2 from their reference starting structure

Root mean square deviations (RMSD) of the backbone atoms from the reference starting structure were calculated for all the extended simulations of ACP-mupA3a and ACP-mupA2 using the GROMACS module *g_rms*. Below is an example command to run *g_rms*, which requires a MD simulation input file and a MD simulation output trajectory file. The RMSD was calculated at every 10 ps of the simulation. The RMSD plots were visualized using the program Grace (`http://plasma-gate.weizmann.ac.il/Grace/`).

```
g_rms -s md.tpr -f md.xtc -o rmsd.xvg
```

Root mean square fluctuations (RMSF) were calculated for the backbone atoms, averaged per residue, for all the extended simulations of ACP-mupA3a and ACP-mupA2 using the GROMACS module *g_rmsf*. Below is an example command to run *g_rmsf*, which requires a MD simulation input file and a MD simulation output trajectory file. The RMSF was calculated from points in the trajectory at 10 ps intervals. A combined scatter graph was plotted for all the simulations using a spread sheet program.

```
g_rmsf -s md.tpr -f md.xtc -o rmsf.xvg -res
```

2.3.2.10 Calculating RMSD of the ACP-mupA3a and ACP-mupA2 from the reference FAS ACP structure

The GROMACS suite has a module to calculate RMSD of a set of atoms from a reference structure, but it does not perform well with a reference structure which is not identical to the structure in the simulation. The task in this analysis was to calculate the RMSD of the conformational states of the ACP-mupA3a and ACP-mupA2 (wild and mutant) structures at every ns of the simulation from the reference FAS ACP structure (PDB ID 1L0I). To perform this task not natively supported by GROMACS, conformational states at every ns of the simulation were extracted using VMD (Visual Molecular Dynamics) (Humphrey *et al.* 1996). I wrote a Perl script (Appendix I Section A.5) to extract the individual conformational states from a multiple structure concatenated file and aligned each of them with the reference FAS ACP structure using the Matt structural alignment program (Menke *et al.* 2008). My script also extracts the RMSD value from each alignment result file and outputs the results on the screen with the time frame of the simulation. A scatter graph was plotted of the RMSD values against time for each simulation system using a spread sheet program.

2.3.2.11 Calculating cavity volume during the course of ACP-mupA3a/mupA2 simulations

To detect the formation and change in the volume of the proposed cavity in the ACP-mupA3a / mupA2 structure during the course of molecular dynamics simulations a third party GROMACS plugin *trj_cavity* was used (Paramo *et al.* 2014). It takes a GROMACS topology and trajectory as the input and an optional seed value in the form of Cartesian coordinates to initiate the cavity detection at a particular position on the protein. Here, the seed value was calculated using the PASS program from the MetaPocket server (Huang 2009). Another important parameter is a dimension value which is 5 by default which means that the algorithm requires void space in 5 of the 6 possible directions (that is +/- x, y, z coordinates) till it encounters protein atoms. The program gives the flexibility of choosing the grid size in Åwhich is 1.4 by default, to represent the size of a water molecule, and the atomic radii. In this analysis a gird size of 1.3 and the

atomic radii from the AMBER99SB-ILDN force field were used. The grid size of 1.3 was decided after the observation that with a grid size of 1.4 there was no cavity detected in many of the frames where a human would deem there to be one. As the proposed cavity is highly surface exposed the void spaces were not deep enough to find protein in all the five directions. The grid size of 1.3 did not eliminate the possibility of no cavity detection but upon visual inspection of the trajectories there were fewer frames found with no cavity mapped. Upon decreasing the dimension size to 4 and keeping the grid size to default 1.4 more frames were detected with cavities but there were more instances of spill overs. Here, spill overs mean that the probe couldn't find protein in five directions and hence ran off the intended cavity space until it finds the edge. This could be controlled with a cutoff distance, various cutoff distances were tried but there wasn't much improvement. Upon decreasing the grid size lower than 1.3 lead the probe go through the protein interior thus mapping the whole or the majority of the structure as a potential cavity. Grid sizes lower than 1.3 with an increased atomic radii were also tried but again there wasn't much improvement and at the same time the system seem to be more unrealistic. Therefore, after comparing different parameter sets the final set of parameters were decided to be dimension of 5, 1.3 grid spacing, 9 cutoff distance, AMBER99SB-ILDN atomic radii and a corresponding seed value for different structures. The command below shows an example of the parameters assigned. A scatter graph was plotted of the cavity volumes against time for each simulation system using a spread sheet program.

```
trj_cavity -s em.gro -f md_cat-fit.xtc -seed 46.804 50.162 30.501 -o
    cavity_max_dim5_ff1.3.pdb -ov volume_max_dim5_ff1.3.xvg -mode max -dim 5
    -ff_path amber99sb-ildn.ff/ -ff_radius -spacing 1.3 -cutoff 9
```

2.3.2.12 Calculating hydrogen bonds and solvent accessible surface area (SASA) during the course of ACP-mupA3a and ACP-mupA2 simulations

To detect and quantify the interaction of the phosphopantetheine and the acyl chain with the protein surface and the solvent, the number of hydrogen bonds and solvent accessible surface area (SASA) were calculated. Two built in modules in the GROMACS suite *g_hbond* and *g_sas*

were used for calculating the hydrogen bonds and SASA respectively. New GROMACS index groups were made for the phosphopantetheine and acyl chains to represent them as different moieties, from the atoms starting from phosphate till the sulphur as the phosphopantetheine and from the first carbon after the sulphur till the terminal carbon as the acyl chains respectively. The hydrogen bonds were calculated during the course of simulation at 10 ps intervals between each index group and the protein as well as with the solvent for both the phosphopantetheine and the acyl chains (excluding phosphopantetheine). Separate scatter graphs were plotted of the number hydrogen bonds detected and the SASA values against time for each simulation system using a spread sheet program. Below are the example commands to calculate hydrogen bonds and SASA using g_hbond and g_sas module respectively.

```
g_hbond -f md_1-cat.xtc -s em.tpr -n index.ndx -num sol_hbnum.xvg -dist
    sol_hbdist.xvg
g_sas -s em.tpr -f md_3-cat.xtc -n index.ndx -o sas_area.xvg -tv sas_volume.xvg
```

2.3.2.13 Calculating the distance between the two loops on the MupH surface

To detect the loop movement at the MupH cavity during the simulation g_dist module of the GROMACS suit was used. The distance between the CA atom of three residues on the loop I (L150, M151 and I152) were measured from each of the three residues on the loop II (P208, D208 and S209), at every 5 ns of the simulation. Six index groups were created for the three residues on the loop I and the three on the loop II to differentiate them as different entities for the distance measurement. In case of the MupH dimer complex the residues were selected on chain A. The distance between one residue on the loop I with each of the three residues on loop II was represented as an average. A scatter graph was plotted for the three distance averages calculated between the loop I and loop II for every simulation. Below is an example command to calculate the distance between two atoms or groups in GROMACS using g_dist module.

```
g_dist -f md.xtc -s md.tpr -n loopres.ndx -o dist_0_P207_L150.xvg
```

2.3.3 Interface prediction

At present the available protein interface prediction methods can be divided into two classes. The first class of methods predicts the residues most likely to be at an interface using multiple sequence alignment. Since there is evolutionary pressure for conservation of the interface, for example the Evolutionary Trace method (ET) (Lichtarge *et al.* 1996), which is ideally mapped onto the surface of the protein structure. The second class of the methods do not rely on any evolutionary information and are solely based on geometrical, physiochemical and statistical properties of the surface of the protein structure, for example the PIER (Protein Interface Recognition for Structural Proteomics) method (Kufareva *et al.* 2007). In the present work both ET and PIER have been used to predict the most likely interacting interfaces, which were either used to drive the docking experiments and/or to validate the docking results (Section 3.2.6.2). Predicting interacting interfaces can help to reduce the sampling required for docking experiments as the surface areas that are least likely to be at the interface can be excluded from the analysis, although here they were used to support results obtained by docking calculations that did not incorporate those predictions.

2.3.3.1 Evolutionary trace (ET)

Evolutionary trace (ET) analysis ranks the positions in a multiple sequence alignment of a family of protein sequences according to their evolutionary importance. ET analysis is based on two hypotheses. First, in an evolutionary related family of proteins the functionally important residues, such as in active sites and at protein-protein interaction interfaces, are conserved. Second, some mutations may lead to new functions, or functional variants and sequences showing these features may form a clade with the residue associated with the subtype of function under pressure to be conserved (Lichtarge *et al.* 1996).

To generate an evolutionary trace from a multiple sequence alignment the first step is to generate a the phylogenetic tree and partition it into different clades based on sequence identity. On a dendogram a clade is defined as a set of sequences arising from the same node. This partitioning of the clusters is carried out by the partition identity cutoff (PIC) (Du and Alkorta

1994). Once each cluster is identified through partitioning, a consensus sequence is derived for each and every clade. This consensus sequence represents a residue type for the conserved positions or a blank (neutral) for positions which are not conserved within a clade. In the next step these consensus sequences are aligned to determine residues conserved within clades but which might vary between clades, and this is the ET. If any of the consensus sequences contain a neutral or a gap position then that position is considered to be neutral in the trace. Finally the functionally important residues obtained from the trace are mapped onto the structure.

Partitioning plays an important role in the ET analysis. A few large clusters are obtained at a lower PIC value which could be good enough to highlight broad functional features of a protein family. However, to highlight finer functionality within a subgroup higher PIC values are needed. The number of sequences in the alignment is also an important criterion for a successful trace. A high PIC value will generate a large number of clusters with a single sequence which are excluded from the analysis. When there are very few sequences in a cluster these may not represent the true variation within the group and many positions that are in reality neutral will be highlighted as functionally important.

Oliver Lichtarge and co-workers (Lichtarge *et al.* 1996) have shown that the evolutionary trace identifies ligand binding sites in the three protein domains SH2, SH3 and the DNA binding domains of nuclear hormone receptor. The results obtained at lower and higher PIC values show agreement with experimental observations.

As explained earlier the consensus sequence derived for each and every clade represents alignment positions which are conserved within a clade and if these are conserved within each clade, but not necessarily between them, ET ranks them as important. However, this procedure suffers from the limitation that the positions which are not considered to be conserved and hence not included in the ET ranking, may also be important, on the basis that they might not be conserved at that position in only one sequence or clade. This lack of conservation may arise for example due to a sequencing error, even if its not a sequencing error throwing away the whole column because one sequence did not have the identical residue might lose important information. To over come this limitation and make the ET more robust, a hybrid

method called real value evolutionary trace (rvET) was developed (Mihalek *et al.* 2004). Real value evolutionary trace is a combination of ET with information entropy. Information entropy allows the ranking of the importance of a position by counting the frequency of different residue types in a column. Hence by combining the ET with information entropy those columns which lack absolute consensus but are highly conserved may also be ranked as important, the rvET score is lower for columns with partial conservation in a clade as compared to the columns in which a particular residue is absolutely conserved in one clade and differ in another clade.

Real value evolutionary trace analysis can be performed using ET viewer (Morgan *et al.* 2006) which is an easy to use graphical user interface java applet. The applet runs the trace remotely on a server in the Lichtarge group and requires a reference protein structure and a sequence alignment. Both the structure and the sequence alignment can be uploaded by the user manually or be obtained automatically by the server from the PDB website and through an automated blast search respectively. In the present work both the structure and the alignment were uploaded manually and the rvET was carried out with default parameters.

2.3.3.2 Protein Interface Recognition for Structural Proteomics (PIER)

PIER (`http://abagyan.ucsd.edu/PIER/`) is a computational method to predict the interface residues based on the local statistical properties of the protein surface at the atomic group level (Kufareva *et al.* 2007). PIER predictions do not analyse sequence conservation information although in the original publication of the method evolutionary information was supplemented but this produced marginal or no additional improvement in the predictions. The PIER prediction model is based on the data obtained from 748 protein complexes available in the PDB at the time of its publication. This diverse set consists of 490 homodimeric, 62 heterodimeric, and 196 transient interfaces. In order to validate the model Kufareva *et al.* (2007) randomly divided the dataset of 748 complexes into three equally distributed sets. Each set was used to train a model and which was tested on the other two sets. PIER has shown a precision (positive predictive value) of 60% at the recall (sensitivity) rate of 50% in identifying the interface residues. Where precision can be defined as the ratio of the total number of true positives (TP) to the size of the predicted patch (TP+FP), FP means false positive, and the recall rate is the ratio of total num-

ber of true positives (TP) to the size of the true interface (TP+FN), FN means false negative. Through the statistical analysis of the data set, Kufareva *et al.* (2007) identified 12 significant atom groups (Table 1 in the original paper (Kufareva *et al.* 2007)) with a high probability of being found at an interface. The analysis revealed that stable permanent complexes are mostly formed by side chain interactions. Whereas in transient complexes backbone - backbone and backbone - side chain interactions are dominant.

The PIER algorithm predicts interface residues in a four step process. *a*) It generates the surface patches on the given protein based on solvent accessible surface area. *b*) For each patch it assigns 12 physical descriptors for the 12 significant atomic groups based on the total solvent accessible area of all atoms of a group. *c*) It calculates the PIER value for each patch. *d*) Transfer the calculated patch PIER value to the residues. The PIER value represents the chances of a residue to be on the interface, higher the score for a residue more likely it to be at the interface. In the present study the PIER values are mapped onto the structures using the b-factor column of the PDB file. The structures are coloured with the rainbow colouring scheme. Where the residues coloured as red are most relevant and violet as the least.

2.3.4 Molecular docking

In silico docking is a computational method to predict the most likely stable conformation of two molecules with respect to each other. Molecular docking can be performed to predict atomic resolution models of protein small molecule, protein-DNA, protein-RNA and protein-protein complexes. Broadly, for any of the above mentioned purposes, docking can be of two types. First, rigid body docking in which the two interacting molecules are considered as solid structures and the resulting complex is a function of shape complementarity between the two monomers. The second approach called the flexible docking allows for certain or all the parts of the interacting molecules to be flexible and dynamic. Both approaches have their own merits and demerits, rigid body docking is faster than flexible body and can be applied to larger complexes however; flexible docking is more accurate but can be quite computationally expensive for large molecules (Teague 2003). Both processes require a search algorithm which sets the

search space for all the possible molecular orientations and an energy term which calculates the intermolecular energy of the pair, this energy term is usually based on a molecular mechanics force field. A clustering algorithm can also be used to rank and cluster the complexes into clusters of similar orientation and comparable energies. Both rigid body and flexible docking can be of a global or local type. In global docking one molecule is kept stationary while the other molecule is allowed to roll over the entire surface of the stationary molecule to find the best match. The global docking algorithms do not require any prior knowledge of the interacting interface or residues for example the ZDOCK server (`http://zdock.umassmed.edu/`) (Pierce *et al.* 2014). Whereas, in local docking the docking interface or patch is roughly predefined and the docking algorithm tries to optimize the match for example the HADDOCK server (`http://haddock.science.uu.nl/`) (Vries *et al.* 2010). This predefined patch would usually be residues identified to be on the interface through various protein protein interaction studies, mutation experiments or Bioinformatics analysis such as ET or PIER.

More recent docking software also supplements the scoring procedure with biochemical and/or biophysical data such as chemical shift perturbation data obtained from NMR experiments, mutagenesis data or Bioinformatics analysis. HADDOCK (High Ambiguity Driven Docking) (Dominguez *et al.* 2003) is one such software which carries out the docking by expressing the interface information as ambiguous interaction restraints (AIRs). An AIR can be defined as the ambiguous distance between all the interacting residues. HADDOCK has continuously been ranked as one of the best software available for docking in the CAPRI competitions (`http://www.ebi.ac.uk/msd-srv/capri/`). In practice the HADDOCK software can be accessed through an easy to use webserver (Vries *et al.* 2010) (`http://haddock.science.uu.nl/`) which gives access ranging from easy to guru level interface. The latest version of HADDOCK can also perform multi body interface docking which can be used to dock upto six proteins structures (Karaca *et al.* 2010).

HADDOCK requires a list of "active" and "passive" residues which are likely to be on the interface. For example the active residues can be the ones which show significant chemical shift perturbation upon complex formation in an NMR experiment and also have a high solvent

exposure in the free form of protein. Whereas, the passive residues are the one which surrounds the active residues and have a less significant chemical shift but still with high solvent exposure. The docking is driven by assigning an AIR between any atom in active residue of one protein to the atoms in the active and passive residues of the other protein with the aim to maximize the interaction. The AIRs are defined as an ambiguous intermolecular distance with a maximum value of 3 Å. Apart from assigning active and passive residues to generate AIRs HADDOCK can also utilize unambiguous distance restraints within the same structure or in between the two structures.

The overall docking protocol consists of three stages. In the first stage the orientation of the two molecules is randomized followed by docking by rigid body energy minimization. This step is faster and generates approximately 1000 docking solutions of which the top 200 complexes in terms of intermolecular energies are further refined. The second stage is a three step semi rigid simulated annealing refinement. In the first step the two proteins are considered as rigid bodies, in the second step the side chains at the interface are allowed to be flexible and in the third step both the side chain and the backbone is allowed to be flexible. The resulting complexes are then subjected to the steepest descent energy minimization. The final stage of the docking is carried out with refinement at the interface in Cartesian space using molecular dynamic simulation in TIP3 explicit solvent with OPLS all atom force field. The final complexes are clustered using pairwise backbone RMSD where a cluster consists of at least two complexes with the RMSD smaller than 1 Å at the interface.

2.3.4.1 ACP-mupA3a/b + MupH docking

ACP-mupA3a was docked using HADDOCK server in two different ways to a MupH structure modelled as described in Section 2.3.1.1. The Easy interface was used for docking with a list of active and passive residues. The active and passive residues (explained in Section 2.3.4) were determined using the PIER program (as described in Section 2.3.3.2). The list of active and passive residues obtained from PIER are mentioned in the results section of Chapter 3, in Table 3.4.

In the second method the expert interface was used to dock ACP-mupA3a/b with MupH

using distance restraints. A distance restraint of 2.0 Å was used between the phosphorous of phosphopantetheine bound in the active site of MupH and the Oγ of the serine (S38/142) residue of ACP-mupA3a or mupA3b. An additional restrain of 9.13 Å (Steussy *et al.* 2005) was placed between the sulphur of the thioester linkage in the ligand and the Cα of the catalytic cysteine (C115) of MupH.

2.3.4.2 Docking the natural substrate into KS-mupA2 dimer

The natural substrate of the KS-mupA2 was docked into the modelled KS-mupA2 homodimeric structure (modelling described in Section 2.3.1.3) using the HADDOCK server Multi-body interface, with distance restraints. Docking was performed to mimic the decarboxylation stage of the Claisen condensation (see Section 1.2.4.3 for KS reaction mechanism). A distance restraint of 1.8 Å between the SG of the catalytic cysteine C158 and the first carbon of the substrate was used to mimic the acyl chain transfer step. To mimic the oxyanion hole formation, the backbone NH atoms of residues C158 and A403 were used to restrain the first carbonyl of the substrate with a distance restraint of 2.93 Å each. To mimic the decarboxylation step, atom HD1 of H293 and HE2 of H333 were used to restrain the carbonyl of the malonate and atom OZ of F219 was used to restrain the carboxylate of the malonate with a distance restraint of 3.0 Å each. To keep the phosphopantetheine attached to the malonate in a fully extended state a modelled ACP-mupA2 structure was also docked with a distance restraint of 2.0 Å between the atom OG of S38 on the ACP-mupA2 and the phosphate of the phosphopantetheine.

2.4 Kalimantacin ACP swap experiment

2.4.1 Bacterial strains and Plasmids

The wild type mupirocin producer *Pseudomonas fluorescens* NCIMB 10586 strain was used as a control. A *Pseudomonas fluorescens* Δ*acp4* strain, which was deleted for the *acp4* (ACP-mupA3b) was used as the starting point for ACP swap experiments since it carries just the ACP-mupA3a in the 3^{rd} module of MmpA. A *Pseudomonas fluorescens* Δ*mupH* strain, which has a deletion in *mupH*, was used as the host strain for swapping the ACP-mupA3ab with the

ACP-k24a from the kalimantacin cluster. *Escherichia coli* DH5α was used as the host for all cloning experiments, *Escherichia coli* S17-1 was used for the conjugal transfer of plasmids to *Pseudomonas fluorescens*. *Bacillus subtilis* 1064 was used as the detection organism for the overlay bioassay. Plasmid pAKE604 was used to clone the kalimantacin ACP together with the ACP-mupA3ab flanking regions from the *mup* cluster. Previously-prepared pJH10 derivatives were used to express *mupH*, *batC* and *batC* L218M mutant *in trans*. Table 2.1 lists the bacterial strains and plasmids used in this study along with their genotype.

Table 2.1: *Bacterial strains and plasmids used in this study*

Bacterial strain	**Genotype and properties**	**Reference**
Pseudomonas fluorescens NCIMB 10586	Wild type (WT) strain which produces mupirocin	G. T. Banks
Pseudomonas fluorescens Δ*acp4*	*Pseudomonas fluorescens* NCIMB 10586 with *acp-mupA3b (acp4)* deleted	(Rahman *et al.* 2005)
Pseudomonas fluorescens Δ*mupH*	*Pseudomonas fluorescens* NCIMB 10586 with *mupH* deleted	Wu *et al.* (2007)
Escherichia coli DH5	Φ80lacZΔM15, *recA1*, *endA1*, *gyrA86*, *thi-1*, *HsdR17*(r_k^-, m_k^+), *supE44*, *rrelA1*, *deoR*, Δ*(lacZYA-ArgF)U169*	Gibco BRL
Escherichia coli S17-1	*RecA pro hsdR RP4-2-Tc::Mu-km::TAN7*	Simon *et al.* (1983)
Escherichia coli S17-1 + *mupH*	*E. coli* S17-1 with *mupH* in pJH10 plasmid	Haines *et al.* (2013)
Escherichia coli S17-1 + *batC*	*E. coli* S17-1 with *batC* in pJH10 plasmid	Haines *et al.* (2013)
Escherichia coli S17-1 + *batC* L>M	*E. coli* S17-1 with *batC* L>M in pJH10 plasmid	Haines *et al.* (2013)
Bacillus subtilis 1064	*trpC2amyE::(spec* P_{xyl}*-gfp-lacI) chr::pSG1196(rmD0lacO cat)*	Moir *et al.* (1979)
Plasmid		
pAKE604	Size 7.2 kb, pMB1 replicon, Amp^R, Kan^R, *oriT*, *lacZα*, *sacB*	El-Sayed *et al.* (2001)
pJH10	Size 14.5 kb, IncQ. pOLE1 IncC1 deleted. Tet^R (from pDM1.2) *oriT*	El-Sayed *et al.* (2001)

2.4.2 Cell culture media and growth conditions

Four types of culture media were used in this project:

L-broth which contains yeast extract (10 g/l), tryptone (5 g/l), NaCl (10 g/l) and glucose (1 g/l);

L-agar which contains all the ingredients from the L-broth along with agar (15 g/l);

M9-Minimal medium which contains (per liter) 1.5 % w/v water agar (400 ml), 2XM9 salts (400 ml), 1M thiamine (800 μl), 1M $MgSo_4$ (800 μl), 1M $CaCl_2$ (800 μl) and 40% glucose (4 ml);

Mupirocin secondary stage medium (SSM), for HPLC, which consists of (per litre) soy flour (25 g), spray dried corn liquor (2.5 g), $(NH_4)4SO_4$ (5 g), $MgSO_4.7H_2O$ (0.5 g), Na_2HPO_4 (1 g), KH_2PO_4 (1.5 g), KCl (1 g), $CaCO_3$ (6.25 g), glucose (4 %). The SSM was prepared, autoclaved and pH adjusted to 7.5 prior to adding glucose.

For selection of resistance all the media were supplemented with the antibiotics ampicillin (50 μg/ml dissolved in water), tetracycline (15 μg/ml dissolved in 70% ethanol) and kanamycin (50 μg/ml dissolved in water) used individually or in combination. *Pseudomonas fluorescens* strains were grown at 30°C whereas *Escherichia coli* and *Bacillus subtilis* strains were grown at 37°C.

2.4.3 Competent cell preparation

Competent *E. coli* bacteria were prepared using the calcium chloride method (Cohen *et al.* 1973). *E. coli* cells were grown overnight in 5 ml L-broth at 35°C at 200 rpm. Overnight cultures were diluted 1 in 100 using fresh L-broth and were grown again at 37°C at 200 rpm until the optical density of the culture at 600 nm reached 0.4-0.6. Cells were pelleted by centrifugation at 5000 X g for 7 min at 4°C. After discarding the supernatant, the cell pellet was re-suspended in pre chilled 100 mM calcium chloride (2 ml per 5 ml culture) and incubated on ice for 20 min. Cells were pelleted again by centrifugation at 5000 X g for 7 min at 4°C and the pellet was re-suspended in pre chilled 100 mM calcium chloride (0.5 ml per 5 ml culture) to make the cells competent. These competent cells were stored in 4°C and used within two weeks.

2.4.4 Polymerase Chain Reaction

Standard polymerase chain reactions (PCR) were performed to amplify regions of chromosomal or plasmid DNA, both for cloning or for testing the cloned products. For cloning purpose Q5 High Fidelity (HF) kit was used, because this enzyme has proof-reading ability which ensures accurate DNA replication. However, it also generates DNA fragments with blunt ends. For testing purposes the Taq Polymerase kit from Invitrogen was used. Taq does not have proof reading ability but it is cheaper than Q5 HF enzyme and hence better for testing. Table 2.2 lists the ingredients needed for a single PCR reaction using a Q5 or Taq DNA polymerase.

Table 2.2: *PCR reaction master mix for Q5 and Taq DNA polymerase*

Q5		**Taq**	
Ingredient	Quantity	Ingredient	Quantity
Water	25.5 / 14.5 μl	Water	12.4 μl
Q5 buffer	10 μl	50% glycerol	10 μl
10 mM dNTPs	4 μl	10 X PCR buffer	5 μl
10 μM forward Primer	5 μl	50 mM MgCl2	1 μl
10 μM reverse primer	5 μl	10 mM dNTPs	4 μl
Q5 polymerase	0.5 μl	10 μM forward Primer	6 μl
Template DNA	1 μl	10 μM reverse primer	6 μl
Enhancer (optional)	0 / 10 μl	Taq DNA polymerase	0.6 μl
		Template DNA	5 μl
Total volume	50 μl		50 μl

For amplifying the three fragments, i.e. ACP-K24a and the two ACP-mupA3ab flanking regions from the *mup* cluster, six primers were designed. The primers were designed (Table 2.3) to perform Gibson assembly in the next step therefore the gene-specific primer sequence (highlighted in blue) also included the overlap region (highlighted in red) (more details in section 2.4.6). All the primers were obtained from Alta Biosciences, University of Birmingham.

To amplify the kalimantacin ACP-K24a, purified kalimantacin chromosomal DNA supplied by our collaborators the group of Prof. Rob Lavigne in Belgium, was used as the template. The reaction mix was prepared for the Q5 polymerase as mentioned in the Table 2.2, with

Table 2.3: *Primers for amplifying ACP-K24a, left and right flanking regions.*

Primer	Sequence	Length	Product size
acpK24a-F	CGCGGATACTCAGCCAACAGGACATCGTCGTA	32	297+(17x2)
acpK24a-R	GCAGCCGTTGCTTGAGTCGACGGATAACGGTT	32	
mup-LF	ACCCGGTGAATTCTCTAGAACGCGAGCCAGACCTGCAAGC	40	487+20
mup-LR	TTGGCTGAGTATCCGCGGCG	20	
mup-RF	ACTCAAGCAACGGCTGCG	18	532+20
mup-RR	GGTAATCCCGGATCCCCGGGAACAGCATGGTGCAATCGC	39	

F and R denotes forward and reverse; mup-L/R denotes left and right flanking regions. Product size is the intended length of the amplified fragments including the length of the homology region (red).

and without the optional enhancer. The annealing temperature for the primers were calculated using the NEB TM calculator (`http://tmcalculator.neb.com`). Since the primers were designed to be used with the Gibson assembly in the next step two annealing temperatures were calculated, the annealing temperature for the gene specific primer sequence was 57°C and the annealing temperature for the whole sequence including the overlap sequence was 72°C. Table 2.4 shows the programme used for the kalimantacin ACP-K24a amplification using Q5 polymerase kit.

Table 2.4: *The PCR program used for amplifying ACP-K24a with the Q5 polymerase kit.*

Steps	Temperature (°C)	Time (sec)	Cycle
Initial denaturation	98	30	1
Denaturation	98	10	2
Annealing	57	20	2
Extension	72	30	2
Denaturation	98	10	28
Annealing & Extension	72	40	28
Long extension	72	120	1

To amplify DNA for available bacterial strains, individual colonies were picked and suspended in 50 μl of sterile distilled water. The suspension was boiled for 10 mins and then centrifuged to 15000 X g for 5 mins, 2 μl of the supernatant was used as the template DNA

and the rest of the ingredients were as described in the Table 2.2. For the left arm, annealing temperature calculated for the gene specific primer region including the overlap region, was 72°C. Table 2.5 shows the program used for the left arm PCR.

Table 2.5: *The PCR program used for amplifying the left arm with the Q5 polymerase kit.*

Steps	**Temperature (°C)**	**Time (sec)**	**Cycle**
Initial denaturation	98	30	1
Denaturation	98	10	30
Annealing and extension	72	60	30
Long extension	72	120	1

To amplify the right arm, the annealing temperature initially calculated for the gene specific primer region was 68°C and including the overlap region was 70°C. However, I found that while using two different annealing temperatures at different stages of the PCR cycle, working at the calculated temperatures sometimes does not produce good results. To explore temperatures that might produce better results, a gradient PCR was setup. In 1°C steps on the PCR block and 6 samples were placed with gaps to give 60, 62, 64, 66, 68 and 69°C. Table 2.6 shows the programme used for the right arm.

Table 2.6: *The PCR programme used for amplifying the right arm with the Q5 polymerase kit.*

Steps	**Temperature (°C)**	**Time (sec)**	**Cycle**
Initial denaturation	98	30	1
Denaturation	98	10	2
Annealing	65.5 (5.5 gradient)	20	2
Extension	72	50	2
Denaturation	98	10	28
Annealing	67.5 (5.5 gradient)	20	28
Extension	72	50	28
Long extension	72	120	1

2.4.5 Agarose gel electrophoresis

Agarose gel electrophoresis were performed to verify the size of the DNA fragments generated either by PCR or restriction digestion reactions. 100 ml of 1% agarose gels were prepared by melting 1 g of agarose in 98 ml of distilled water followed by addition of 2 ml of 50X tris acetate EDTA (TAE) buffer and 2 μl of ethidium bromide solution (concentration 2.0 μg/μl). 1 kb DNA Ladder from Thermo Scientific was used as the molecular weight marker.

2.4.6 Gibson Assembly

Gibson assembly is a method that can efficiently join multiple overlapping DNA fragments in a single isothermal reaction as illustrated in Figure 2.1. It requires that the DNA fragments contain ≈20-40 base pair overlap with adjacent DNA fragments. The Gibson assembly master mix consists of three different enzymes in a single buffer, which include: an exonuclease, which creates single stranded 3' strand which serves as the overlapping region complementary to the adjacent DNA fragment; a polymerase enzyme, which extends the gaps within each annealed fragment; and a DNA ligase which seals the nicks in the assembled DNA. To get the DNA fragments for the Gibson assembly, i.e. ACP-K24a from the kalimantacin cluster and the two five hundred base long DNA fragments from both the sides of ACP-mupA3a (left and right arm), PCR was performed as explained in the section 2.4.4.

Once fragments of the desired size were obtained, the three PCR products were purified using the GE Healthcare Life Sciences Illustra-GRX PCR DNA and gel band purification kit and the concentration of DNA in the solution was estimated using the Nanodrop instrument. Simultaneously suicide plasmid vector pKAE604 was digested using Hind III and Sal I restriction enzymes. The digested product was purified and the concentration of the plasmid DNA was measured. For Gibson assembly NEB recommends a total DNA concentration of 0.2 to 1.0 pM when 4 to 6 DNA fragments are to be assembled. Therefore, for 4 fragments 0.25 pM is required and the weight in ng for each of the four fragments can be calculated with equation 2.3. This allowed the volume required in μl of the extracted DNA solutions to be determined. NEB recommends incubation of the Gibson Assembly master mix for an hour at 50°C.

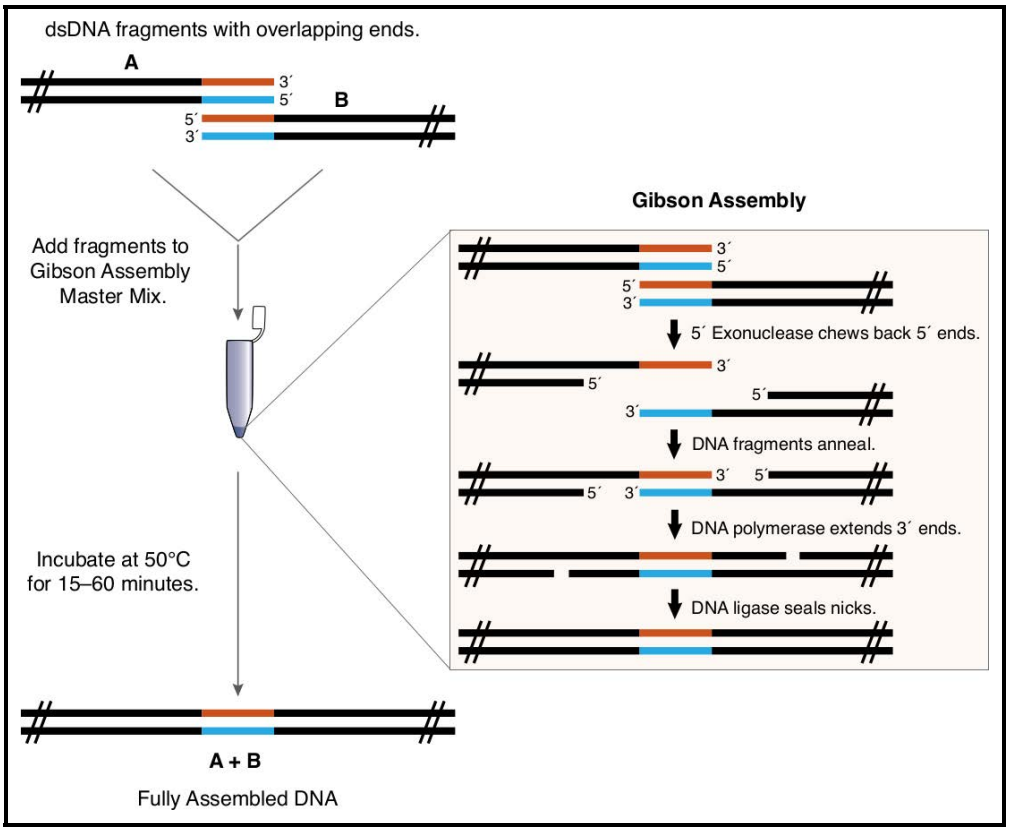

Figure 2.1: *Overview of the Gibson assembly method. Picture reproduced from the NEB Gibsson assembly master mix instruction manual NEB #E2611S/L*

$$weight\ in\ ng = \frac{pmols \times base\ pairs \times 650\ daltons}{1000} \tag{2.3}$$

2.4.7 Transformation and validation

Following the Gibson assembly the ligated product was transformed into freshly prepared *E. coli* DH5α competent cells. As recommended by the NEB Gibson assembly protocol, the ligated product was diluted 4 fold (5 μl of ligation in 15 μl of deionised distilled water) prior to transformation. For transformation 2 μl of the diluted product was added to 100 μl of the competent cells and was held on ice for an hour. After an hour the cells were heat shocked for 90 sec at 42°C and then briefly returned to ice. The culture held on ice was mixed with 500 μl of L-broth and was incubated at 37°C for an hour. Finally 200 μl and 400 μl of the culture was plated on to L-agar plates with kanamycin and grown at 37°C overnight.

PCR was performed on the colonies picked from the transformation plates, using the left arm forward primer and the right arm reverse primer. The method for PCR was same as mentioned in the section 2.4.4 using the Taq polymerase enzyme. Table 2.7 shows the PCR programme used for validating the ligation of the four fragments using Gibson assembly. The PCR product were analysed by 1 % agarose gel electrophoresis.

Table 2.7: *The PCR programme used for validating the product of Gibson assembly with the Taq DNA polymerase kit.*

Steps	**Temperature (°C)**	**Time (sec)**	**Cycle**
Initial denaturation	94	180	1
Denaturation	94	30	30
Annealing and extension	72	90	30
Long extension	72	600	1

Successful PCR products were purified and sequenced using primers designed previously by Dr. Joanne Hothersall. Those primers bind on the region outside the multiple cloning site in the pAKE604 plasmid. The validated sample(s) was used to transform freshly prepared *E. coli* S17-1 competent cells and were grown on L-agar plates with kanamycin selection and a sample

of the successful transformant was stored in the -80°C freezer.

2.4.8 Conjugal transfer of the suicide vector into *P. fluorescens*

For mating, overnight cultures were prepared for the transformed *E. coli* S17-1 and the two *P. fluorescens* host strains. 1 ml of each of the two *P. fluorescens* host strains was mixed separately with 1 ml of the transformed *E. coli* S17-1 strain. 1 ml of the culture mixture was filtered through a millipore nylon filter using a syringe and the filters were placed face up on to an L-agar plate without any antibiotic and grown overnight at 30°C. The bacteria were washed off the filters with 1 ml of saline. 10^{-5} fold serial dilutions were made using 100 μl of the cells washed in saline followed by plating 100 μl of each dilution onto minimal medium plates with kanamycin selection. The plates were grown at 30°C for three to four days till the colonies were visible and big enough to be picked easily.

To detect the successful transfer of the plasmid into the *P. fluorescens* ΔACP4 and ΔMupH cells, colonies were subjected to PCR (as described in Table 2.7) with ACP-K24a specific primers using Taq DNA polymerase. The PCR products were analysed using 1% agarose gel.

2.4.9 Sucrose selection and excisant validation

After mating, the *P. fluorescens trans* conjugants were subject to sucrose selection. For the *P. fluorescens* ΔACP4 trans-conjugants, colonies were grown overnight in L-broth without any selection. For each of the overnight cultures, five serial dilutions of 10^{-1} were made for a final dilution of 10^{-5} and 100 μl of culture from each of the five dilution tubes was plated on L-agar plates with sucrose selection. The cells were grown at 30°C for 24 hrs. To confirm that the plasmid had successfully excised out of the *P. fluorescens* strains, colonies were patched onto two plates carrying ampicillin and kanamycin respectively. The colonies which grew on the ampicillin plates but not on the kanamycin plates were the ones which had successfully excised the plasmid. Successful excisions were subjected to PCR (as described in Table 2.7) with ACP-K24a specific primers using Taq DNA polymerase, in order to detect the integration of the ACP-K24a into *P. fluorescens* ΔACP4 chromosome.

The previous steps validated the integration of the ACP-K24a into the *P. fluorescens* ΔACP4

chromosome but did not detect whether the integration had happened at the correct position in the *mup* cluster. To validate that ACP-K24a had integrated at the correct position another PCR using Taq DNA polymerase was carried out. Primers designed previously by Dr. Anthony Haines, which bind to the positions outside the left and right arms were used with the annealing temperature of 53°C (as described in Table 2.7, with separate annealing and extension steps). *P. fluorescens* NCIMB 10586 and *P. fluorescens* ΔACP4 were used as the controls. The PCR products were further subjected to restriction digests using enzymes Blp I and Stu I. Blp I cuts in the middle of the ACP-K24a but not does not cut anywhere in ACP-mupA3a and Stu I cuts in the middle of the ACP-mupA3a as well as at the rear ends of the two arms. A 20μl restriction digestion reaction mixture consists of DNA (5μl), Buffer (2μl), BSA (1μl), enzyme (conc. 5 units/μl) and enough sterile distilled water to make the total volume of 20μl. The restriction digest fragments were analysed by electrophoresis on 1% agarose gel, an uncut fragment was used as the control.

For *P. fluorescens* ΔMupH trans-conjugants sucrose selection was performed following the same steps as that for the *P. fluorescens* ΔACP4 trans-conjugants. However, to validate the excisions and integrants, samples were subjected to PCR using the primers designed to bind to the region outside the two arms. The steps of PCR with ACP-K24a specific primers and subsequent restriction digestions were excluded, since running a PCR using only the outer primers and sequencing the PCR products gave the same result in fewer steps and less time as compared to the above mentioned steps.

2.4.10 Overlay Bioassay for *in trans* expression of MupH, BatC and BatC L218M mutant

For the bioassay 20 ml measured L-agar plates were made. The *P. fluorescens* ΔH-6d strain carrying one of MupH, BatC or Batc L218M *in trans* and the Δ4-1a strain carrying the BatC *in trans* were grown overnight in 5 ml L-broth with ampicillin selection. *P. fluorescens* NCIMB 10586, ΔMupH, ΔACP4, ΔH-6d and Δ4-1a were also grown without any plasmids, as controls. 100 μl of the overnight cultures were diluted with 900 μl of L-broth. To keep the concentration

of the cells the same in all the cultures, the optical density of the samples was measured at 600 nm. Using equation 2.4, further dilutions were made keeping the total volume of the sample as 1 ml.

$$volume\ of\ the\ sample\ to\ dilute = \frac{smallest\ OD_{600}}{individual\ sample\ OD_{600}} \times 1000 \tag{2.4}$$

10 μl of the diluted samples were spotted onto the 20 ml L-agar plates and were grown for 24 hr on the bench at room temperature. Simultaneously an overnight culture of *Bacillus subtilis* 1064 strain was grown at 37°C. To prepare the overlay medium 4 ml of *Bacillus subtilis* culture were mixed in 100 ml of molten L-agar and 500 μl of TTC (5% w/v). 15 ml of the overlay medium were poured over the spots from the 24 hr grown *P. fluorescens* strains and were allowed to settle till the agar solidified, followed by a 24 hr incubation at 30°C. The diameters of the clearance zones were measured in two different directions, subtracting the diameter of the central disk.

2.4.11 High performance liquid chromatography analysis

To order to detect the pseudomonic acids produced by the *in trans* expression of *mupH*, *batC* and *batC* L218M in the ACP-K24a mutant *P. fluorescens* strains, high performance liquid chromatography (HPLC) was performed. *P. fluorescens* NCIMB 10586 and *P. fluorescens* ΔH-6d and Δ4-1a with an empty pJH10 plasmid were used as controls. The seed cultures were prepared by growing the strains for 16 hr in L-broth at 25°C, 200 rpm. From the seed cultures, 1.25 ml of each sample was inoculated into 25 ml of the secondary stage medium (SSM). The inoculated cultures were grown at 22°C, 200 rpm for 40 hr. The cultures were pelleted in Falcon tubes by centrifugation at 5000 X g, at 25°C for 40 min. The supernatant was separated and filtered using 0.2 μm Acrodiscs Nylon filters (Pall-Gelman, labs) before injection into the HPLC machine.

The HPLC machine (Gilson) was supplied with HPLC grade water (Fisher) and 100% aceto nitrile (Fisher) as mobile phase. The two solvents were added with 0.01% formic acid (v/v) for pH adjustment and were degassed using an aspirator to remove any dissolved air prior to

injection into the machine. The machine was set at 0.002 AUFS (absorbance units full scale) sensitivity and the detection was carried out at the wavelength of 233 nm. A Supelco reverse phase C18 column (15cm X 4.6 mm, 5μm), which had hydrophobic alkyl chains covalently attached to the silica beads, was used to bind the hydrophobic compounds. Elution was performed by using a 5-70% acetonitrile-water gradient at 1ml/min flow rate for 1 hr. Unipoint software was used to run the programme and analyse the chromatograms.

Chapter 3

ACP-HCS interaction in β-branching

3.1 Introduction

Some type I Polyketide biosynthesis systems, for example producing mupirocin, thiomarinol, kalamanticin, or myxovericins, are found to incorporate a branch on the third (i.e. β) carbon in the growing polyketide chain, referred to as a β-branch. This β-branching mechanism is thought to be catalysed via an "HMG-CoA synthase (HCS) cassette", consisting of an HMG-CoA synthase homologue and further auxiliary enzymes. In the mupirocin system the HCS cassette is a set of 5 proteins of which MupH (the HMG-CoA synthase homologue) is the first enzyme to interact with an acyl carrier protein (ACP) from the MmpA subunit. This interaction between ACP and MupH initiates the β-branching reaction.

At the start of the present work little was known about the HCS cassette structure and function or how it interacts with the proteins in PKS? It was not understood what allows MupH to recognize the ACPs with substrates for β-branching, as opposed to the several other ACPs involved in a typical PKS pathway where branching is not required. We also did not know whether HCS proteins have a subtype or if they always work as one set of only five proteins as found in the mupirocin cluster. However, a recent study of the myxovirescin system (Simunovic *et al.* 2006) shows two HCS clusters involved in β-branching at two position in the synthesis pathway. These two stages are catalysed by two non-complimentary pairs of HMG-CoA synthase homologue and ACP interactions, which suggests the possibility of HCS subtypes. In a

study of the curacin system, it was found that the HCS proteins can also work in conjunction with a halogenase. This halogenase activity adds a halogen (chlorine in case of curacin) on the 4^{th} (γ-carbon) carbon of the growing polyketide chain (Busche *et al.* 2012).

A number of questions arise. *1*) How different are the ACPs which interact with the HCS cassette from other ACPs found in the pathway? *2*) Is it possible to predict the molecular features responsible for the ACP-HCS recognition? *3*) And can these molecular features be exploited for the engineering of β-branching function in mupirocin biosynthesis pathway or other systems? I have used computational methods to address these exciting questions. The work started using the preliminary data available from Prof. Christopher M. Thomas lab.

In a sequence analysis carried out by Dr. Anthony Haines on the ACPs from various PKS systems, in β-branching ACPs a conserved tryptophan motif is found (Figure 3.2) at the position 6 residues downstream of the catalytic serine. This position is not found to be conserved in the non-branching ACPs. It was also found that upon mutating this conserved tryptophan to leucine in the mupirocin system the production of pseudomonic acid A is significantly lowered. Thus, it was hypothesized that this conserved tryptophan could be the recognition motif. These preliminary observations lead Dr. Mathew Crump from the University of Bristol to solve the solution structure of the ACP di-domain from module 6 of the MmpA subunit in the mupirocin biosynthesis pathway. This NMR structure (PDB ID 2L22) has been used in the present work to carry out the predictions on the ACP-HCS interaction mechanism.

3.1.1 HMG-CoA synthase cassette

The HCS cassette in the mupirocin biosynthesis pathway, is a set of 5 proteins consisting of an ACP (mACPc), a 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) synthase homologue (MupH), a decarboxylase (MupG) and two proteins (MupJ, MupK) from the crotonase super-family. MupH homologues, our focus here (Wu *et al.* 2007), are found in various polyketide biosynthesis pathways and are thought to be involved in β-branching mechanisms along with the other enzymes in the HCS cassette. It is hypothesized that acyl carrier proteins from the ACP di-domain (ACP-mupA3a and ACP-mupA3b), in the MmpA subunit of the mupirocin

biosynthesis pathway, make the first point of contact with the MupH, initiating the β-branching reaction. This interaction between the ACPs and MupH is thought to be governed by a set of specificity determinants in the interacting residue pairs, which allows the ACPs involved in the β-branching systems to recognise and interact with the proteins in the HCS cassette. To understand better what governs the interaction between the ACPs and MupH, a structural model of the ACP-mupH complex determined either through experimental or computational methods could help to design mutagenesis experiments. The solution structure for the ACP-mupA3ab di-domain has been recently resolved (PDB ID 2L22), however there is no structure available for MupH. To predict a reliable structure of MupH and to understand its structural properties it was necessary to first understand the structural properties and the reaction mechanism catalysed by its homologue HMG-CoA synthase.

3.1.1.1 HMG-CoA synthase reaction mechanism

HMG-CoA synthases (EC 2.3.3.10) are 42 KDa proteins found in a wide range of organisms from bacteria to mammals and play a central role in fatty acid, polyketide, and isoprenoid biosynthesis. The enzyme belongs to the thiolase superfamily and can be broadly classified into the bacterial isoforms, the eukaryotic cytosolic isoforms and the mammalian specific mitochondrial isoform (Shafqat *et al.* 2010).

HMG-CoA synthase catalyses a three step reaction that involves a conserved Cys-His-Glu catalytic triad and an acyl-enzyme intermediate. As in *Enterococcus faecalis* HMG-CoA synthase (PDB Id 1X9E) the first step is a **deacetylation** of the substrate Ac-CoA, H233 is thought to act as a catalytic base or H-bond donor for the nucleophilic C111, which attacks the carbonyl carbon of Ac-CoA , thereby transferring the acetyl group to the Cys-S atom and releasing the reduced CoASH (Figure 3.1) (Steussy *et al.* 2005).

In the second step, the methyl group of acetylated-Cys is deprotonated by the general base G79 to form a carbanion which attacks the distal (β) carbonyl of the incoming AcAC-CoA (second substrate), following the **condensation** of Ac-Co and AcAC-CoA forming an enzyme:HMG-CoA intermediate.

In the final step the resultant enzyme:HMG-CoA intermediate is **hydrolyzed** to release the

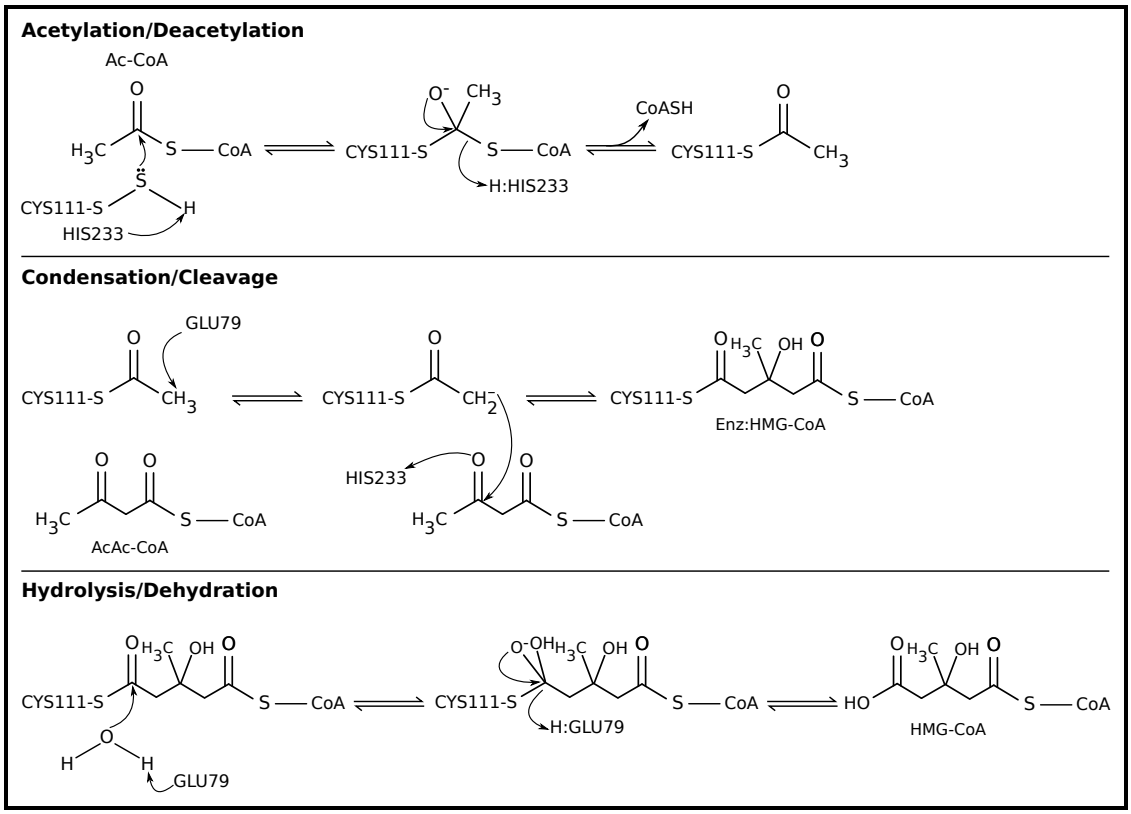

Figure 3.1: *The reaction mechanism of HMG-CoA synthase. proposed in (Steussy* et al. *2005)*

.

product HMG-CoA and regenerate the reduced Cysteine. G79 is shown to mediate the hydrolysis step. The similar reaction mechanism is also observed in *Staphylococcus aureus, Brassica juncea* and Human HMG-CoA synthases (Theisen *et al.* 2004; Shafqat *et al.* 2010).

It is assumed that the β-branching mechanism of MupH would be similar to that of HMG-CoA synthase and thus can be used to guide the modelling.

3.2 Results

3.2.1 ACP sequence analysis

The initial sequence analysis carried out by Dr. Anthony Haines on the seven well-characterized PKS clusters known to be involved in β-branching (Figure 3.2) revealed the presence of a conserved tryptophan 6 residues downstream of the catalytic serine. A W was not seen at this position in the non-branching ACPs. Dr. Haines further used the sequence motif (DSXXXXXW) as a search pattern for PHI-BLAST and found further ACPs that he confirmed from the literature

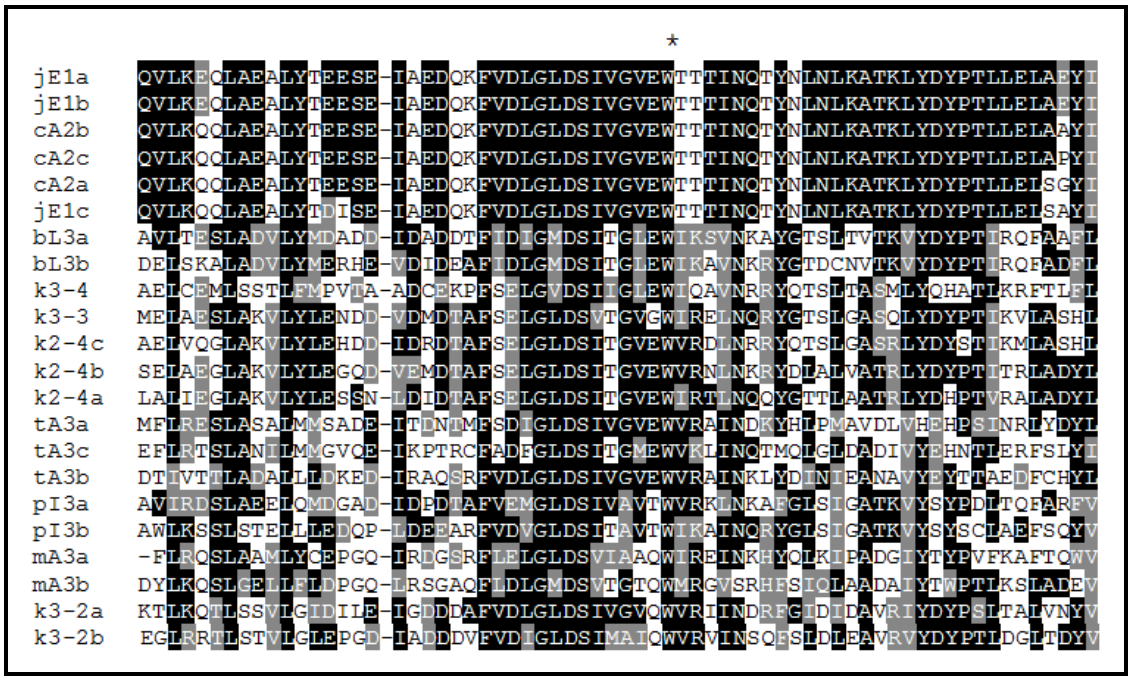

Figure 3.2: *Alignment of the ACP sequences from HCS cassette containing systems provided by Dr. Anthony Haines. The conserved W is indicated by a *.*

were associated with β-branching (Haines *et al.* 2013). This sequence motif predicted ACPs from all the known systems involved in β-branching with the exception of two ACPs from virginiamycin and leinamycin clusters each. Which suggests that although this sequence motif is a strong predictor of β-branching associated ACPs it might not be enough to predict all of them. Therefore in the present study a stronger predictor, based on the statistical method of Hidden Markov models (HMM), was developed to classify ACPs into β-branching and non-β-branching types.

The sequences provided by Dr. Anthony Haines were used to build HMMs using 38 and 178 sequences from 15 well characterised pathways for the β-branching and non-β-branching ACPs respectively. These models were tested using a test set of ACPs (provided by Dr. Anthony Haines) from other clusters which were not part of the training set. These results were plotted on a graph of non-β-branching HMM score against β-branching HMM score (Figure 3.3). The graph was divided by the $y = x$ line where the models predict an ACP as having the same likelihood of being a β-branching ACP or standard ACP. The majority of ACPs from the remaining clusters with at least one identified β-branch-associated ACP locate to one or other

of these clusters (Figure 3.3).

β-branching ACPs from the virginiamycin cluster were identified as the outliers which are just above the $y = x$ line and adjacent to the 'branching' cluster. Other outliers were the two β-branching ACPs from the leinamycin cluster one of which was just below the line and the other actually fallen within the non-branching cluster. Except for leinamycin, classifying ACPs using HMMs agrees with the available information about their likely presence in the β-branching or non-β-branching modules. The model for non-β-branching ACPs (standard) was used to fetch the ACP sequences from the UniProtKB/TrEMBL (20127441 seq) and Refseq microbial (6408654 seq) databases, database version on 9th March, 2012. I developed a couple of Perl scripts (see Appendix I) to remove all the sequences which were shorter than 60 aa or duplicates or sequences without the phosphopantethinylated serine which resulted in a set of 16,490 unique sequences. To ensure that these sequences cover the full length of the model, they were extended by 7 residues on both the ends. The extended sequences were scored using both the HMM models and a scatter diagram was plotted (Figure 3. Scatter diagrams showing the separation of ACPs into two clusters by their fit to the β-branch-associated 'branching' ACP HMM and the non-branching 'standard' ACP HMM. (b)). In Figure 3.3 the scatter plots shows the separation of ACPs into two clusters by their fit to the β-branch-associated 'branching' ACP HMM and the non-branching 'standard' ACP HMM. These scatter plots were rendered by Dr. Anthony Haines using the data provided by the HMM analysis carried out by Rohit Farmer (this graph is also presented in (Haines *et al.* 2013)).

The newly found ACPs in the screen could be classified as likely β-branching or non-branching. On the graph, close to the $y = x$ line, an HMM score of above 45 represents the ambiguous regions where the branching state of an ACP cannot be predicted below 45 seems to be indicative of non branching. One ACP from the leinamycin cluster was found in this region, the other in the non-branching region. The two β-branching ACPs from the myxovirescin, cluster, which are associated with different MupH homologues, can be clearly identified. This suggests that the HMM model can be used to identify the ACPs associated with the β-branching but cannot be used to predict the ACP subtypes if they exist.

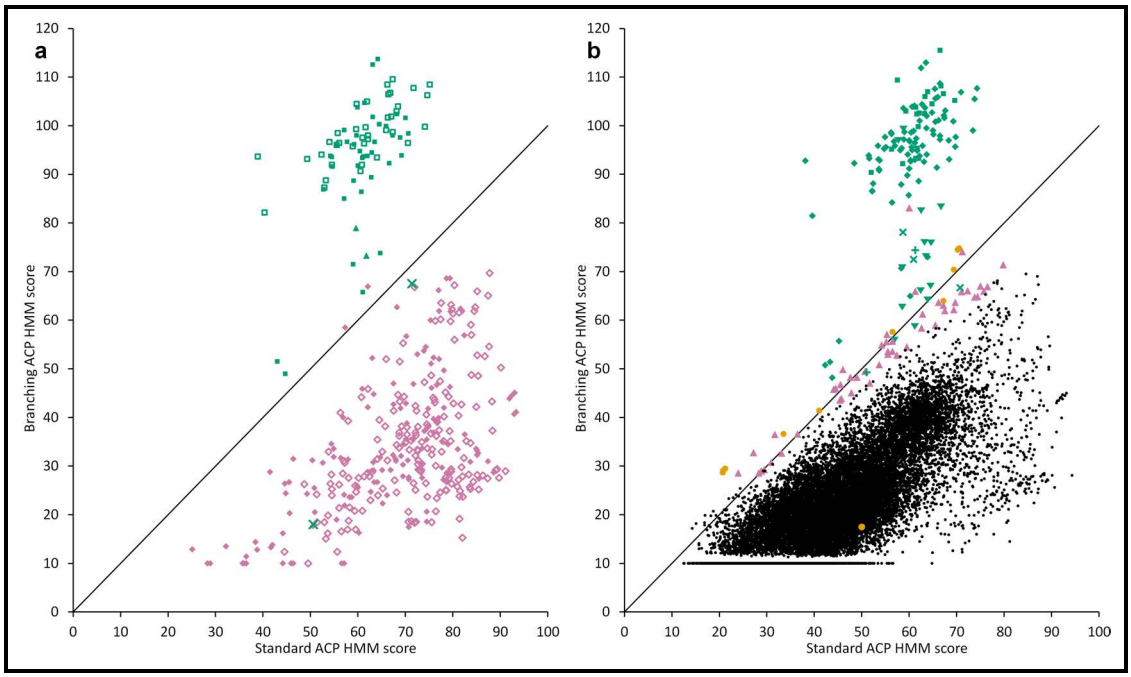

Figure 3.3: *Scatter diagrams showing the separation of ACPs into two clusters by their fit to the β-branch-associated branching ACP HMM and the non-branching standard ACP HMM. (A) ACPs from 26 pks clusters with at least one known or predicted branching ACP. (□) Training set for HMM using known branching ACPs. (■) predicted branching except (▲) virginiamycin cluster branching ACPs and (×) leinamycin cluster β-branch-associated module ACPs (◊) Training set for HMM using non-branching ACPs (♦) predicted non-branching ACPs. (B) 16,490 ACP-like sequences identified by screening the TrEMBL and RefSeq protein databases using the standard ACP HMM. Sequences which did not pass the branching HMM cut-off were conferred a score of 10 so they could be plotted. (♦) branching ACP. (■) unlisted variants in similar clusters (×) known branching, (▼) predicted branching, identified in this screen (+) ACPs which may add branches in a non-type I-pks pathway. (●) insufficient sequence context or conflicting information. (▲) predicted non-branching (·) not examined. Graph and parts of the figure legend copied from Haines* et al. *(2013).*

```
>ACP-tmlD3a-59.8
                                0     6    11 14    20     27
SVICEALSDALKVPKKMIDPTEAFSDYGLDSITGVNVAQTISSVLNVDLKTTALFDYVCIDQLARYV
                                      WI   N  Y     A      P
>gen1-68.7
SVICEALSDALKVPKKMIDPTEAFSDYGLDSITGVNWAQTISSVLNVDLKTTALFDYVCIDQLARYV
>gen2-72.7
SVICEALSDALKVPKKMIDPTEAFSDYGLDSITGVNWAQTISSVLNVDLKTTALFDYPCIDQLARYV
>gen3-76.6
SVICEALSDALKVPKKMIDPTEAFSDYGLDSITGVNWAQTISSVYNVDLKTTALFDYPCIDQLARYV
>gen4-80.0
SVICEALSDALKVPKKMIDPTEAFSDYGLDSITGVNWAQTINSVYNVDLKTTALFDYPCIDQLARYV
>gen5-82.9
SVICEALSDALKVPKKMIDPTEAFSDYGLDSITGVNWIQTINSVYNVDLKTTALFDYPCIDQLARYV
>gen6-85.7
SVICEALSDALKVPKKMIDPTEAFSDYGLDSITGVNWIQTINSVYNVDLKATALFDYPCIDQLARYV
```

Figure 3.4: *The ACP-tmlD3a sequence scored 59.8 against β-branching HMM model and would require 5 mutations at the positions 6, 7, 11, 14 , and 20 counting from the active site serine (0) to reach the score 82.9 and 6th mutation at position 27 to score 85.7. Figure 3 .9 shows the relative positions of the residues required to be mutated on the ACP-tmlD3a:MupH complex*

3.2.1.1 Minimum changes required to shift ACP-tmlD3a from non-β-branching to β-branching cluster.

It was observed in the HMM analysis that the β-branching ACPs and the non-β-branching ACPs (standard) cluster in different zones, which raises a question as to the minimum changes necessary to shift an ACP from standard ACP cluster to β-branching ACP cluster, as scored by the HMM models? These mutations might allow us to make a non-branching ACP function like a β-branching ACP.

To address this question I wrote a Perl script (Appendix I, Script A.3) to mutate each and every position in the sequence to the other 19 proteinogenic amino acids. A non branching ACP (ACP-tmlD3a) from the thiomarinol cluster was taken as the sequence of reference. ACP-tmlD3a was chosen because experiments from Prof. Thomas group showed no mupirocin production upon replacing ACP-mupA3a/b with ACP-tmlD3a/b. The script generated all possible substitutions at each of the 67 positions in the sequence, giving 19 X 67 = 1273 sequences.

From the previous analysis it was observed that the β-branching ACP sequences typically score in the range of 82.2 to 109.6 using β-branching HMM model (wacp.hmm). This suggests

that if any sequence scores 82.2 or above using β-branching HMM model, should fall under the β-branching cluster. Utilizing this observation the 1273 sequences were scored using β-branching model thus generating a first generation of mutated sequences. The sequence with the highest score was selected for the next generation of mutations and the process was iterated till a score of 82.2 or above was reached. For ACP-tmlD3a it took 5 generations (iteration) to reach the score of 82.9 and 85.7 in 6 generations.

This experiment suggests that it would need to make a minimum of 6 mutations at the positions shown in Figure 3.4 and listed in Table 3.1, to shift the ACP-tmlD3a from the non-branching ACP cluster to the β-branching ACP cluster. Unsurprisingly, the highest scoring amino acid change in the first generation of mutated sequences was tryptophan (W). All the other mutations were observed downstream from the active site serine are mainly towards helix III (Figure 3.5). The same method could be applied to the other ACPs as well however, the number of mutations (iterations) required may be different.

Table 3.1: *Muations required for ACP-tmlD3a sequence to score more highly with the HMM trained on branching ACPs than with the non-branching ACPs*

Mutation	Branching ACP HMM	Standard ACP HMM
ACP-tmlD3a	59.8	79.4
V36W	68.7	76.1
V57P	72.7	79
L44Y	76.6	79.3
S41I	80	80.1
A37I	82.9	80.1
T50A	85.7	78.9

3.2.2 ACP structure analysis

The NMR determined apo ACP-mupA3ab (PDB ID 2L22) structures consists of a typical four helical bundle. The NMR experiments showed that the conserved tryptophan identified in the sequence analysis lies buried in the ACPs core between helix I and helix II rather than forming an exposed patch. The burial of the tryptophan inside the ACP core raised the question of how this residue permits an interaction between the ACPs and the HMG-CoA synthase homologue.

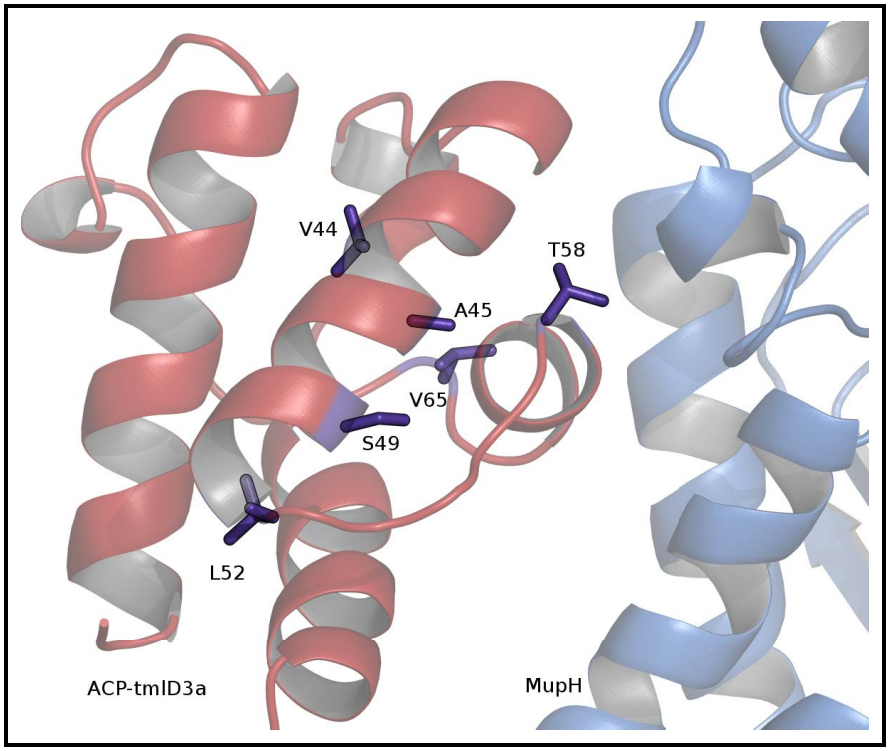

Figure 3.5: *Mutations required reaching the score of 82.2 or above when scored with β-branching HMM model mapped on the structure. ACP-tmlD3a (red) in complex with MupH (Blue) superimposed on the mupA3a:MupH complex 1 from cluster 1. The residues displayed as sticks are the positions for the mutations V44W, V65P, L52Y, S49I, A45I and T58A that are needed for the ACP-tmlD3a sequence to score more highly in the HMM trained on β-branching ACPs than in the HMM trained on non-β-branching ACPs (see Table 3.1)*

As mentioned earlier that on the basis of the W to L mutation experiments which showed a substantial decrease in the pseudomonic acid A production.

The orientation of the tryptophan seen in ACP-mupA3a is almost perpendicular to that in ACP-mupA3b, Figure 3.8 shows ACP-mupA3a and ACP-mupA3b superimposed on helix II, however, both the configurations are in trans Figure 3.6 shows that the conformation of W is consistent with the ensemble for each ACP. Figure 3.9 shows the 20 NMR structures of an ACP homologue from the curacin system (PDB ID 2LIU). Figure 3.10 shows the structural comparison of the curacin ACPs (2LIU, 2LIW) with the mup ACPs. The tryptophan side chains can be seen to form a continuum rather than being biased towards a preferred orientation. Notably in all the above mentioned ACP structures the tryptophan side chain tends to orient similarly within a given NMR ensemble but is different between ensembles. However, the curacin ACP (2LIU and 2LIW) ensembles are more similar to each other than ACP-mupA3a and ACP-mupA3b to each other. Table 3.2 lists the rotameric values for the tryptophan side chain in the above mentioned four ACP structures. The tryptophan side chain atom numbering scheme is based on the Recommendations for the presentation of NMR structures of proteins and nucleic acids (Markley *et al.* 1998). Another notable difference in the ACP-mupA3a/b structure is the position of the helix III (Figure 3.11).

Table 3.2: *Backbone dependent tryptophan side chain rotameric values*

ACPs	Phi	Psi	Chi1	Chi2	
	(C'_{i-1}-N_i- C^{α}_i-C'_i)	(N_i-C^{α}_i-C'_i-N_{i+1})	(N-C^{α}-C^{β}-C^{γ})	(C^{α}-C^{β}-C^{γ}-$C^{\delta 1}$)	(C^{α}-C^{β}-C^{γ}-$C^{\delta 2}$)
mupA3a	-62	-38.5	-158.8	77.3	-103.8
mupA3b	-59.5	-54.7	-178.2	10.9	-170.6
2LIU	-65.3	-36.5	-168.6	59.8	-118.6
2LIW	-66.8	-34	-169.3	45.2	-132.4

These measured rotamer values for the tryptophan side chain atoms were compared with the values given in the default set of the 2010 backbone-dependent rotamer library (Shapovalov and Dunbrack 2011) from Ronald Dunbrack's group. By looking into the Dunbracks rotamer library it was not obvious what causes the difference in the orientation of the tryptophan as the

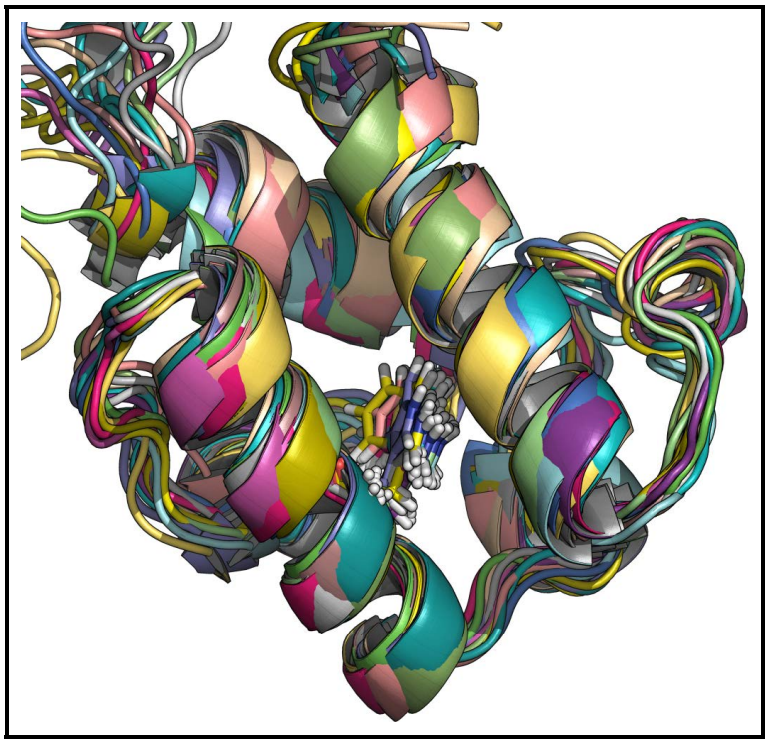

Figure 3.6: *The 20 ACP-mupA3a NMR models superimposed on each other. Tryptophan highlighted as sticks.*

calculated values lie within the rotamer probability distribution. To supplement the observation the NMR structure of ACP-mupA3a/b had all side chains removed and the SCWRL4 software (Krivov *et al.* 2009), which uses the Dunbrack rotamer library, was used to put the side chains back on the ACP backbone. SCWRL4 placed the side chains at similar positions to those of the original side chains in the NMR structure for both ACP-mupA3a/b, which suggests nothing unusual about the W residues.

3.2.2.1 Affect of W to L mutation on ACP molecular dynamics

To investigate the possible effect of the W to L mutation on the structural dynamics of ACP-mupA3a, molecular dynamics simulations of the wild type and mutant ACP in water were carried out (see methods section 2.3.2.2) . Figure 3.12 shows the averaged root mean square fluctuation (RMSF) of 20 simulations for wild and mutant type each. The RMSF values suggest significantly greater motion in the mutant than in the wild type, with the largest effects being around helix III and in the loop between helix I and helix II.

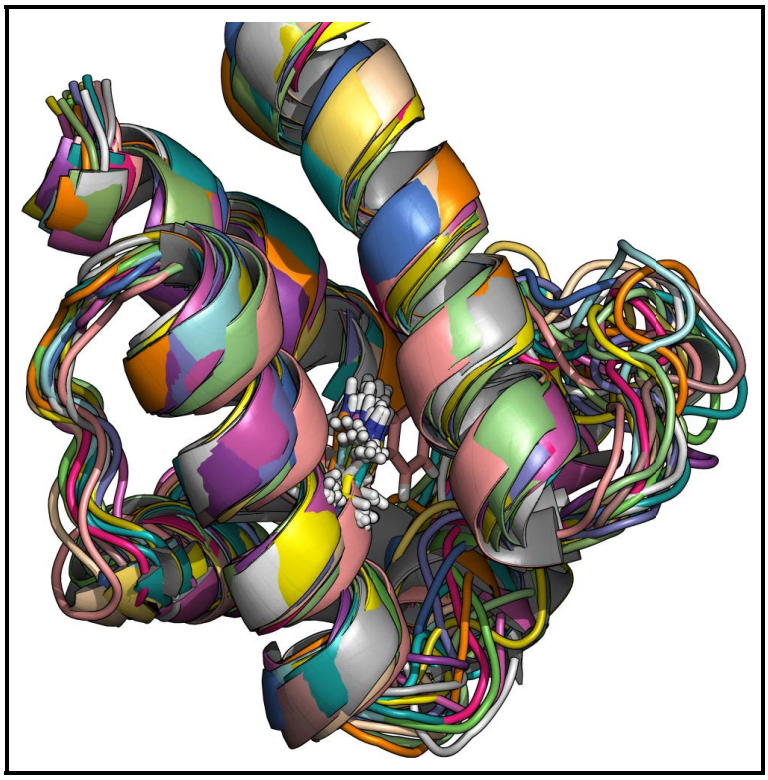

Figure 3.7: *The 20 ACP-mupA3b NMR models superimposed on each other. Tryptophan highlighted as sticks.*

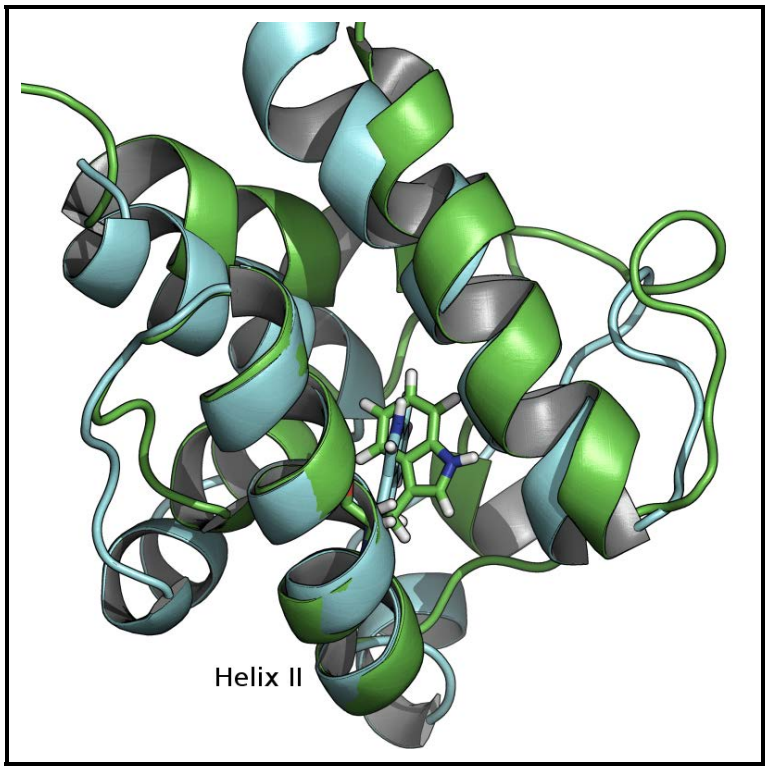

Figure 3.8: *ACP-mupA3a (green) and ACP-mupA3b (cyan) superimposed on helix II. Tryptophan highlighted as sticks*

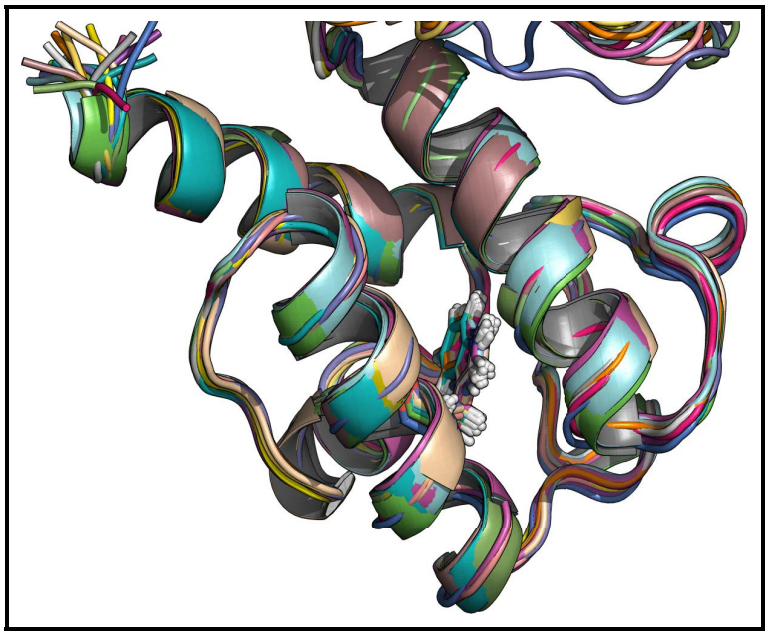

Figure 3.9: *Curacin ACP responsible for halogenase activity via β-branching mechanism (PDB ID 2LIU), NMR models superimposed on each other. Tryptophan highlighted as sticks.*

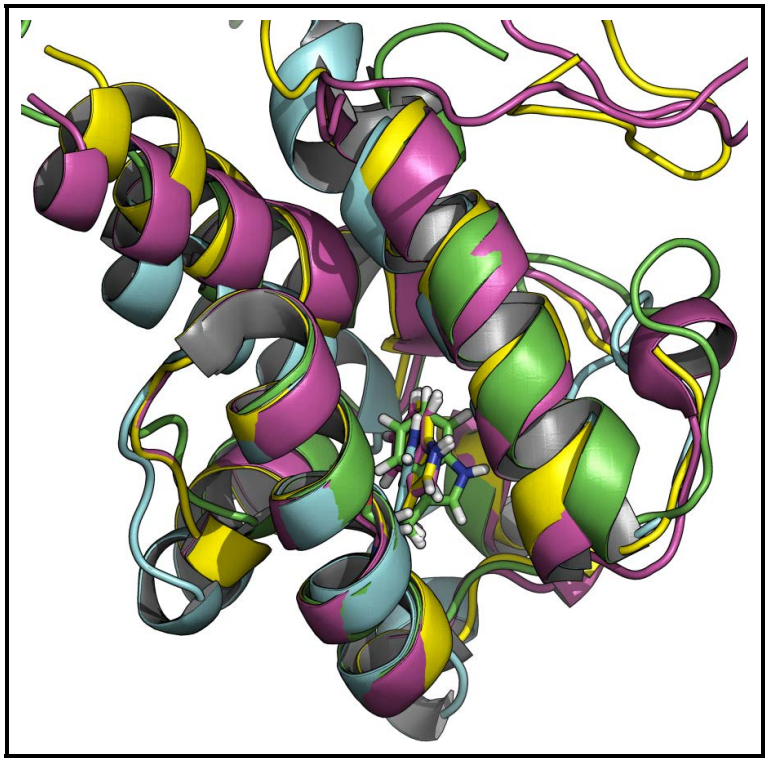

Figure 3.10: *Superimposed representative structures of ACP-mupA3a (green), ACP-mupA3b (cyan), curacin ACPS 2LIU (magenta) and 2LIW (yellow). Tryptophan highlighted as sticks.*

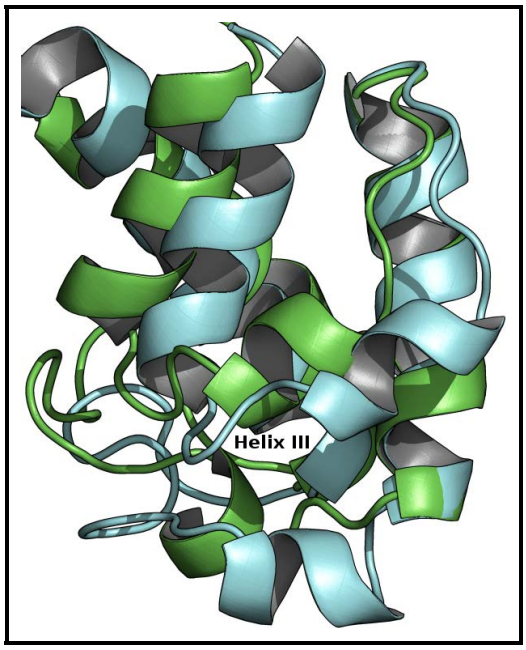

Figure 3.11: *Superimposed ACP-mupA3a (green) and ACP-mupA3b (cyan) highlighting the difference in helix III position.*

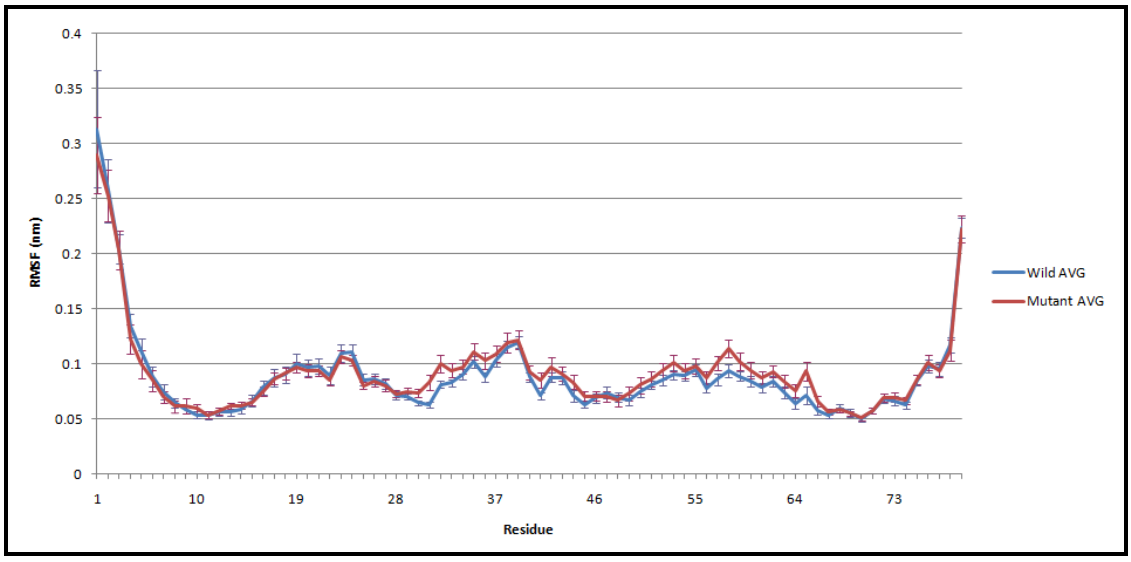

Figure 3.12: *Graph of the average over 20 simulations of the root mean squared fluctuation (RMSF) of the backbone calculated for each simulation of the wild type (blue) and W44L mutant (red) of ACP-mupA3a. The error bars represent the least significant difference at the 95% confidence level; i.e. where the error bars of the two lines do not overlap they are signifcant in a t-test with 95% confidence. Residues 29-35 (central part of the loop between helix I and helix II), 41, 44 (in helix II), 57-59 (N-ter of helix III), and 64 (C-ter of helix 3) show significantly greater motion in the mutant than in the wild type, with the largest effects being around helix III and in the loop between helix I and helix II. Graph reproduced from Haines* et al. *(2013).*

3.2.3 MupH structure prediction

As the MupH structure was not available in the PDB database, homology modelling was used to predict its three dimensional structure. A structural homologue from *Staphylococcus auereus* (PDB ID 1X9E) was used as the template for the protein structure prediction as well as for the docking of the mupirocin intermediate in the predicted structure (details in section 2.3.1.1). The predicted structures were verified for their stereochemical qualities and energy minimized to stabilise the docked ligand inside the active site (Figure 3.13). To verify the correct orientation of the ligand inside the MupH active site the residues within 5 Å of the docked ligand (Figure 3.14) were compared with conserved and functionally important residues in the MupH homologue structures (Figure 3.15).

3.2.4 Similarity between MupH and HMG-CoA homologues in sequence and structure

Despite the low level of sequence identity among MupH homologues, the key catalytic and substrate binding residues mentioned in the literature are highly conserved (Figure 3.16). Residues that are conserved and identified as important in the literature are listed in Table 3.3. The conserved residues can be divided into four categories, (1) catalytic triad (2) residues responsible for substrate orientation in the active site (3) tunnel residues (4) gate Keeper residues (Misra and Miziorko 1996; Bahnson 2004; Theisen *et al.* 2004; Steussy *et al.* 2005; Shafqat *et al.* 2010).

3.2.4.1 Catalytic triad and the essential residues responsible for substrate orientation in the active site

In the sequence and structure comparison of all the HMG-CoA homologue structures from *Staphylococcus auereus, Enterococcus faecalis* and *Homo sapiens* used in this study and MupH, the **catalytic triad** (Cys - His - Glu) is found to be conserved. The residues in the active site responsible for the correct orientation of the substrate are also absolutely conserved with some minor variation in MupH.

The acetyl moiety bound to the catalytic cysteine (Figure 3.18) is constrained by interactions

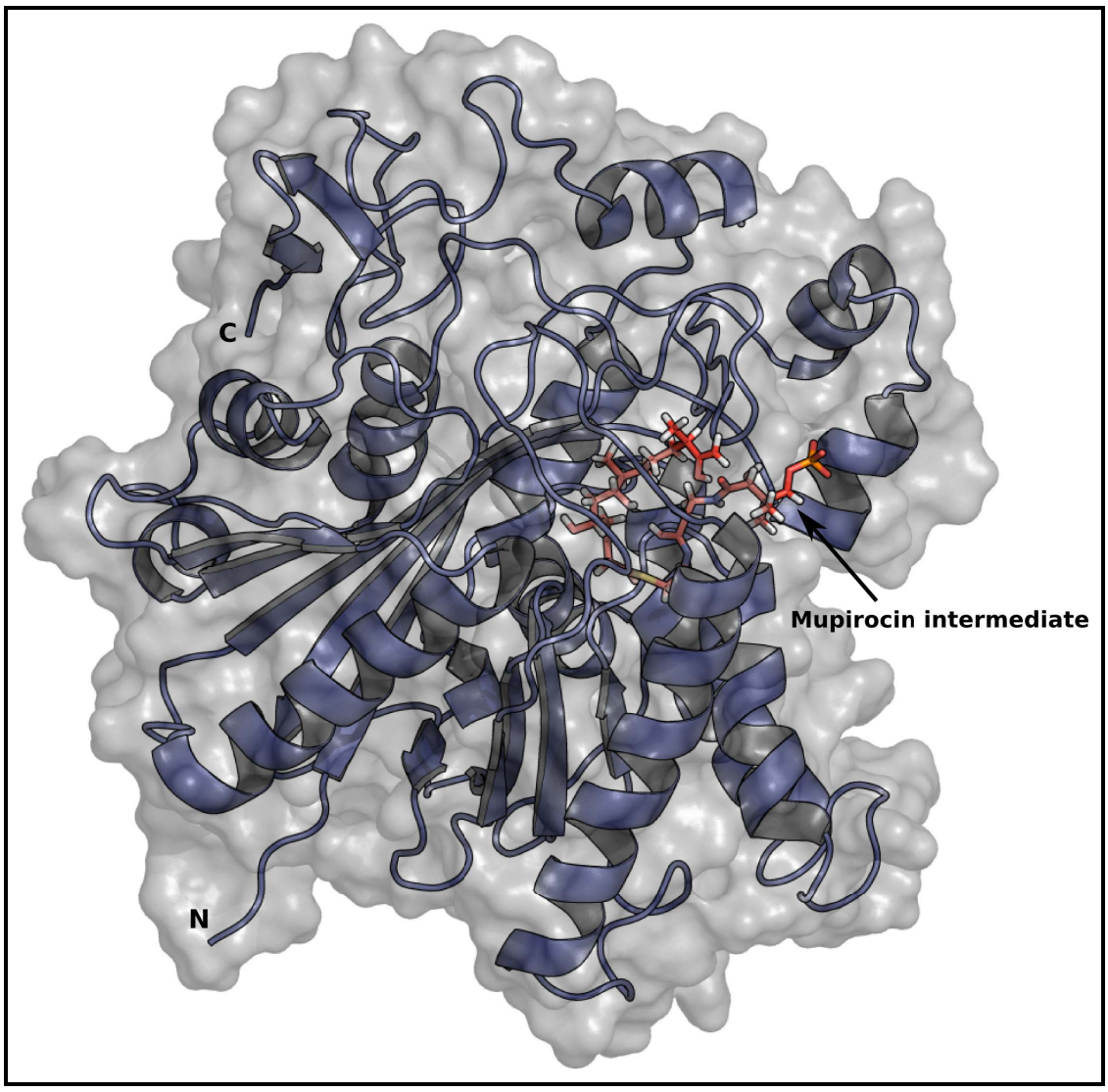

Figure 3.13: *Homology model of MupH complexed with the mupirocin intermediate based on the HMG-CoA synthase X-ray structure from Enterococcus faecalis (PDB ID 1X9E).*

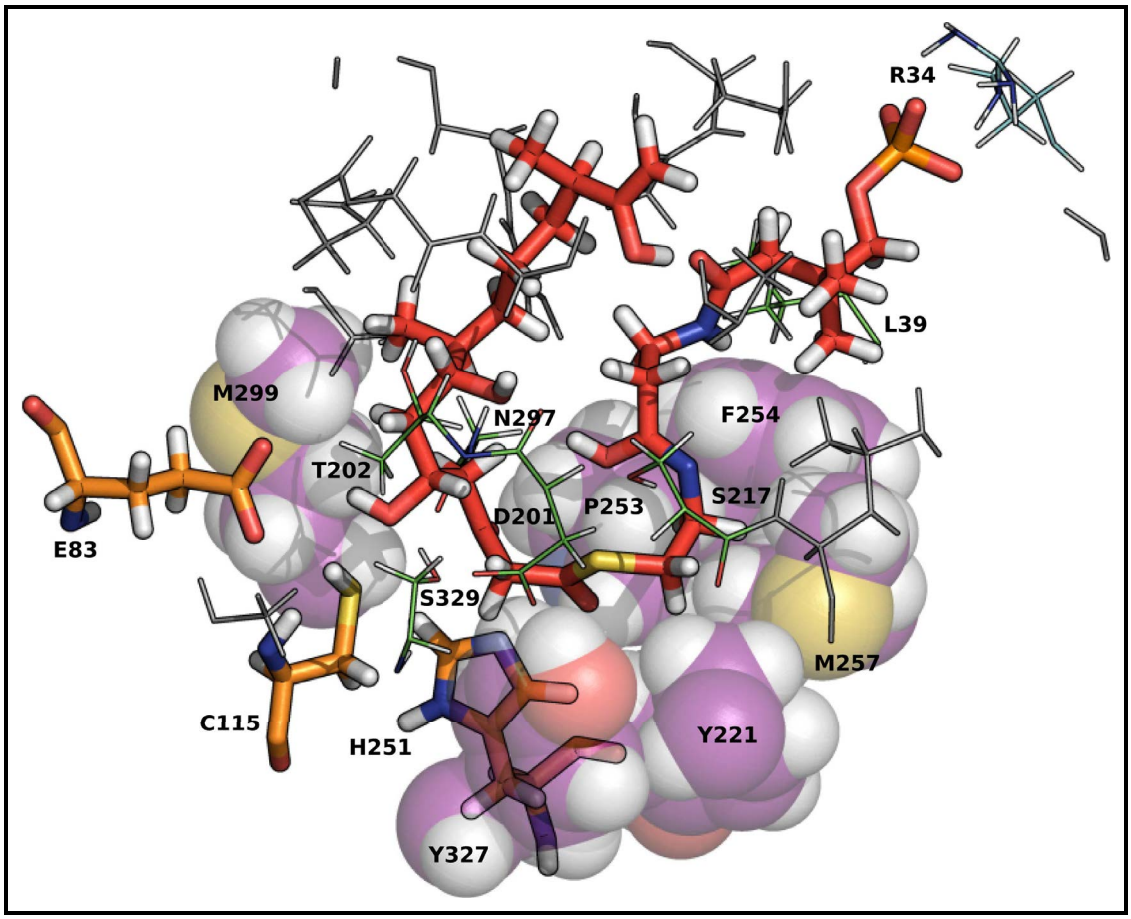

Figure 3.14: *Residues within a 5 Å radius of the mupirocin intermediate. The residues of the catalytic triad are shown as sticks, other active site essential residues are shown as lines with green backbone, tunnel lining residues are shown as spheres, the gate keeper residues are shown as line with cyan backbone. The term catalytic triad, other essential residues, tunnel lining residues and the gate keeper residues are further discussed in detail in the Section 3.2.4. C115 and E83 are placed close the β-carbon thus the model is consistent with the catalytic activity.*

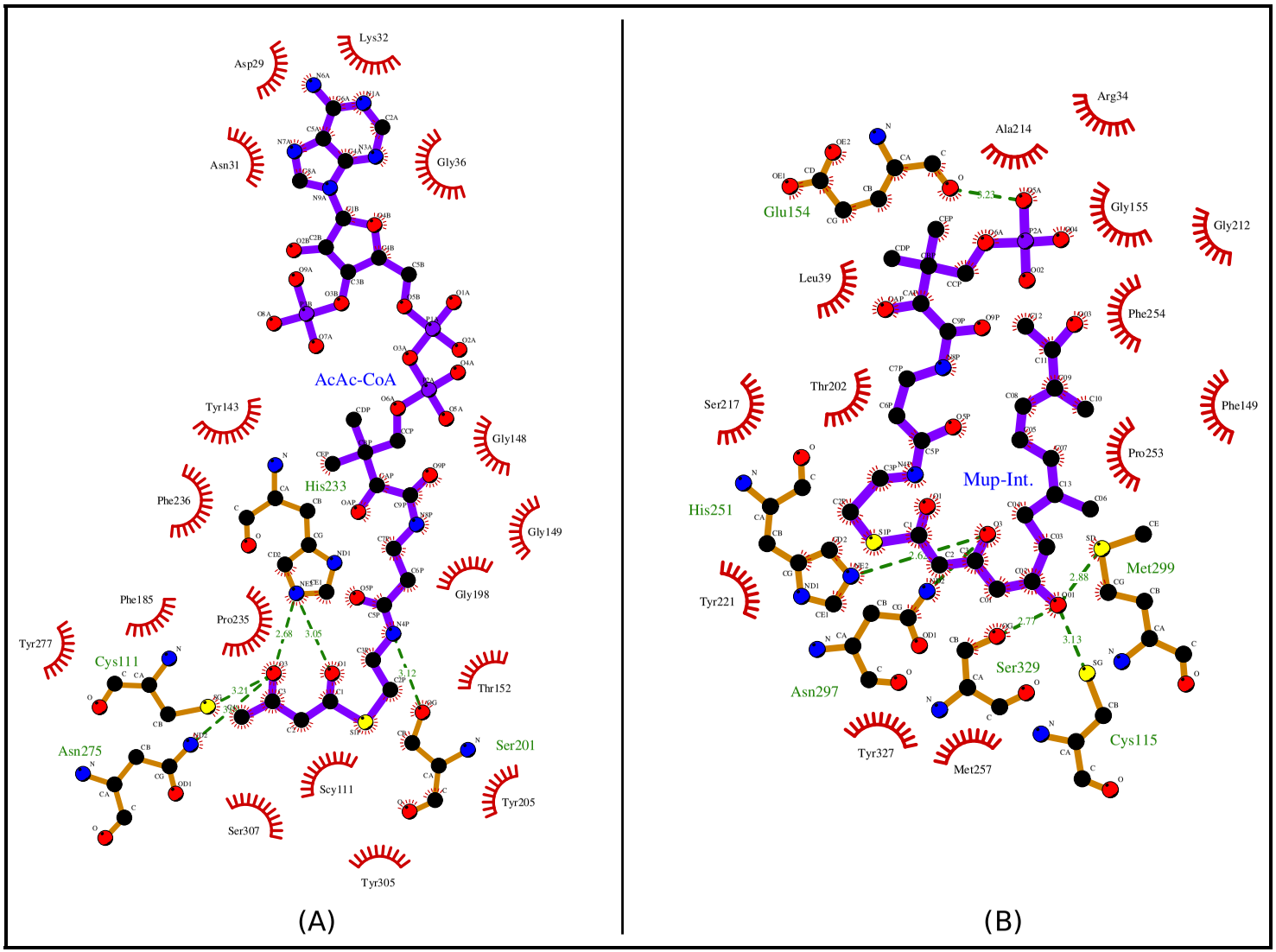

Figure 3.15: *Protein-ligand interaction plot for HMG-CoA homologues. (A) crystal structure of Staphylococuccous aureus HMG-CoA (PDB ID 1XPK) complexed with acetoacetyl-CoA and (B) modelled MupH complexed with PKS bound mupirocin intermediate. Green dotted lines shows hydrogen bonds and red arches around the residues shows hydrophobic interactions. Plot produced by LigPlot (Wallace* et al. *1995).*

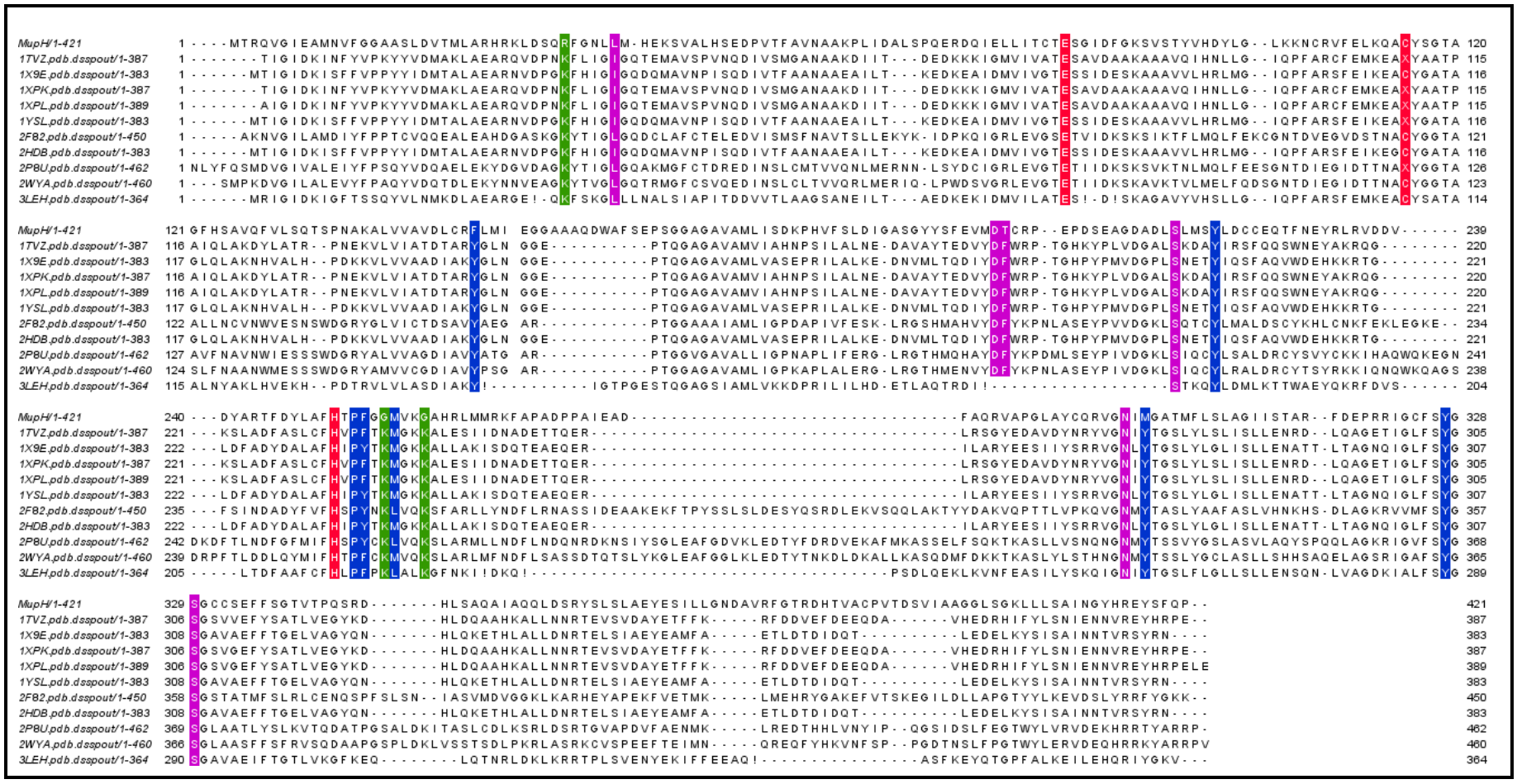

Figure 3.16: *Sequence alignment of the templates used for the MupH modelling. Red: Catalytic Triad, Magenta: residues responsible for the substrate orientation in the active site, Blue: Tunnel residues, Green: Gate keeper residues. Sequences are taken from the template structures ! indicates a break in the chain associated with unresolved residues.*

Table 3.3: *List of all the residues found conserved in the alignment with annotations from the literature*

	Homo sapiens	*Staphylococcus aureus*	*Enterococcus faecalis*	*Pseudomonas fluorescens*
MupH orthologue				x
HMG-CoA orthologue	x	x	x	
PDB ID	2WYA	1XPL, 1XPK	1X9E, 1YSL	MupH
Active Site	GLU 132	**GLU 79**	**GLU 79**	GLU 83
	CYS 166	**CYS 111**	**CYS 111**	CYS 115
	HIS 301	**HIS 233**	**HIS 233**	HIS 251
Other essential active site residues	ASN 380	**ASN 275**	**ASN 275**	ASN 297
	ASP 240	ASP 184	**ASP 184**	ASP 201
	PHE 241	PHE 185	**PHE 185**	THR 202
	SER 258	SER 201	SER 201	SER 217
	SER 414	**SER 307**	**SER 308**	SER 329
	LEU 88	ILE 37	ILE 37	LEU 39
Tunnel Residues	TYR 200	TYR 143	**TYR 143**	PHE 149
	TYR 262	TYR 205	TYR 205	TYR 221
	PHE 304	PHE 236	TYR 236	PHE 254
	TYR 382	TYR 277	TYR 277	MET 299
	TYR 412	TYR 305	TYR 306	TYR 327
	PRO 303	PRO 235	PRO 235	PRO 253
	MET 307	MET 239	MET 239	MET 257
Gate Keeper Residues	LYS 83	LYS 32	LYS 32	ARG 34
	LYS 266	LYS 238	LYS 238	GLY 256
	LYS 310	LYS 242	LYS 242	GLY 260

* Residues highlighted in bold are conserved and identified as important in the literature (Theisen *et al.* 2004; Steussy *et al.* 2005; Shafqat *et al.* 2010). The non-bold residues are from the alignment

with Tyr 143, Phe 185, His 233, Asn 275, Ser 307 (*Staphylococcus aureus*, PDB ID 1XPK). Tyr 143 and Phe 185 residues are in hydrophobic contacts with the ligand as depicted in Figure 3.15(A). In the MupH these residues are mutated to Phe 149 and Thr 202 respectively, however they are still in contact (Figure 3.15(B)). The others are conserved, which may help to constrain the acetyl moiety and the rest of the substrate in the PKS bound intermediate. It was hypothesised that the orientation of the acetylated cysteine in the *Staphylococcus auereus* (PDB ID 1XPK) HMG-CoA synthase structure helps to form a hydrogen bond between the backbone amide of Ser 307 and the carbonyl oxygen of the thioester for stabilizing the oxyanion formed during the transfer of acetyl to the catalytic cysteine (Figure 3.1) (Theisen *et al.* 2004). In MupH this Ser 329 is also conserved which suggests a similar role.

The HMG moiety of HMG-CoA occupies the catalytic pocket such that its β-hydroxyl group hydrogen bonds to His 301 as in Human HMG-CoA synthase structure (PDB ID 2WYA) and Asn 380 while the terminal carboxyl interacts with Glu 132 and the backbone amide of Ser 414 (Shafqat *et al.* 2010). In Figure 3.15(B), MupH His 251 and Asn 297 can be seen forming hydrogen bonds with the β-hydroxyl group of the ligand. However, since there is no terminal carboxyl in the mupirocin intermediate close to the corresponding Glu and Ser there is no interaction found.

3.2.4.2 Tunnel residues

A set of hydrophobic residues populates the middle portion of the active site **tunnel** forming a hydrophobic lining. No direct hydrogen bonds are made between the enzyme and the CoA as it passes through this hydrophobic sleeve. MupH is also found to have similar residues (Figure 3.16) and the homology model indicates a hydrophobic interaction with the ligand along the tunnel sleeve (Figure 3.15(A)). These residues seem to interact with the phosphopantetheine arm of the CoA specifically, as the gate keeper residues are responsible for stabilizing the nucleotide moiety in the CoA. In avian HMG-Coa synthase, several of these residues have been mutated to leucine with only modest changes in enzyme kinetics.

3.2.4.3 Gate keeper residues

In the standard HMG-CoA enzymes, the solvent exposed outer edge of the active site tunnel is populated by a number of basic residues, Lys 32, Lys 238, Lys 242 (PDB ID 1X9E), which hydrogen bond with ribose phosphate from the CoA moiety but not the phosphopantetheine moiety. These residues are not found to be conserved in MupH as shown in the alignment (Figure 3.16). This variation in the gate keeper residues may be because the likely substrate of MupH, unbranched monic acid is directly attached to the catalytic serine of the ACP via the phosphopantetheine arm and lacks the nucleotide moiety and associated phosphates that are found in CoA.

Figure 3.17 represents the sequence alignment between the MupH orthologs as found by PSI -blast search against the NCBI's protein database for sequences with sequence similarity above 60% and with keyword polyketide in the query, to make sure that the sequences are true MupH orthologs. The alignment highlights all the key conserved residues as found in the alignment of HMG-CoA orthologs. The alignment also highlights the conservation of the residues among the MupH orthologs which are found to be mutated in the HMG-CoA ortholog alignment.

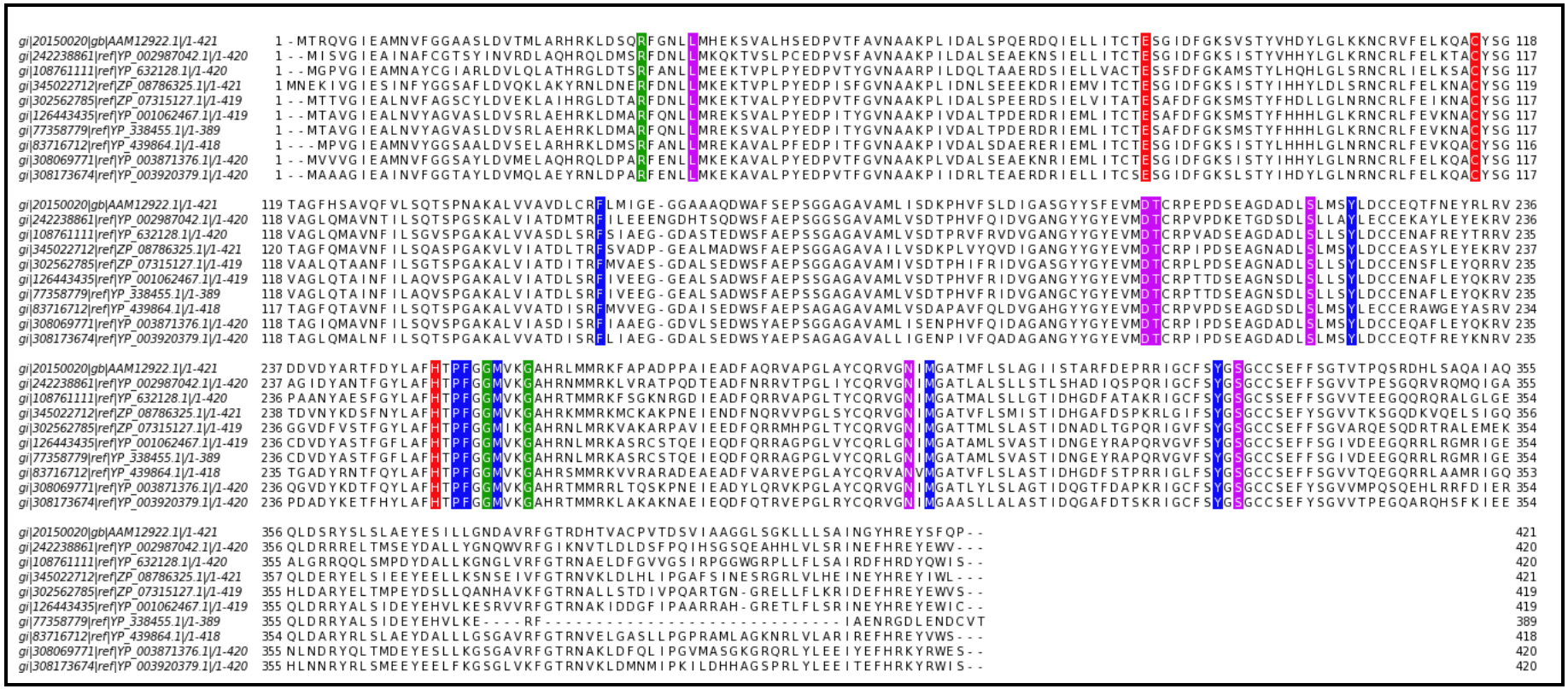

Figure 3.17: *Sequence alignment of the MupH orthologs. Red: Catalytic triad, Magenta: residues responsible for the substrate orientation in the active site, Blue: Tunnel residues and Green: Gate keeper residues.*

3.2.5 Proposed MupH reaction mechanism

The similarities between the sequence and structure of HMG-CoA synthase and MupH, and the conservation of key catalytic residues, may suggest a similar reaction mechanism for both enzymes. The acetyl moiety attached to mAacpC condenses with the β-ketothioester moiety bound to PKS (i.e. mACP3a and mACP3b). In the proposed mechanism, MupH His 251 (catalytic base) seems to attack catalytic Cys 115 which in turn would attack the carbonyl carbon of Ac-mAcpC, thereby transferring the acetyl group to the Cys-S atom and releasing the mAcpC-SH.

In the second step, the methyl group of acetylated-Cys is deprotonated by the general base Glu 83 to form a carbanion which attacks the β-carbonyl of the incoming thioester moiety of ACP-mupA3a/b bound intermediate (second substrate), followed by the condensation of Ac-S-Cys and β-ketothioester moiety of the PKS bound intermediate, which forms an enzyme:glutaryl thioester. Ser 329 seems to play a primary role in the formation of an oxyanion hole that stabilises the tetrahedral intermediate in this reaction (Wu *et al.* 2007). The resultant enzyme:glutaryl thioester is hydrolysed to release the product glutaryl thioester and regenerate the reduced cysteine. Glu 83 mediates the hydrolysis step.

The glutaryl thiolester is dehydrated by MupJ thus producing glutaconate intermediate, which on MupK mediated decarboxylation gives the 3-methylbut-2-enoyl thioester moiety in monic acid precursor (Figure 3.18).

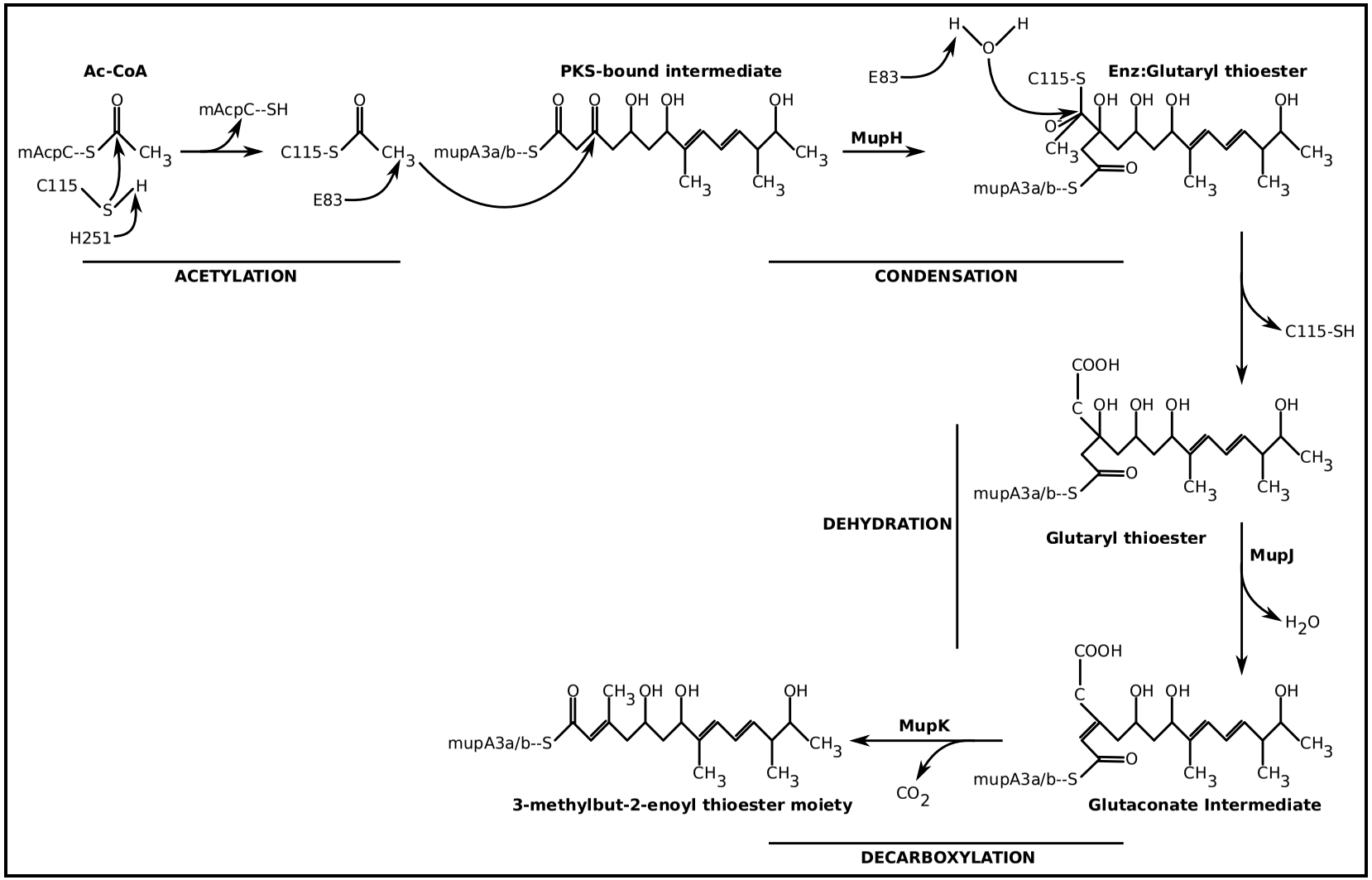

Figure 3.18: *Proposed reaction mechanism of MupH in the HCS cassette.*

3.2.6 MupH and ACP (mupA3a and mupA3b) interaction

The modelled MupH and the NMR determined ACP structures were docked to predict the probable interaction between them. The HADDOCK (Vries *et al.* 2010) program was used to carry out the docking analysis in two different ways (as described in Section 2.3.4.1). In the first method HADDOCK utilizes a set of active and passive residues in which it aims to maximise the interaction of each active residue with as many of the atoms of all passive and active residues as possible. Active residues should have a high degree of evidence that they are at the interface of the complex, e.g. residues with high chemical shift perturbation during NMR titration. Passive residues are typically residues with weaker evidence of involvement in the interface, or surface exposed residues neighbouring the active residues. The definitions thus make no presupposition that an interaction will occur and the system of ambiguous interaction restraints (AIR) is used to optimise the interaction of the active residues with all other residues (see Section 2.3.4 for details).

The active residues were the catalytic serine (Ser 38) of ACP-mupA3a and Arg34, Leu 39, Glu 154 and Ala 214 of MupH, these residues were chosen based on the expected position in the MupH acive site tunnel opening of the phosphopantetheine, which must be covalently bound to serine of ACP-mupA3a. Passive residues were residue contiguous with the active residues and predicted by the program PIER as being part of the interacting interface (Table 3.4) . PIER uses the physicochemical properties of the atomic groups at the protein surface. Here all the residues for the ACP-mupA3a and MupH which had a PIER Value of 30 or above were defined as interface residues. Figure 3.19 shows the docked state of MupH and ACP-mupA3a along with the ligand inside the active site of MupH.

In the second method the docking of each of the ACP-mupA3a and ACP-mupA3b with the MupH ligand complex was carried out using the HADDOCK webserver with a distance restraint of 2.0 Å between the phosphorous of phosphopantetheine bound in the active site of MupH and the OG of the serine (S38/142) residue of ACP-mupA3a or mupA3b. To anchor the ligand in the active site, an additional restraint of 9.13 Å was placed between the sulphur of the thioester linkage in the ligand and the Cα of the catalytic Cysteine (C115) of MupH.

Table 3.4: *Interface residues as predicted by PIER and used in HADDOCK as Active and Passive residues.*

ACP	MupH
Active	**Active**
SER 38	ARG 34, LEU 39, GLU 154, ALA 214
Passive	**Passive**
ALA 15, MET 17, LEU 18, TYR 19, ILE 40, TYR 62, THR 63	LEU 30, PHE 35, LEU 38, GLY 153, GLY 155, GLY 156, SER 168, GLY 212, ASP 213, ASP 215, SER 217, LEU 218, PHE 254, MET 257

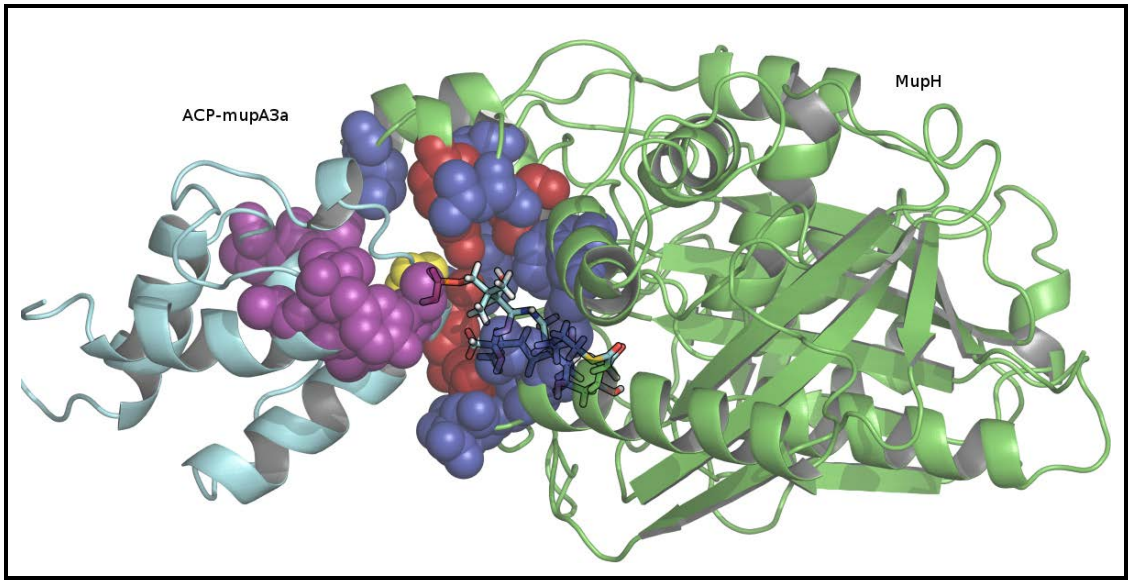

Figure 3.19: *The ACP (cyan) - MupH (green) complex structure predicted by HADDOCK. Yellow: active residue on ACP, Purple: passive residues on ACP, Red: active residues on MupH and Blue: passive reisudes on MupH . The ligand (sticks) shows the location and probable orientation of the phosphopantetheine + mupirocin intermediate in the active site.*

HADDOCK produced a single cluster for ACP-mupA3a and 3 clusters for ACP-mupA3b, each cluster represented by four complexes.

3.2.6.1 Interface analysis

The residues at the interface of the ACP and MupH docked complex using distance restraints were determined using a PyMol script (`http://www.pymolwiki.org/`) (Table 3.5). The interacting pairs of residues between ACP-mupA3a/b and MupH were determined using the CONTAC module of WHATIF (Vriend 1990) (Table 3.6). Two of the ACP-mupA3b clusters mimic

the binding seen in the ACP-mupA3a:MupH complex, whilst the other ACP ACP-mupA3b cluster binds to MupH at the opposite side of the active site entrance (a rotation of approximately 180 degrees around the ligand bind site, see figure 3.20). Despite this variation, the residues Ser 38/142, Val 39/143, Asp 59/163, Tyr 62/166, Thr 63/167 of ACPs ACP-mupA3a/mupA3b were found at the interface of all 16 complexes, the latter three residues being on the surface of Helix III, and S38 being the active site serine. A representative complex is shown in figure 3.21 with an emphasis on helix III residues and their corresponding interacting residues Arg 34, Met 257 and Arg 263 on MupH. The central position of helix III of the ACPs in the complex, higher levels of conservation around this helix, and its shifted orientation compared to other ACPs all point to it being critical to the recognition of the ACPs by MupH and its orthologues.

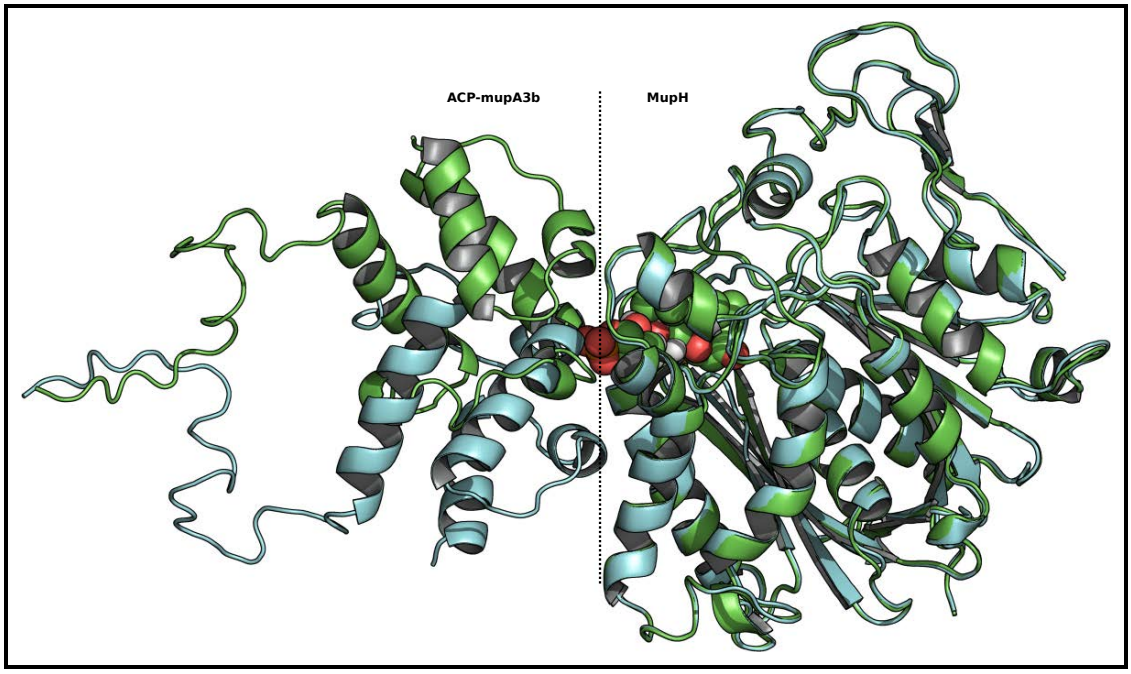

Figure 3.20: *The representative complex of MupH:ACP-mupA3b. The representative complex of MupH:ACP-mupA3b from the cluster 1 (green) with the representative complex from cluster 2 superposed (cyan). The cluster 2 representative is similar in position to the binding orientation of cluster 3 and cluster 4 of MupH:ACP-mupA3b and of the representative cluster of MupH:ACP-mupA3a. The docked states of ACP-mupA3b in cluster 1 are almost 180 degrees rotated around the ligand binding pocket of MupH compared to the other clusters.*

Table 3.5: *List of residues found at the interface of the predicted ACP-mupA3a:MupH complexes and of predicted ACP-mupA3b:MupH complexes*

MupH (total 16 complexes, 4 with ACP-mupA3a and 12 with ACP-mupA3b)		Residues in ACP-mupA3a that are found in cluster 1 of ACP-mupA3b:MupH complexes	Residues in ACP-mupA3b that are found in clusters 2 and 3 of ACP-mupA3b:MupH complexes [a]	Residues in ACP-mupA3b that are found in cluster 1 of ACP-mupA3b:MupH complexes [b]
Residue	*Frequency*			
ARG 34	16	ARG 30	PHE 135	GLY 139
LEU 38	16	LEU 32	GLY 139	MET 140
LEU 30	15	**SER 38**	**SER 142**	**SER 142**
ALA 214	15	**VAL 39**	**VAL 143**	**VAL 143**
ARG 263	15	ALA 42	**ASP 163**	**ASP 163**
ASP 31	14	ARG 46	**TYR 166**	ILE 165
GLN 33	14	ALA 58	**THR 167**	**TYR 166**
ASP 215	13	**ASP 59**		**THR 167**
ASN 37	12	**TYR 62**		PRO 169
MET 219	12	**THR 63**		
LEU 218	11	TYR 64		
GLY 256	10	PRO 65		
GLY 154	9	TRP 73		
GLY 260	9			
ARG 267	9			
LEU 222	6			
MET 257	6			
GLY 153	5			
ILE 152	4			

[a] Residues in common with the interface of the predicted complexes of the ACP-mupA3b:MupH in cluster 1, 2 and 3 and the ACP-mupA3a:MupH cluster are highlighted in bold. In addition all ACP-mupA3b clusters have G139 at the interface.

[b] Each cluster is comprised of 4 complexes, ACP-mupA3b clusters 2 and 3 bind MupH in a similar orientation to ACP-mupA3a, ACP-mupA3b cluster 1 is rotated through 180 degrees around the phosphopantetheine with respect to the other clusters.

Table 3.6: *List of interacting pairs in ACP-mupA3a and ACP-mupA3b with MupH.*

MupH residues	mupA3a residues	mupA3b equivalent	Frequency (Max. 4)	MupH residues	mupA3b residues clusters 2 and 3	Frequency (Max. 8)	MupH residues	mupA3b residues cluster 1	Frequency (Max. 4)
GLN 33	ARG 46	ARG 150	4	ARG 263	ASP 163	8	ARG 34	GLY 139	4
ARG 34	SER 38	SER 142	4	ARG 34	VAL 143	7	**ARG 34**	**SER 142**	4
ASP 215	ARG 30	GLN 134	4	ALA 214	GLY 139	7	LYS 29	THR 167	3
ASP 215	LEU 32	LEU 136	4	GLN 33	ARG 150	6	GLN 33	PHE 135	3
MET 257	TYR 62	TYR 166	4	ASP 31	ARG 150	5	ARG 34	LEU 138	3
LEU 218	LEU 32	LEU 136	3	ASP 215	PHE 135	5	LEU 38	GLY 139	3
GLY 260	TYR 62	TYR 166	3	ASN 37	TYR 166	4	ASP 31	PHE 135	2
ARG 263	ASP 59	ASP 163	3	ASP 215	LEU 136	4	ARG 28	TYR 166	2
ARG 263	THR 63	THR 167	3	**ARG 34**	**SER 142**	3	LEU 30	TYR 166	2
ARG 267	TRP 73	GLU 177	3	ASN 37	ASP 163	3	ARG 34	TYR 166	2
MET 219	ARG 30	GLN 134	2	LEU 38	TYR 166	2	LEU 38	MET 140	2
ARG 263	TYR 62	TYR 166	2	GLU 154	MET 140	2	ALA 214	VAL 143	2
ASP 31	ALA 42	THR 146	1	ALA 214	MET 140	2	LYS 29	PRO 169	1
ASP 31	ARG 46	ARG 150	1	LEU 218	TYR 166	2			
ARG 34	VAL 39	VAL 143	1	LYS 259	ASP 163	2			
ASN 37	ASP 59	ASP 163	1	ASP 31	GLN 147	1			
LEU 38	ASP 59	ASP 163	1	GLN 33	GLN 147	1			
LEU 38	TYR 62	TYR 166	1	ARG 34	THR 146	1			
ASP 215	GLU 33	ASP 137	1	ARG 34	THR 144	1			
LEU 218	TYR 62	TYR 166	1	LEU 38	VAL 143	1			
PHE 254	TYR 62	TYR 166	1	ASP 215	GLY 139	1			
ARG 263	TRP 73	GLU 177	1	ASP 215	TYR 166	1			
ARG 267	GLN 77	ARG 181	1	LEU 216	PHE 135	1			
ARG 267	THR 76	ARG 180	1	MET 257	TYR 166	1			
				GLY 260	ASP 163	1			
				ARG 263	THR 167	1			

Residue pairs underlined are found at the interface of at least one ACP-mupA3a:MupH and ACP-mupA3b:MupH (clusters 2 and 3) complexes. Residue pairs in bold are found in both modes of ACP-mupA3b:MupH interaction.

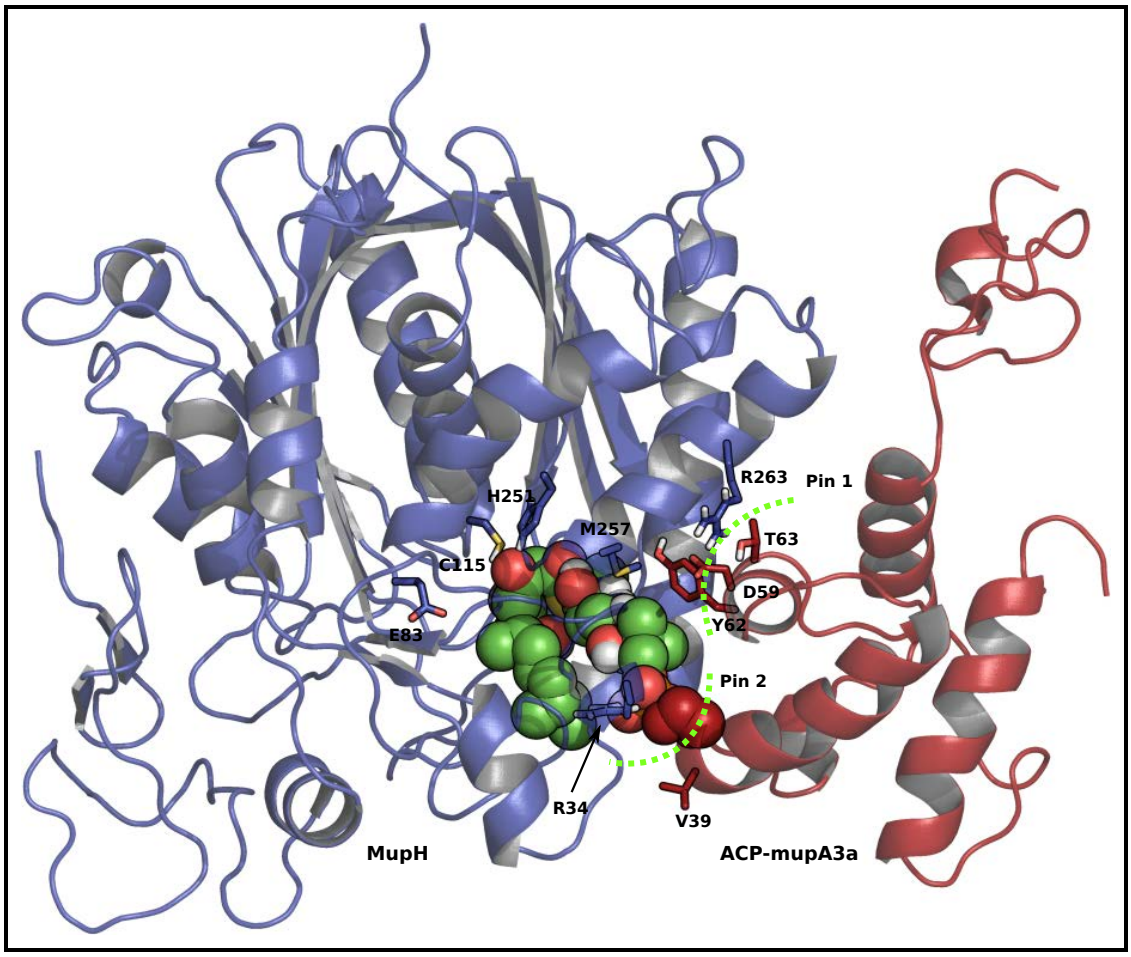

Figure 3.21: *A representative complex with ACP-mupA3a. A representative complex with ACP-mupA3a (red), polyketide intermediate bound to S38 in ACP (spheres) and MupH (blue) showing interface residues on ACP V39, D59, Y62 and T63 and their interacting partners R34, M257 and R263 on MupH. C115, H251and E83 are the catalytic triad in MupH.*

3.2.6.2 Real value evolutionary trace and PIER analysis

It is commonly observed that functionally important residues such as catalytic residues are evolutionary conserved within a protein family. This evolutionary pressure is also seen in the residues which are involved in protein-protein interactions responsible for their function. Therefore, the conservation of the interface residues predicted in the HADDOCK docking experiment was determined using real value evolutionary trace analysis. Real value evolutionary trace analysis (Lichtarge *et al.* 1996) for ACP-mupA3a/mupA3b and MupH was performed using Evolutionary Trace Viewer (Morgan *et al.* 2006) (`http://mammoth.bcm.tmc.edu/traceview/index.html`). This trace gauges conservation within the context of the phylogenetic tree such

that residues that are conserved within a clade but vary between them score highly as do totally conserved residues (details in Section 2.3.3.1). The multiple sequence alignment for real value evolutionary trace analysis was carried out using ClustalW with 95 and 75 sequences for ACP-mupA3ab and MupH respectively. Figure 3.22 and Figure 3.23 show the real value evolutionary trace scores mapped on a representative complex structure of ACP-mupA3a and MupH. The scores are coloured through the spectrum from red to blue, with red indicating predicted importance and blue lack of importance. The ACP is shown as a pink ribbon. The figure shows three views of the same complex; Front View shows the hot (red) patch at the ACP+MupH complex interface. Left view is the clockwise 90 degree rotation from the front view. And the right view is the anti-clockwise 90 degree rotation from the front view. Right View also shows a hot patch at the MupH structure. This hot patch could be a dimer forming interface as the HMG-CoA homologues commonly exit as homodimers, Figure 3.23 shows the structure of the HMG-CoA homologue dimer (PDB ID 1X9E) superimposed, on the MupH+ACP complex. The strong evolutionary conserved patch at the interface supports the existence of MupH as a homodimer.

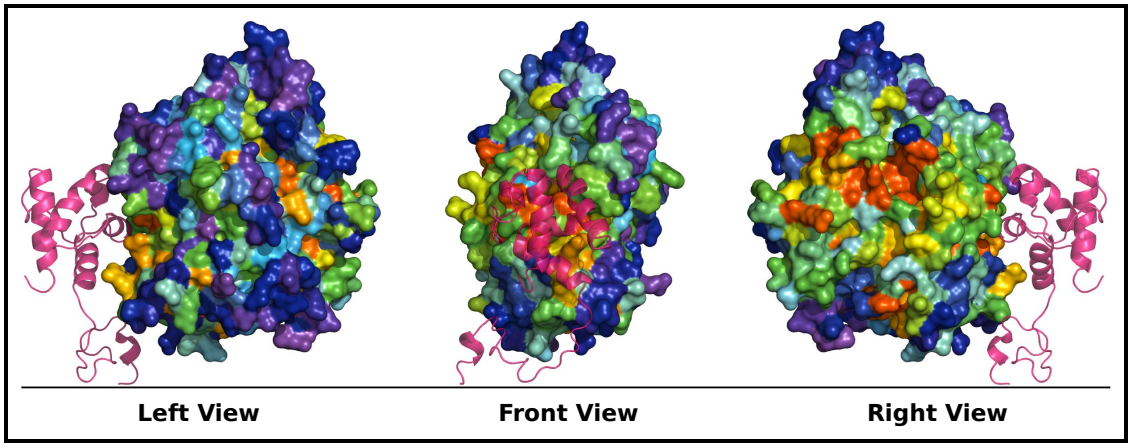

Figure 3.22: *Real value evolutionary trace of MupH homologues. The different view shows the functional patch on the predicted interface of MupH. The left view is a 90 degree rotation anticlockwise with respect to the centre, the right view a 90 degree rotation clockwise. Residues on the surface of the MupH are coloured through the spectrum from red to blue, with red indicating predicted importance, blue lack of importance, the ACP is shown as a pink ribbon.*

The PIER values calculated in the previous step for MupH were also mapped on the structure and coloured through the spectrum from red to blue, Figure 3.24(A). It was interesting to

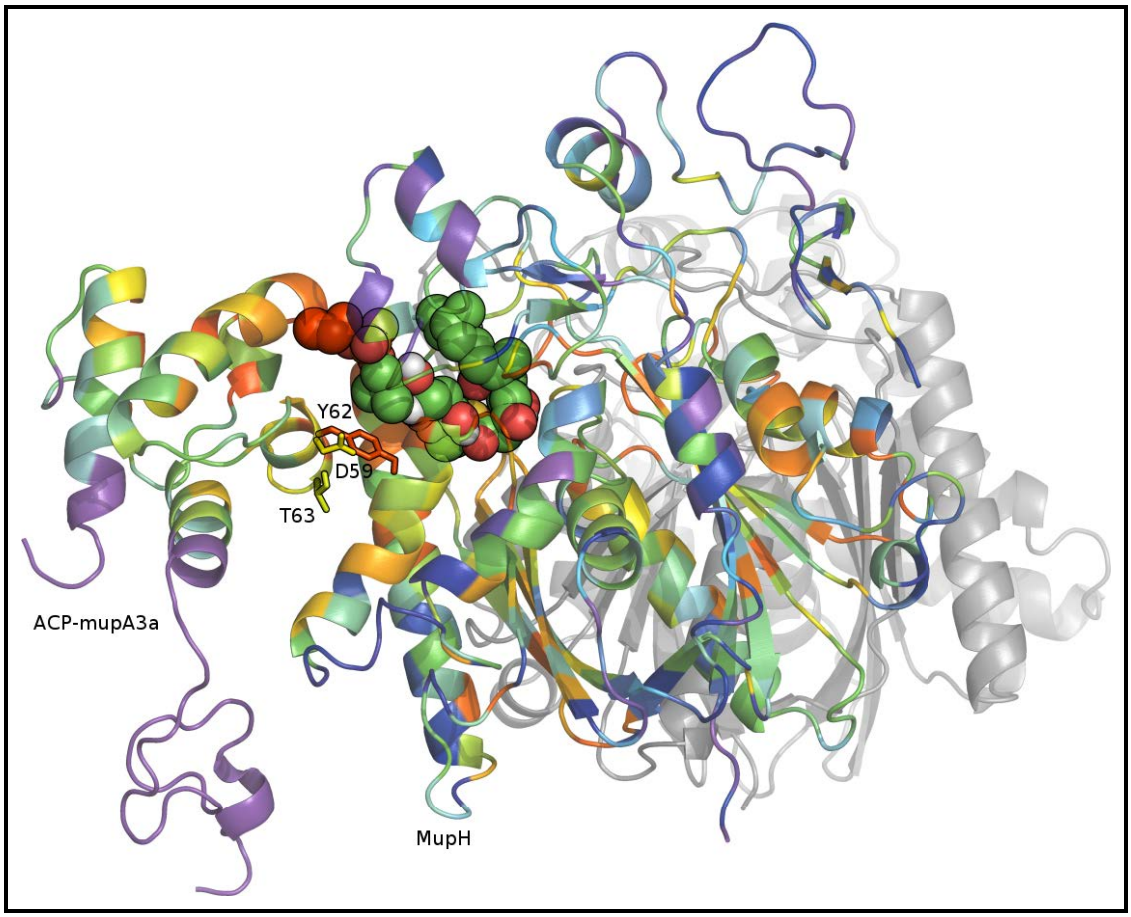

Figure 3.23: *HmgCoA synthase homo-dimer (grey) from Enterococcous faecalis (PDB accession code 1X9E) superimposed on the MupH:ACP-mupA3a complex 1 from cluster 1. Residues are coloured by their real value evolutionary trace score, which highlights residues likely to be of functional importance; a rainbow colouring is used from red (most important) to violet (least important). As well as highlighting the ACP:MupH interface, the analysis also highlights some residues on the MupH surface that are distal from the proposed ACP binding site; these are where most HmgCoA synthases form homo dimmers. (Y 62, D 59 and T 63 are identified as being in the top 5.71%, 22.85% and 23.80% respectively of residues for evolutionary importance).*

see that the residues predicted the important in the real value evolutionary trace analysis also agree with the PIER calculations and similar hot and cold patches can be seen on the surface. However, PIER scores show an additional hot patch on the left view which was not identified in the real value evolutionary trace. For comparison the MupH homologues from kalamanticin (BatC) and thiomarinol (TmlH) clusters were also submitted to PIER analysis Figure 3.24(B) and (C) respective. The structures of BatC and TmlH were predicted using the same template and alignment as that used for MupH and were superimposed on the MupH+ACP-mupA3a representative complex. These complexes also highlighting similar regions are being important for

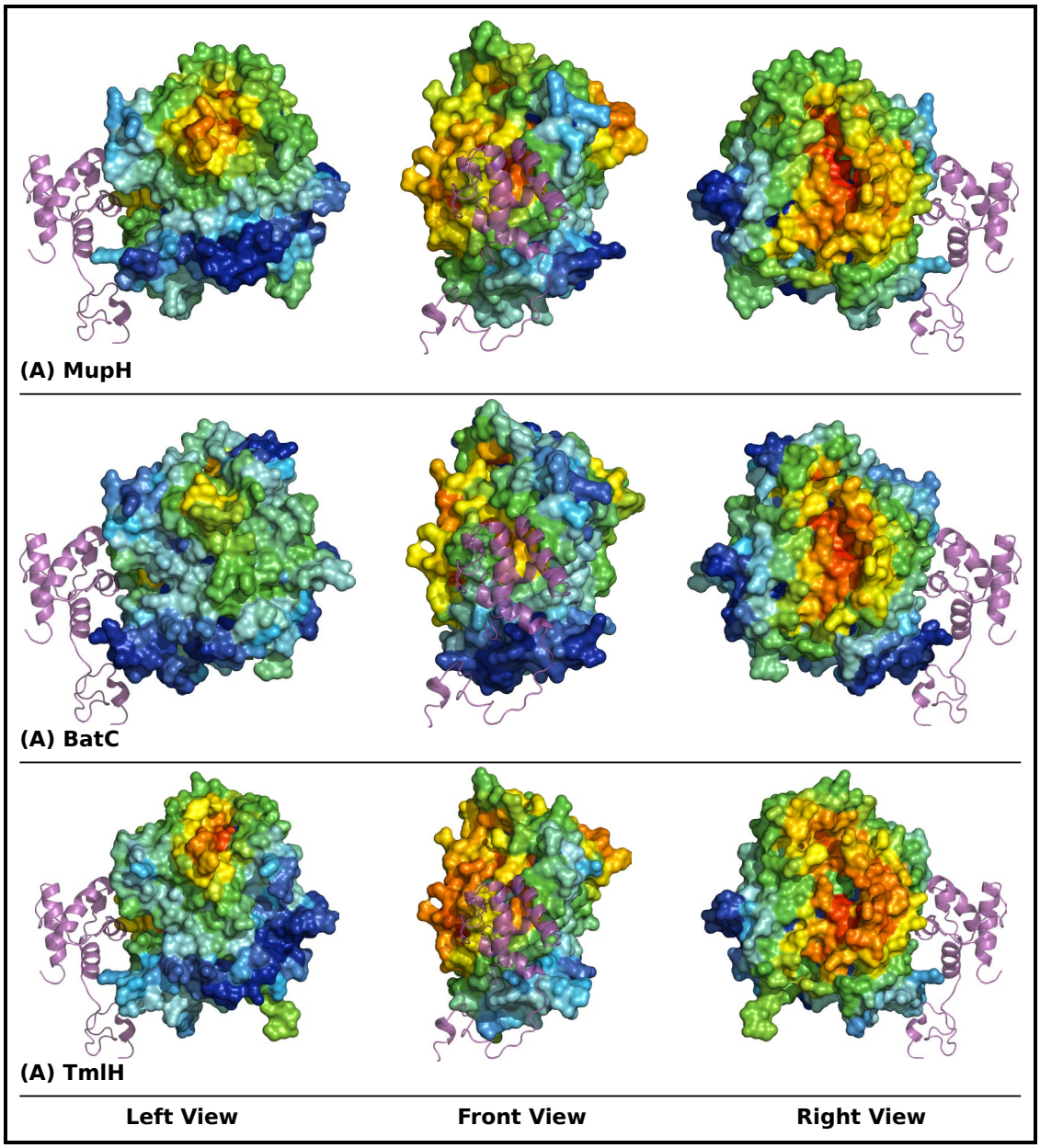

Figure 3.24: *PIER analysis of MupH, BatC and TmlH. The different view shows the functional patch on the predicted interface of (A) MupH, (B) BatC and (C) TmlH. The left view is a 90 degree rotation anticlockwise with respect to the centre, the right view a 90 degree rotation clockwise. Residues on the surface of the MupH are coloured through the spectrum from red to blue, with red indicating predicted importance, blue lack of importance, the ACP is shown as a pink ribbon.*

interface formation, including the additional hot patch of unspecified function.

3.2.6.3 Loss of function with Y to F/A mutation in ACP-mupA3a

The above mentioned docking results which were also supported by PIER and real value evolutionary trace analysis emphasized on the importance of helix III for the formation of the

ACP:MupH interaction. This helix III carried Y62 which was conserved in 95% of the β-branching ACPs and it can be seen pointing towards MupH at the interacting interface and possibly making a favourable methionine to aromatic contact, methionine aromatic interaction were recently reported as important interaction in protein structure (Valley *et al.* 2012). M257 was conserved in all the MupH homologue studied except TaF (myxovirescin system). Mutagenesis experiments were carried out by colleagues in Prof. Thomas group to mutate Y62 to F and A in ΔACP-mupA3b. Phenylalanine was the most commonly occurring residue at this position in non branching ACPs. Mutating Y62 to F and A lowered the mupirocin production by three to four folds and up to ten folds respectively (Haines *et al.* 2013).

3.2.7 BatC complementation failure

BatC is the MupH equivalent protein from the kalamanticin cluster and it was thought that it should be able to complement MupH and produce the beta-branch in mupirocin. However, the complementation experiments showed that *batC* expressed *in trans* in a Δ*mupH* strain greatly decreases mupirocin production. Therefore, to answer the question of why BatC did not complement MupH, bioinformatics and molecular modelling analysis was carried out using the previously modelled BatC structure. Assuming that ACP-mupA3a docks to BatC in the same orientation as it docks to MupH, the BatC structure was superimposed on to the ACP-mupA3a+MupH complex and the ACP-mupA3a+BatC model produced was analysed by CONTAC module from WhatIf package. The contacting residues found were conserved and similar to the ACP-mupA3a+MupH complex (Table 3.7).

Since the contacting pairs of the residues in the BatC+ACP-mupA3a complex were found to be similar to the MupH+ACP-mupA3a complex in the CONTAC analysis it was thought that it is possible that due to superimposing the BatC structure on the MupH+ACP-mupA3a complex the analysis was biased towards the similar positions. In order to address this issue modelled BatC structure was docked to the ACP-mupA3a using the similar distance restraints as it was used for the MupH+ACP-mupA3a complex. The BatC+ACP-mupA3a docked complex came out to be almost 180 degree flip of the MupH+ACP-mupA3a complex. It was not clear why

Table 3.7: *Comparison of the contacting residues in the BatC+ACP-mupA3a pair with the MupH+ACP-mupA3a*

MupH	BatC	ACP-mupA3a
R34	R33	S38
D215	D214	L32
L218	L217	Y62
L219	L218	P65
L222	L221	T63
M257	M256	Y62
G260	G259	Y62
R267	R266	T76

it happened as the BatC interface when superimposed on MupH+ACP-mupA3a complex was visually similar. However, in the MupH docking to ACP-mupA3b one of the clusters out of the four was similar to the BatC docking therefore, the 180 degree flip orientation was plausible but it was not dominating. In BatC docking all the complexes from both the clusters had the same orientation. To answer this question it was thought that it might be the electrostatic potentials on the interface which are different for MupH and BatC. The electrostatic potential shows the electrostatic properties on the surface of a molecule in solution. Therefore, for two molecules to interact with each other they should have complementary electrostatic potential. Similar potentials will obviously repel each other.

To test this hypothesis the APBS (Baker *et al.* 2001) plugin in PyMol was used in conjunction with PDB2PQR (Dolinsky *et al.* 2004). The electrostatic potentials were calculated separately for MupH, BatC and ACP-mupA3a (docked as well as BatC superimposed on MupH+ACP-mupA3a complex) and the structures were superimposed to their respective complexes for comparison. The orientation of the MupH or BatC was kept the same. Three figures were rendered for each of the complex in three different orientations for the solvent accessible and the iso surfaces. The first is the "front view", which shows the interacting interface as flat as possible. The other two are in left and right orientation around Y axis. The degree of rotations used were slightly different for MupH and BatC in order to highlight their interesting features as far as possible.

The electrostatic potentials for both BatC and MupH were found to be very similar with only a few subtle differences. Those few differences might be significant but it was difficult to find any difference in the specificity of MupH and BatC based on the electrostatic potentials on the protein interface.

3.2.7.1 Gain of function with L to M mutation in BatC complementation

The above mentioned docking and electrostatic potential analysis failed to project any striking difference between the interface of MupH with ACP-mupA3a compared to the equivalent interface between ACP-mupA3a and BatC. However, a possible answer to the riddle was in the previously determined contacting pairs of residues between the MupH/BatC and the mup ACPs. CONTAC analysis was carried out twice using a different VDW distance to define a contact each time (i.e. 0.25 Å and 1.25Å). Upon mapping those positions on a sequence alignment of MupH homologue from well studied systems it was revealed that the position 219 on the MupH sequence alternates between methionine and leucine an (observation made by Prof. C. M. Thomas). Interestingly TmlH, which successfully complements MupH, carried a methionine whereas BatC which failed to complement MupH carried a leucine. Looking back at the docked structure it was found that this position was at the edge of the interface and was found to be interacting with the arginine 30 of the ACP-mupA3a in 2 complexes out of 4. This methionine was not found in any of the 12 complexes with ACP-mupA3b. Figure 3.28 shows the position of the residues found in the CONTAC analysis as well as the key residues interacting with the ligand in the MupH active site. Mutating BatC M219L expressed in trans in the mup cluster restored mupirocin production (Haines *et al.* 2013).

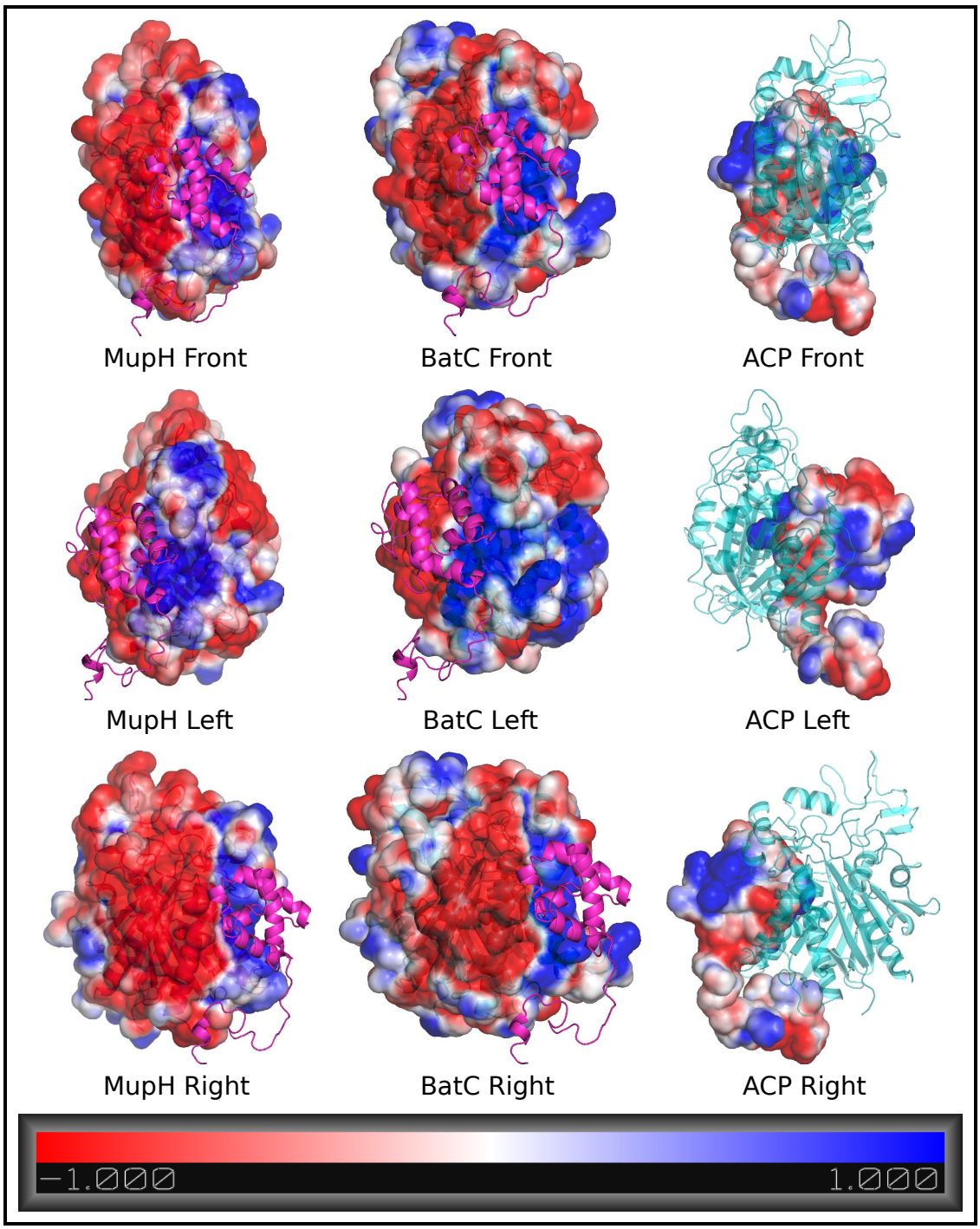

Figure 3.25: *Electrostatic potential mapped on the solution accessible surface of MupH, BatC and ACP. BatC was superimposed on the MupH + ACP-mupA3a complex. The three different orientations (Front, Left and Right) were made to clearly show the mapped potentials at the interface. The scale at the bottom shows the intensity of the negative (red) and positive (blue) potential.*

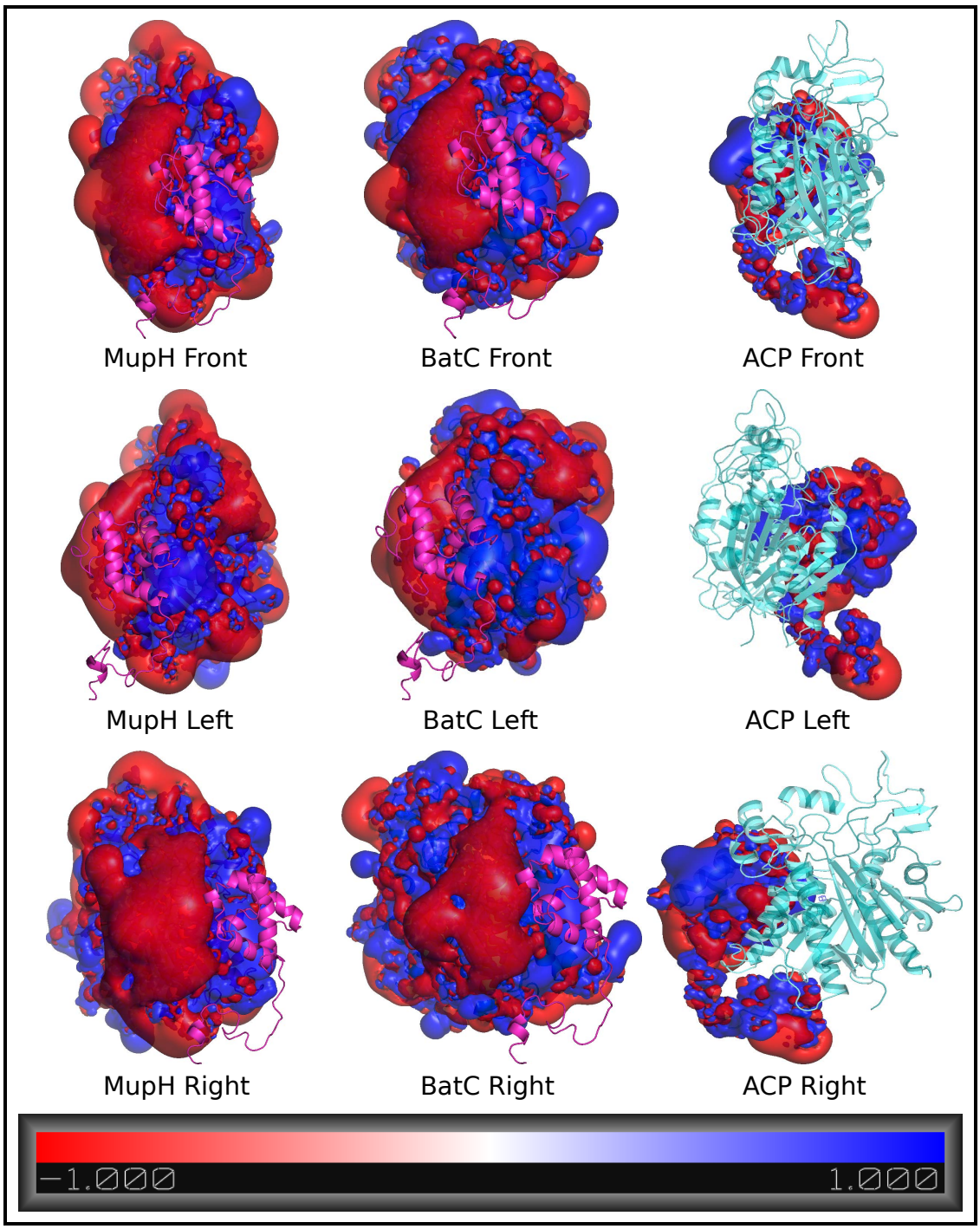

Figure 3.26: *Electrostatic potential mapped as the iso surface on MupH, BatC and ACP. BatC was superimposed on the MupH + ACP-mupA3a complex. The three different orientations (Front, Left and Right) were made to clearly show the mapped potentials at the interface. The scale at the bottom shows the intensity of the negative (red) and positive (blue) potential.*

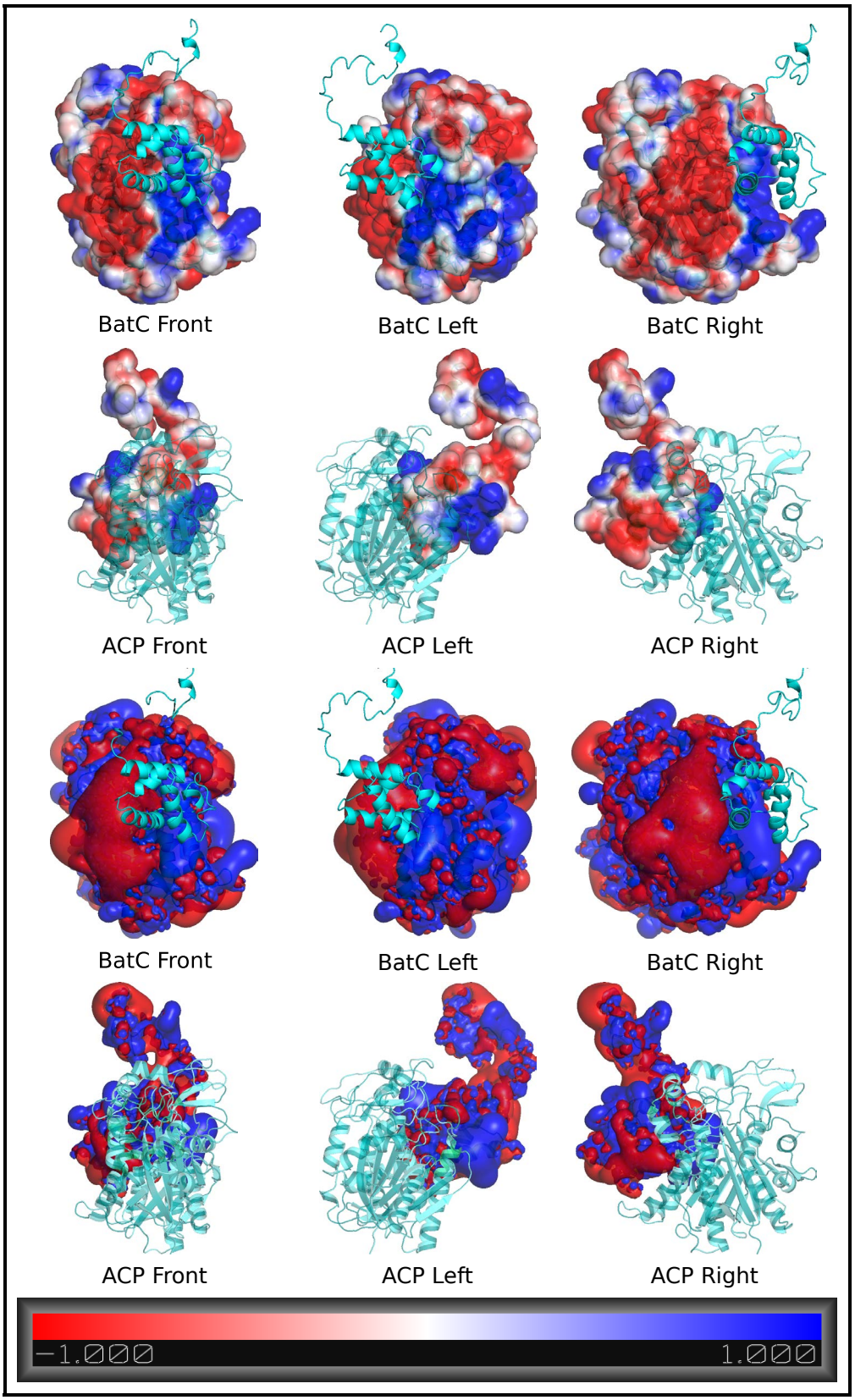

Figure 3.27: *Electrostatic potential mapped on the BatC and ACP-mupA3a docked complex. The orientation of the BatC is kept same as that of the previous figures while the ACP can be seen as docked in almost 180 degree flip orientation. Both the solution accessible surface as well as the iso surface are shown in three different orientations (Front, Left and Right). The scale at the bottom shows the intensity of the negative (red) and positive (blue) potential.*

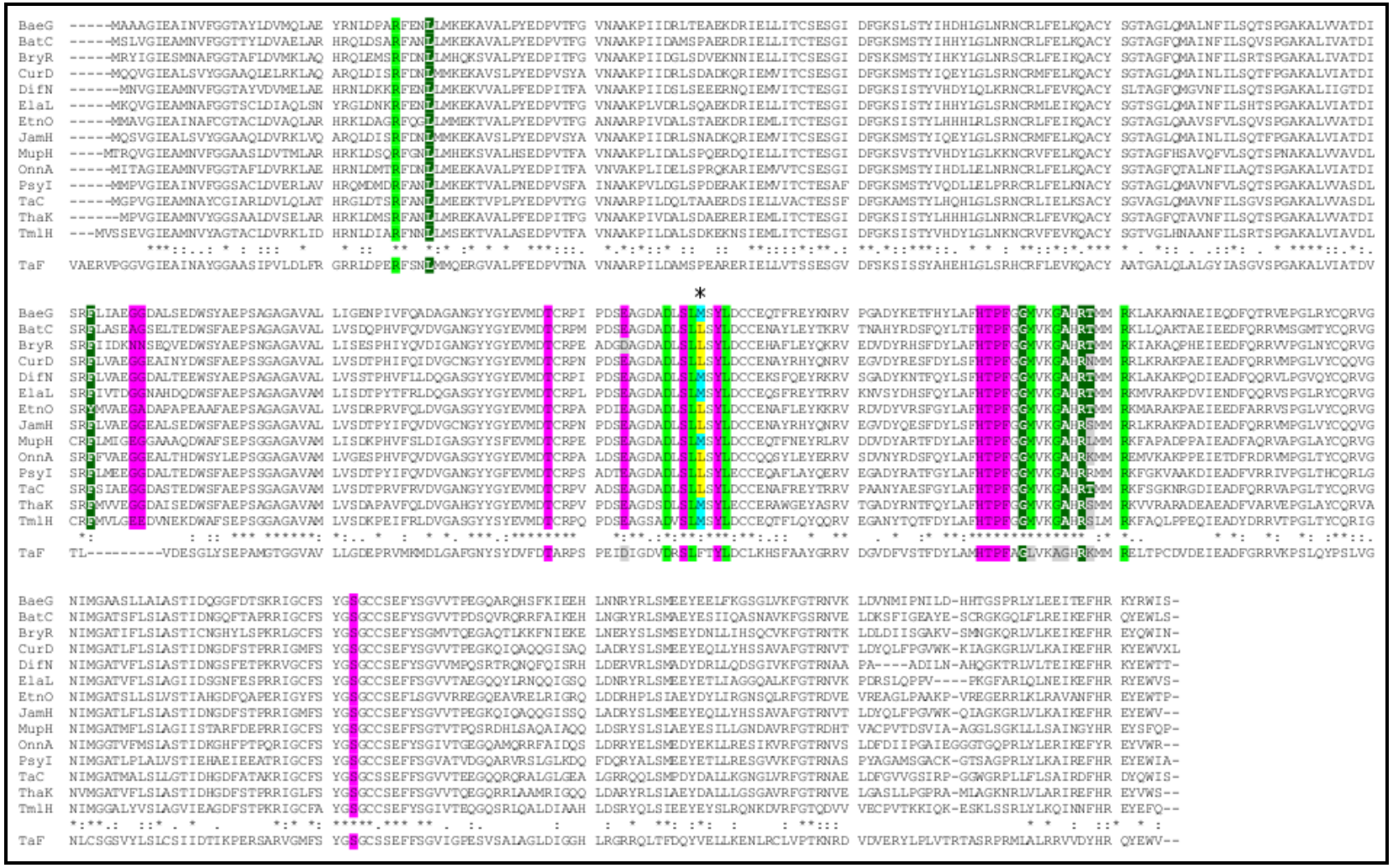

Figure 3.28: *Sequence alignment of MupH homologue from well studied clusters highlighting conserved interface and active site residues. Completely conserved interface residues with VDW distance of 0.25 Å are highlighted in* light green *and residues which alternates between two residues are highlighted as* yellow / cyan*. Residues which are with VDW distance of 1.25 Å are highlighted as* dark green with white text*. The residues which are not on the interface but are in contact with the ligand are highlighted as* magenta*. The * symbol marks the M219.*

3.3 Discussion

The aim of the present work was be able to identify β-branching protein in PKS cluster and to elucidate the specificity mechanism involved in the ACP-HCS interaction, with a longer-term goal being to selectively introduce β-methylation or other HCS modifications such as cyclopropane, as part of the re-engineering of polyketide biosynthetic pathways for novel purposes.

In the mupirocin biosynthesis pathway β-branching is initiated by the interaction between mupA3ab tandem ACPs in module 5 of the MmpA subunit, with MupH, (the HMG-CoA homologue) in the HCS cassette. However, little was known about what governs this interaction. Sequence analysis by Dr. Anthony Haines showed that branching ACPs possess a conserved tryptophan six residues downstream of the catalytic serine but non-branching-ACPs do not. NMR studies on the tandem ACPs from the mupirocin biosynthesis pathway showed that the tryptophan and other highly conserved residues lie in the core (including L10/114, L14/118, L18/122 and F31/135) of the ACPs rather than on or near the surface. These observations raised the question as to what determines the specificity between the branching-ACPs and the MupH in the HCS cassette? Is it just the core motifs which are responsible for this specificity, if so how, or do the residues at the ACP-HCS interface also play some critical role?

In the present work, based on the previous preliminary data on the sequence analysis carried out by Dr. Anthony Hains and the NMR structures solved by Dr. Matthew Crump, further sequence analysis and molecular modelling was carried out to elucidate the specificity determinants in ACP-HCS interaction. Hidden Markov models (HMMs) were created to describe the set of branching-ACPs and the non-branching-ACPs, and were used to predict the clustering pattern between the two classes. It was seen that the HMMs were able to separate the branching-ACPs and the non-branching-ACPs into two distinct clusters with very few exceptions. The HMM models were also used to fetch more sequences from the public databases and the newly found sequences were characterised as branching or non-branching. These HMM models can be used for the annotation of newly found PKS clusters and the β-branching ACP HMM is now in corporated into the SMART database. The HMM models were also used to

calculate the number of mutations that would be required to shift a non-branching-ACP to a branching-ACP cluster i.e. across the HMM score of 82. TmlD3-ACP from the thiomarinol cluster would require 6 mutations, with the mutation of valine to the tryptophan found conserved in the branching-ACPs being the first change.

The observation that the conserved tryptophan lies in the conserved core suggests that it might be responsible for the packing and stability of the ACP structure. Molecular dynamics simulations for the wild type and the mutant ACP in which the tryptophan was mutated to leucine showed relatively high flexibility in and around helix III in the mutant as compared to the wild type ACPs. Mutation studies carried out on T4 lysozyme structures have shown various general properties of protein folding and stability. Mutation of residues packed in the core of T4 lysozyme lowered the melting temperature of the protein whereas changes in the surface residues had little effect. These core residue changes were from "hydrophobic to charged" residue, M102K (PBB ID 1L54), "small to large" residue, A98F and A98W, "large to small" residue, L99G (PDB ID 1QUD) and R95A disrupting various interactions made by arginine (Rennell *et al.* 1991; Tokuriki and Tawfik 2009). A "large to small" residue change in the core of the T4 lysozyme seems to be similar to the W44L change in the β-branching ACPs. And so this suggests that we would expect W44L to destabilise the structure. Computational studies on T4 lysozyme structures also showed that the mutations primarily caused backbone shifts rather than changes in the side chain rotamers (Hurley *et al.* 1992; Dahiyat *et al.* 1997; Mooers *et al.* 2003). Molecular dynamics simulations carried on β-branching ACPs have also showed increased flexibility in the backbone atoms of the helix III.

Docking analysis of ACPs with MupH revealed this helix III to be at the ACP-MupH interface. Helix III contains a tyrosine at the position 62 in ACP-mupA3a, which is highly conserved amongst β-branching ACPs, and which was found to interact with a methionine in a cleft at the MupH interface. This interaction is supported by a recent study done on methionine and aromatic residue interactions and their effect on protein structure stability (Valley *et al.* 2012). Helix III in the HCS interacting ACPs was also identified to be important for halogenase activity in the curacin system (Busche *et al.* 2012), also mediating an interaction with an HMG-CoA

homologue. Experiments done by colleagues in Prof. Christopher M. Thomas' group mutated Y62 on helix III in ACP-mupA3a to F and A in the ΔACP-mupA3b strain and showed reduced mupirocin production by three to four fold and ten fold respectively.

Cross complementation experiments from Prof. Christopher M. Thomas'group found that MupH orthologues TmlH and BatC from thiomarinol and Kalimanticin clusters respectively do restore MupH activity in an NCIMB 10586 Δ*mupH* strain. However, the pseudomonic acid yield by BatC complementation is much lower than TmlH. Analysis of the predicted MupH:ACP interface residues, along with sequence analysis by Prof. Chris Thomas, suggest M219 as a specificity determining residue, being M in MupH and TmlH but L in BatC. *batC* L218M was found to complement Δ*mupH*, with 3 fold more antibiotic production than with wild type *batC* complementation. This observation leads to a further question, if there exists a pair wise specificity between the branching ACPs and the HCS as well as a general determinant of β-branching ACPs the global properties of branching ACPs then replacement of ACP-mupA3ab with the ACP(s) from the kalimantacin system should either fail completely or perform poorly. However, as the mutation in the BatC helped to enhance the favourable interaction with the mupirocin ACPs analogous mutations on the kalimantacin ACP or on the mupirocin ACP should be able to help from favourable interactions between them and MupH or BatC respectively.

Thus it can be concluded that the conserved core residues in the branching-ACPs helps in the correct packing of the branching ACPs. This packing helps the active site serine to present the substrate to the MupH active site and directing the angle of helix III which acts as another anchor at the interface. This two pin model in which the orientation of the active site serine acts as one pin and the tyrosine on the helix III another pin provides the general specificity to the ACP-MupH interaction. However, other residues found at the interface may be responsible for ACP-HCS pair wise specificity as seen in the two different ACP-HCS pairs in the myxovericin system.

Chapter 4

Kalimantacin ACP swap in mupirocin cluster

4.1 Introduction

The predicted structure of the ACP-MupH complex, described in the previous chapter, highlighted various key residues important for the ACP-HCS recognition required for β-branching. Tyrosine 62 on helix III in ACP-mupA3ab was predicted to be at the interface of the interacting models. Mutating Y62 to F or A in the ΔACP-mupA3b strain reduced mupirocin production by four and ten fold respectively, as determined by HPLC. This loss in function confirmed the importance of helix III of the β-branching ACPs for the formation of a functional complex with an HCS protein. However, later experiments showed reduced complementation by *batC* (kalimantacin cluster) in a Δ*mupH* strain. Interface residues identified in the ACP-MupH complex included residue 219 on MupH which was a methionine in MupH and TmlH (thiomarinol cluster) but a leucine in BatC. BatC L218M has an improved ability to complement Δ*mupH*, with three fold more mupirocin production compared to wild type BatC. This observation suggested that although the β-branching ACPs usually have a conserved tryptophan in the core, and a Y on helix III forming the interaction interface, there exists ACP-HCS pairwise subtype specificities. This pairwise specificity can be observed in the myxovirescin system where there are two sets of HCS cassettes that interact with their cognate ACPs.

Extending the observation that *batC* failed to complement Δ*mupH* in the mupirocin cluster due to a difference in the interaction specificity, it seemed that swapping ACP-mupA3ab in the

mupirocin cluster with the β-branching ACP(s) from the kalimantacin cluster should also fail. However, it should be possible for complementation of ACP-mupA3ab by kalimantacin ACP(s) upon changing the residues at its interface, or if the kalimantacin ACPs were expressed together with the wild type *batC* in the mupirocin system. Thus here, the kalimantacin ACP (ACP-K24a; K=kalimantacin, 2=second protein, 4=fourth module, a=first ACP) was amplified from a purified DNA sample obtained from our collaborators. The idea behind selecting ACP-K24a was that it was one of tandem ACPs similar to the branching ACPs in the mupirocin cluster and it seemed to be performing only the addition of a methyl branch without any other modifications that were present in the other β-branching sites in the kalimantacin cluster. Two host strains, *P. fluorescens* Δ*acp4* and *P. fluorescens* Δ*mupH* were used. The *P. fluorescens* Δ*mupH* strain lacks *mupH* in the HCS cassette and was used to express *mupH*, *batC* and *batC* L218M *in trans*, described in detail in section 2.4.1.

4.2 Results

4.2.1 Plasmid preparation and transfer for Suicide Mutagenesis

4.2.1.1 Amplification of DNA fragments for ligation into a pAKE 604 suicide plasmid

Replacement of the branching ACPs in the mupirocin pathway with kalimantacin ACP-K24a was through homologous recombination via suicide mutagenesis. A pAKE604 plasmid was used to clone DNA fragments from the mupirocin and the kalimantacin cluster. To ensure that sequence amplified for ACP-K24a includes all the secondary structures for an ACP, a sequence length equivalent to the length of the solved ACP-mupA3a structure was taken for PCR primer construction. For homologous recombination to happen ACP-K24a needs to be attached to the ≈500 nucleotide long arms either side of the mupirocin cluster ACP-mupA3ab. The left and the right arms were amplified from both the ends of the ACP-mupA3a sequence using the *P. fluorescens* Δ*acp4* strain as the template. The details of the PCR setup and the primers used are described in section 2.4.4.

The amplified fragments were analysed using 1% agarose gel electrophoresis and were pu-

rified using the GE Healthcare Life Sciences Illustra-GRX PCR DNA and gel band purification kit. Figure 4.1 shows the purified fragments for the left arm, right arm and ACP-K24a on the agarose gel. The concentration of the purified samples were quantified using the Nanodrop instrument in order to calculate the volume of samples to be used in the Gibson assembly reaction described later. Simultaneously, a pAKE604 vector was cut using two restriction enzymes (HindIII and SalI) at the multiple cloning site and the concentration of the purified product was also quantified using the Nanodrop instrument.

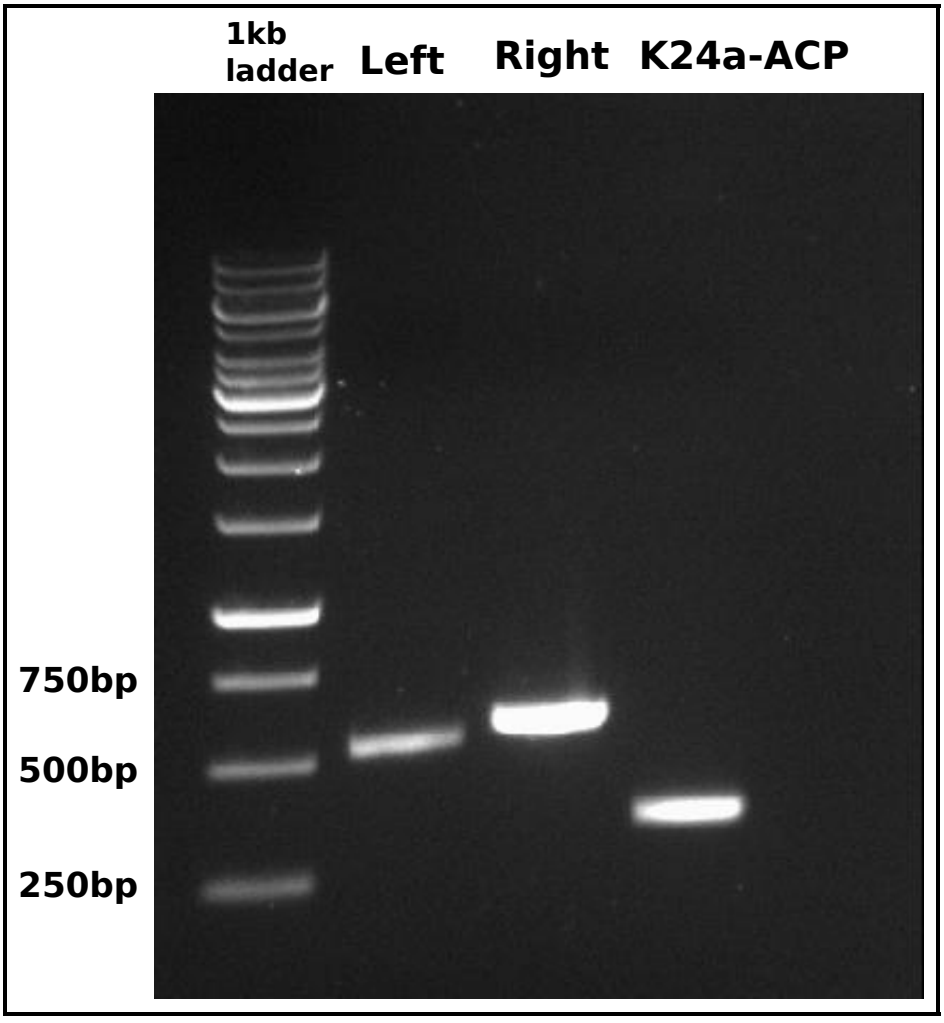

Figure 4.1: *Purified fragments for the left arm, right arm and ACP-K24a on agarose gel. The size of the fragments include the homology regions required for Gibson assembly.*

4.2.1.2 DNA fragments ligation into pAKE604 using Gibson assembly

Gibson assembly is a method to ligate several DNA fragments in a single reaction (details in section 2.4.6). In order to perform the ligation of the ACP-K24a to the two ACP-mupA3ab flanking arms and the multiple cloning site of pAKE604, a Gibson assembly kit from NEB was used. The DNA concentration calculated for ACP-K24a was 68.1 ng/μl, for the left arm it was 50.8 ng/μl, for the right arm it was 98.1 ng/μl and for the digested pKAE604 it was 121 ng/μl. For the ACP-K24a of length 297 the volume required was 0.8 μl, for the left arm of length 487 was 1.5 μl, for the right arm of length 532 was 0.9 μl and for pKAE604 of length 7219 was 9.7 μl. Since the total amount of solution required for one Gibson assembly reaction was 20 μl which include 10 μl of the Gibson assembly master mix the total volume of the three DNA fragments excluding vector was scaled down to 3 μl from 3.2 μl and the volume for the vector was scaled down to 7 μl from 9.7 μl. Volume equal to two Gibson assembly reactions was measured and split into three tubes of 13 μl, 14 μl, 13 μl each and were incubated at 45°C, 50°C and 55°C respectively for one hour. The standard NEB manual protocol did not yield any ligated product so an alternative method for Gibson assembly was tried. The reaction mix prepared for the total volume of two reactions was divided into three tubes, which were incubated at three different temperatures. The three temperatures were 45°C, 50°C (recommend by NEB) and 55°C. The contents of the three reaction tubes were then pooled into a single tube and the mix was then used for transforming the freshly prepare *E. coli* DH5α competent cells (section 2.4.3 and 2.4.7) as recommended in the NEB Gibson assembly protocol. The idea behind pooling all the content into one tube was to reduce the transformations to be done for each temperature, if the Gibson assembly worked for any of the temperatures then the assembled plasmid should be taken up by the competent cells and only the transformed cells would survive the kanamycin selection. A disadvantage of this method is that which temperature(s) actually worked amongst the three remains unknown.

The transformation plates showed fourteen colonies in total which were streaked to single colonies. From the streaked plates fourteen colonies were picked, one for each of the original fourteen streaks, and the PCR was performed using the left arm forward primer and the right

arm reverse primer. Out of the fourteen samples, eleven gave DNA fragments of the required size as analysed on 1 % agarose gel electrophoresis (Figure 4.2). Four of the eleven successful PCR products were purified and two of these were sequenced using primers designed previously by Dr. Joanne Hothersall, which bind to the region outside the multiple cloning site of the pAKE604 plasmid. Of the two samples the forward and reverse reads for the second sample showed the expected sequence. The validated second sample was used to transform freshly prepared *E. coli* S17-1 competent cells.

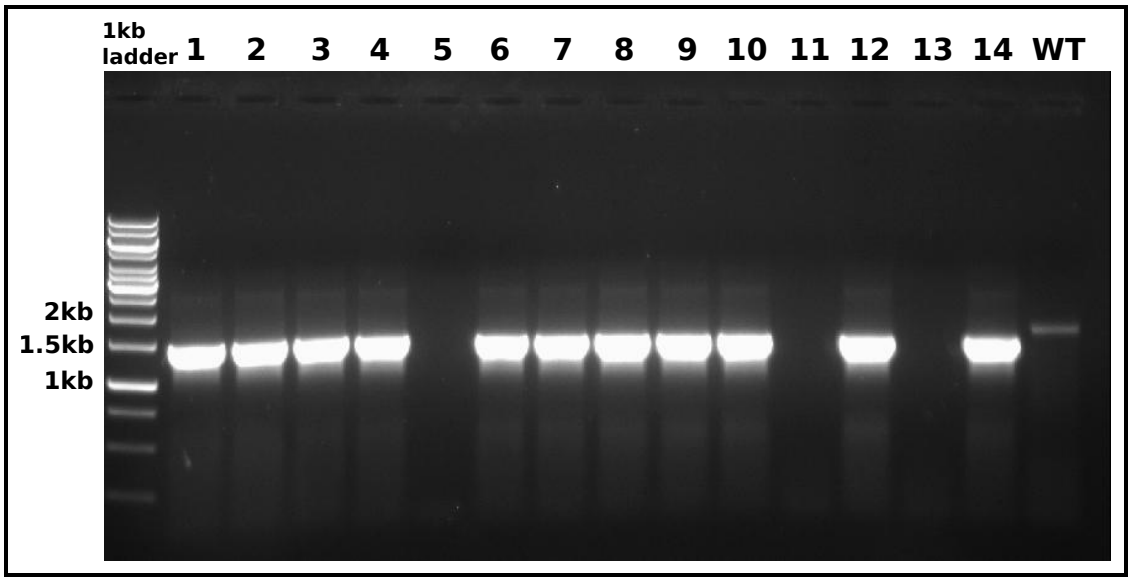

Figure 4.2: *Gibson assembly product on agarose gel. Eleven out of fourteen samples showed the right size band of roughly 1.5 Kb in size. P. fluorescens NCIMB 10586 was used as control and the amplified product contains both ACP-mupA3a/b.*

4.2.1.3 Conjugal transfer of the suicide plasmids into *P. fluorescens* host strains

The suicide plasmid pAKE604 carrying the construct "left arm - ACP-K24a - right arm" in *E. coli* S17-1 was transferred to two different *P. fluorescens* strains through mating. The strains were *P. fluorescens* Δ*acp4* and *P. fluorescens* Δ*mupH* (details in section 2.4.8). Trans conjugants were subjected to PCR using the ACP-K24a specific primers to detect the successful transfer of the plasmid into *P. fluorescens* Δ*acp4* and Δ*mupH* cells.

For *P. fluorescens* Δ*acp4* trans conjugants eight colonies were subject to PCR along with non mated *P. fluorescens* Δ*acp4* and transformed *E. coli* S17-1 strains as negative and positive controls respectively (Figure 4.3). All the colonies as well as the positive control showed the

presence of the plasmid sequence.

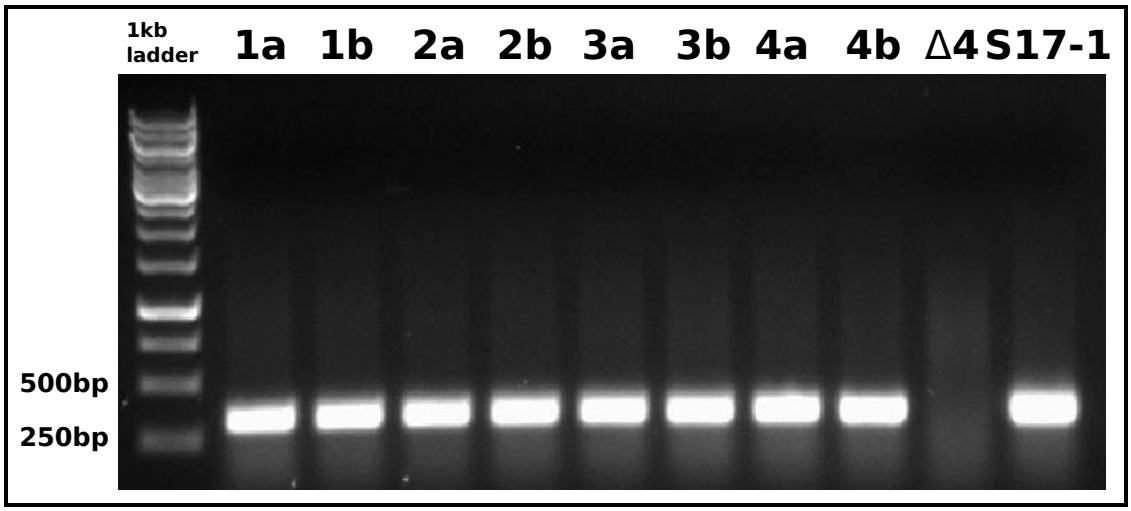

Figure 4.3: *Validation of P. fluorescens Δacp4 trans-conjugants. All the eight colonies as well as the positive control showed the presence of the plasmid carrying the ACP-K24a. No band was detected for the negative control.*

For *P. fluorescens* Δ*mupH* trans-conjugants ten colonies were subjected to PCR along with a transformed *E. coli* S17-1 strain as a positive control (Figure 4.4). Nine out of ten colonies as well as the positive control showed the presence of the plasmid sequence.

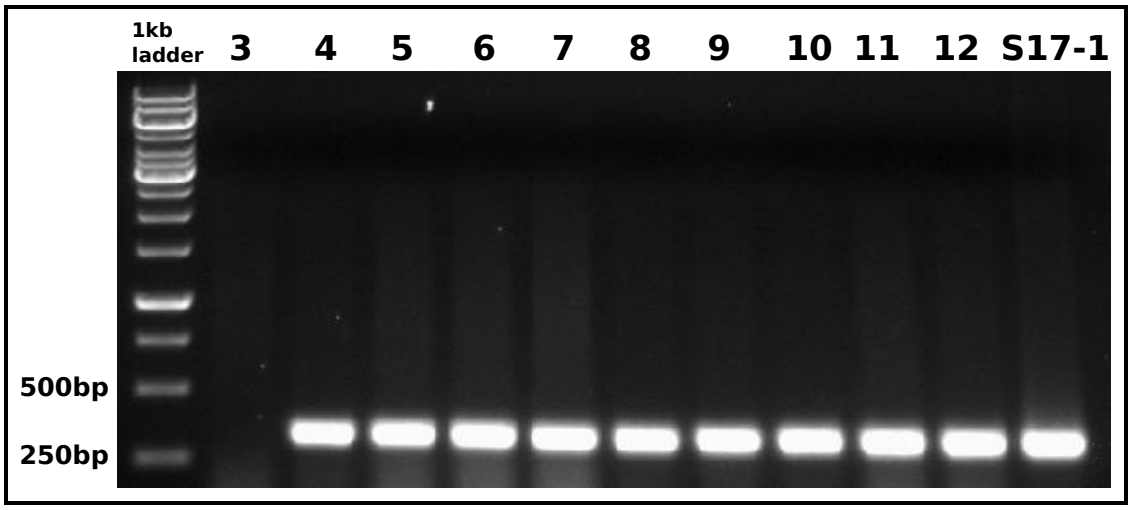

Figure 4.4: *Validation of P. fluorescens ΔMupH trans-conjugants. Nine out of ten colonies as well as the positive control showed the presence of the plasmid.*

4.2.2 Sucrose selection and excisant validation

After mating *P. fluorescens* integrants were grown overnight in L-broth without any selection (see details in section 2.4.9). This would have allowed the integrated plasmid to excise out, leaving behind the desired gene integrated into the chromosome. These overnight cultures were further grown on the L-agar plates with sucrose selection. The pAKE604 plasmid contains a *sacB* gene, which confers sensitivity to sucrose, so that if the plasmids have excised out successfully then the cells would grow, otherwise the *sacB* gene would cause accumulation of the polymer levan in the periplasm which is toxic to Gram negative bacteria. The excision of the plasmid can occur in two ways, by leaving the ACP-K24a behind integrated into the chromosome or by the host chromosome reverting back to its original state. To detect that ACP-K24a was successfully integrated into the chromosome and not taken away at the time of excision, the colonies which grew on ampicillin and not on kanamycin plates were subjected to PCR with ACP-K24a specific primers. For *P. fluorescens* Δ*acp4* excisants, twelve colonies were subjected to PCR and the products were analysed using 1% agarose gel electrophoresis; the right size band was detected in five out of twelve excisants. Assuming that the plasmids excised correctly and were destroyed in the cell, these five bands suggest that ACP-K24a has integrated into the chromosome in the mupirocin cluster (Figure 4.5).

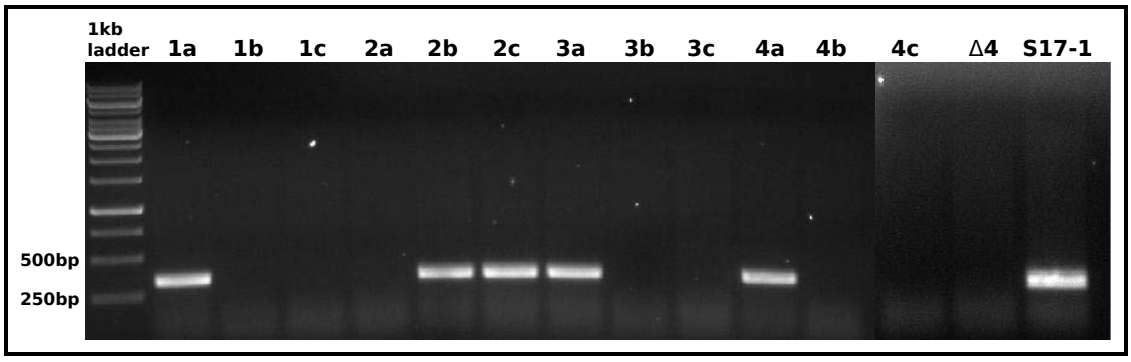

Figure 4.5: *P. fluorescens Δacp4 integrants validation. Five out of twelve colonies as well as the positive control showed the presence of the ACP-K24a.*

The previous steps validated the integration of the ACP-K24a into the *P. fluorescens* Δ*acp4* chromosome but it did not detect whether the integration happened at the correct position in the mupirocin cluster. In order to validate the integration of the ACP-K24a at the correct position

another PCR was carried out on the five integrants. Primers were used that bind to positions outside the left and right arms, which had been previously designed by Dr. Anthony Haines. *P. fluorescens* NCIMB 10586 and *P. fluorescens* Δ*acp4* were used as the controls, with *P. fluorescens* Δ*acp4* expected to give a band equivalent in size to the five integrants since they have a single ACP at the branching position, whereas *P. fluorescens* NCIMB 10586 would give a slightly bigger band because of the two tandem ACPs (Figure 4.6).

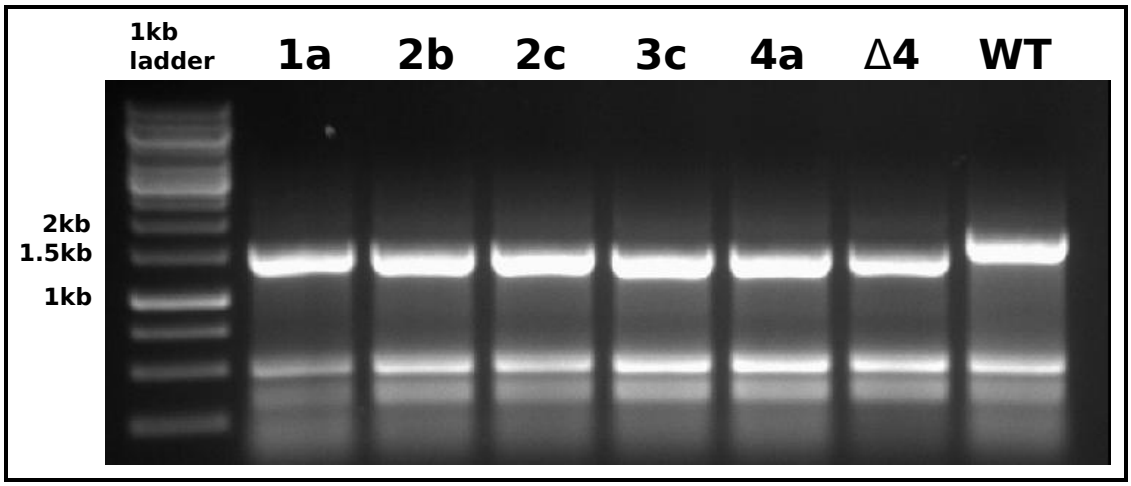

Figure 4.6: *Five P. fluorescens Δacp4 integrants tested for correct location of ACP-K24a integration. The primers bind at locations outside the 500 bp arms thus the five integrants as well as Δacp4, have a band closer to 1.5 kb, which includes two arms of ≈1 kb and an ACP of 331 nucleotides while the wild type P. fluorescens NCIMB 10586 which contains two ACPS and the linker region between them has slightly larger band. The primers also seem to mis-prime some where in the chromosome thus generating a smaller band of ≈500 bp.*

The PCR product for the five integrants were further subjected to restriction digests using enzymes Blp I and Stu I. Blp I cuts approximately at the middle of the ACP-K24a but does not cut anywhere in ACP-mupA3a and Stu I cuts at the middle of the ACP-mupA3a as well as at the rear ends of the two arms. The restriction digest fragments were analysed with 1% agarose gel electrophoresis, an uncut fragment was used as the control. Two samples verified by the restriction digestion of the PCR products were used for sequencing in the forward and reverse direction using the primers used in the previous step. Glycerol stocks of the two verified samples (two replicates for each i.e. Δ 4-1a (tube 1 and 2) and 4-2b (tube 1 and 2)) were stored at -80°C.

For *P. fluorescens* Δ*mupH* trans-conjugants sucrose selection was performed following the

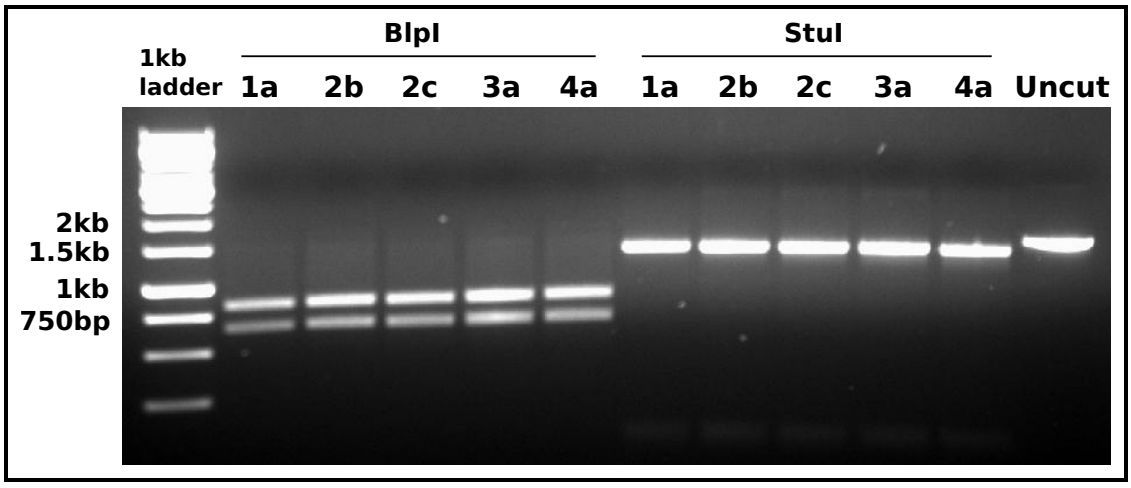

Figure 4.7: *Restriction digestion of the five P. fluorescens Δacp4 integrants using restriction enzymes Blp I and Stu I. BlP I cuts approximately at the middle of the ACP-K24a and Stu I cuts approximately at the middle of the ACP-mupA3a as well as at the rear ends of the two arms. The first five wells shows ACP-K24a cut approximately at the middle whereas the next five shows the fragments which were cut only at the rear ends of the arms. The last well shows the actual size of the uncut DNA fragment.*

same steps as for the *P. fluorescens* Δ*acp4* trans-conjugants. To validate the excisions and integrants, nine samples were subjected to PCR using the primers designed to bind at the region outside the two arms (Figure 4.8), one of which was of the right size and was sequenced. Compared to the treatment of the *P. fluorescens* Δ*acp4* integrants, the steps of PCR with ACP-K24a specific primers and subsequent restriction digestions were excluded, because running a PCR using only the outer primers and sequencing the PCR products gave an equivalent result in fewer steps and less time. Glycerol stocks of the sample (two replicates i.e. Δ H-6d (tube 1 and 2)) were stored at -80°C.

4.2.3 Overlay Bioassay to test for antibiotic production in the constructed strains

To test the two hypothesises that the ACP-K24a from the kalimantacin cluster would not work with MupH but will work with BatC, three previously prepared pJH10 plasmids containing *mupH*, *batC* or the *batC* L218M mutant were expressed *in trans* in the newly constructed *P. fluorescens* ΔH-6d strain. One pJH10 plasmid containing *batC* was also expressed *in trans* in the newly constructed *P. fluorescens* Δ4-1a strain (see details in section 2.4.10). Since, ΔH-6d strain lacks *mupH* in the HCS cassette, the *in trans* expression of any of *mupH*, *batC* or *batC* L218M

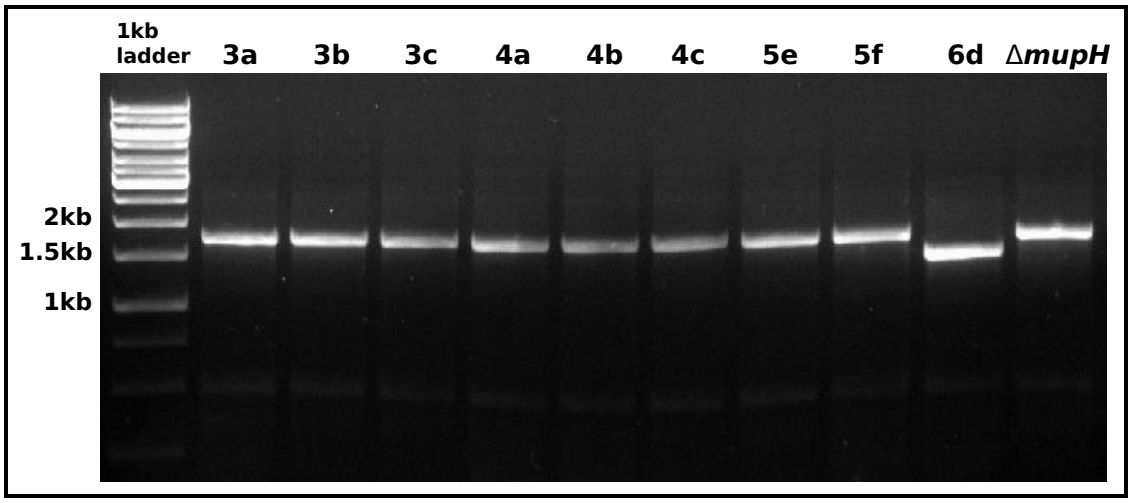

Figure 4.8: *Nine P. fluorescens ΔmupH integrants tested for correct location of ACP-K24a integration. The primers used bind at the locations outside the 500 bp arms only one out of nine showed the band closer to 1.5 kb which includes two arms of ≈1 kb and an ACP of 331 nucleotides while P. fluorescens ΔmupH which contains two ACPS and the linker region between them has slightly larger band along with all the other samples which probably reverted back to the wild type state during plasmid excision. The primers also seem to mis-prime some where in the chromosome thus generating a very faint band of ≈500 bp.*

mutant in this strain would test for a functional interaction with ACP-K24a. It was thought that the interaction of ACP-K24a with *mupH*, whether *in trans* or on the chromosome (as in Δ4-1a strain), would be unfavourable. On the other hand the interaction of ACP-K24a with *batC in trans* should be favourable in the ΔH-6d strain and it should also have some favourable effect in Δ4-1a strain, since the *batC* expressed *in trans* would compete with the chromosomal *mupH* to interact with ACP-K24a. The three plasmids that were previously transformed into the *E. coli* S17-1 strain were used for the conjugal transfer of the plasmids into *P. fluorescens* ΔH-6d and 4-1a cells. The trans-conjugants were purified over minimal media to get rid of the *E. coli* S17-1 strain and were validated by plasmid extraction and gel electrophoresis.

For the bioassay three replicates of each strain were spotted and the antibiotic efficacy was measured as the diameter of the clearance zone, excluding the size of the central disc. The average of two diameters was calculated and the standard deviation between the averages of the three replicates were plotted as error bars in a bar graph (Figure 4.9). A bioassay for each sample was carried out with and without IPTG induction.

The bioassay results showed the Δ4-1a strain had a diameter for the clearance zone that

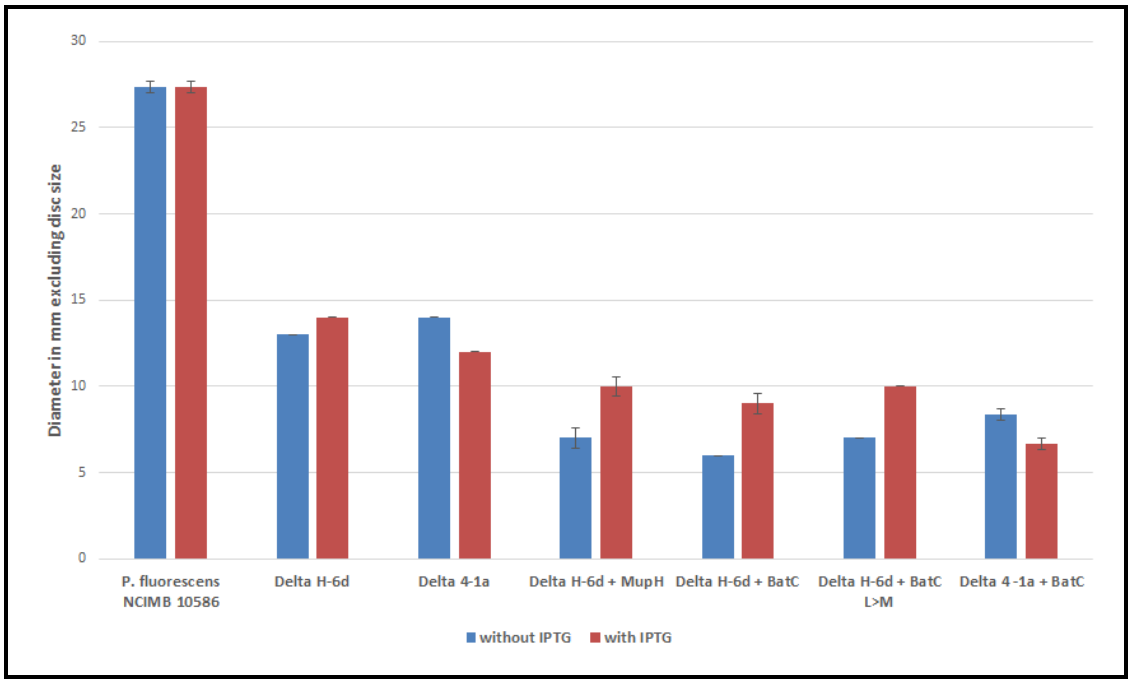

Figure 4.9: *Bioassay results for in trans expression of mupH, batC or the batC L218M mutant with and without IPTG induction. P. fluorescens NCIMB 10586 is the wild type mupirocin producer, ΔH-6d and Δ4-1a carry ACP-K24a in place of mupA3a&b-ACP in the mupirocin cluster. ΔH-6d lacks chromosomal mupH. mupH, batC and batC L218M were the genes on pJH10 plasmid expressed in trans. Error bars represents the standard deviation of three replicates with two diameters measured per strain.*

was 50 % of WT. A similar change was observed in the ΔH-6d strain which does not contain chromosomal *mupH* in the HCS cassette. Upon expressing *batC in trans* in the Δ4-1a strain the clearance zone diameter was found to be almost half of that of the Δ4-1a strain without *batC*, both with and without IPTG induction. Expressing *mupH in trans* in the ΔH-6d strain gave a clearance zone slightly smaller with IPTG and much smaller without IPTG, as compared to the Δ4-1a strain. The clearance zone of the ΔH-6d strain with *in trans* expression of *batC* did not differ greatly from ΔH-6d with *mupH in trans*. A similar clearance zone was also observed for ΔH-6d with *batC* L218M expressed *in trans*. Figure 4.10 shows the plate bioassay for one of the replicates of each sample with and without IPTG induction.

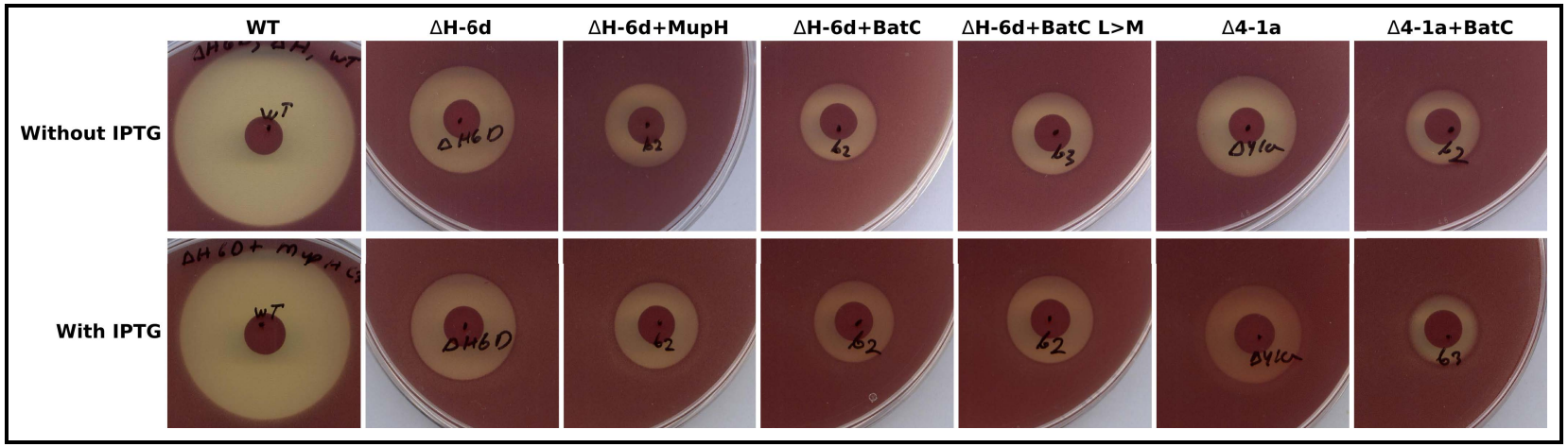

Figure 4.10: *Plate bioassay for one of the replicates of each sample with and without IPTG induction. WT is the P. fluorescens NCIMB 10586 wild type mupirocin producer, ΔH-6d and Δ4-1a carry ACP-K24a in place of mupA3a&b-ACP in the mupirocin cluster. ΔH-6d lacks chromosomal mupH. mupH, batC and batC L218M were the genes on pJH10 plasmid expressed in trans.*

4.2.4 HPLC analysis for *in trans* expression of MupH, BatC and BatC L218M mutant

In order to detect pseudomonic acids produced by the *in trans* expression of *mupH*, *batC* or *batC* L218M in *P. fluorescens* ΔH-6d and Δ4-1a strains, HPCL were performed. *P. fluorescens* NCIMB 10586 and *P. fluorescens* ΔH-6d and Δ4-1a with blank pJH10 plasmid were used as controls (see methods section 2.4.11). The chromatograms plotted in Figures 4.11 to 4.18 were zoomed in on the 5 min to 30 min retention time. Figure 4.11 shows a representative chromatogram of the elution profile of pseudomonic acid A produced by *P. fluorescens* NCIMB 10586. For the three replicates the average retention time and peak area were 20:21 min and 48273600 respectively. Figures 4.12 to 4.15 show the chromatograms of the elution profile for the products produced by *P. fluorescens* ΔH-6d strain , and Figures 4.16 to 4.18 show the chromatograms of the elution profile for the products produced by *P. fluorescens* Δ4-1a strain. No observable peak was detected for any of the known pseudomonic acids in any of these figures, however, a distinct peak for an unknown product was observed at around 17:57 min retention time for all *P. fluorescens* Δ4-1a strain (Figure 4.16 to 4.18).

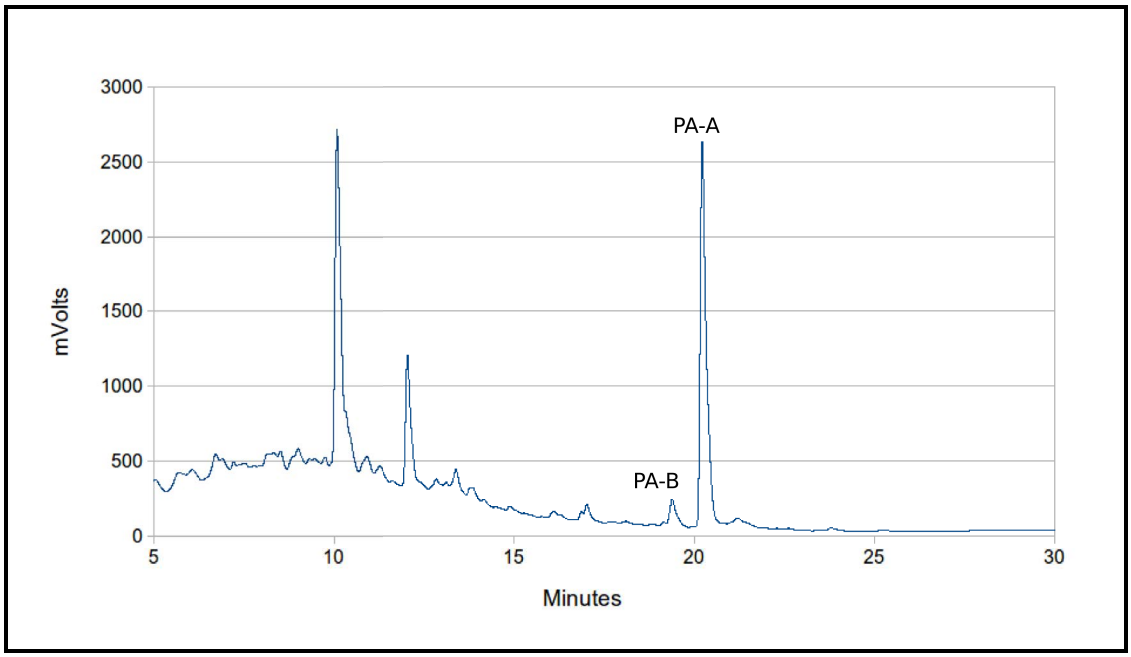

Figure 4.11: *HPLC trace for P. fluorescens NCIMB 10586.*

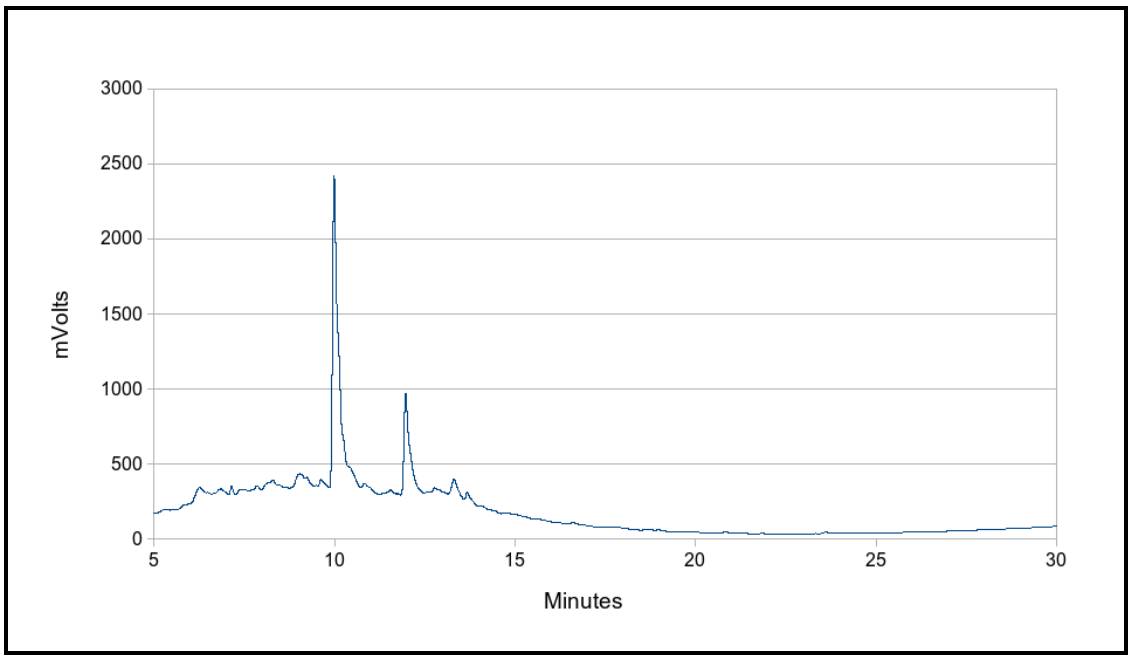

Figure 4.12: *HPLC trace for P. fluorescens ΔH-6d strain with blank pJH10 plasmid.*

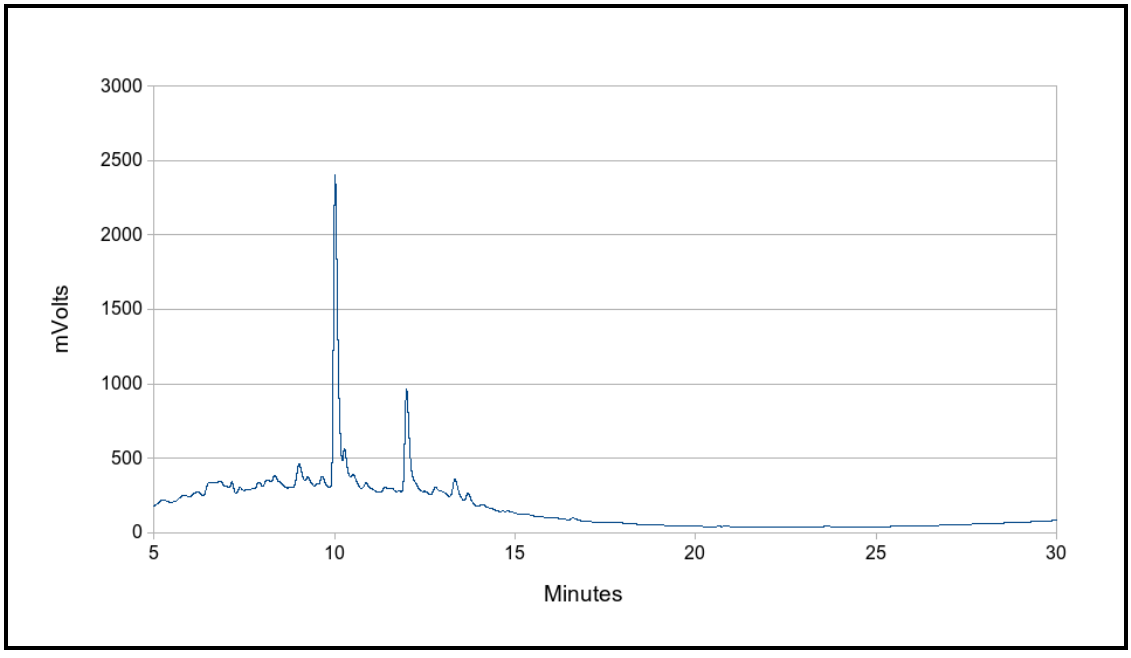

Figure 4.13: *HPLC trace for P. fluorescens ΔH-6d strain with mupH expressed in trans.*

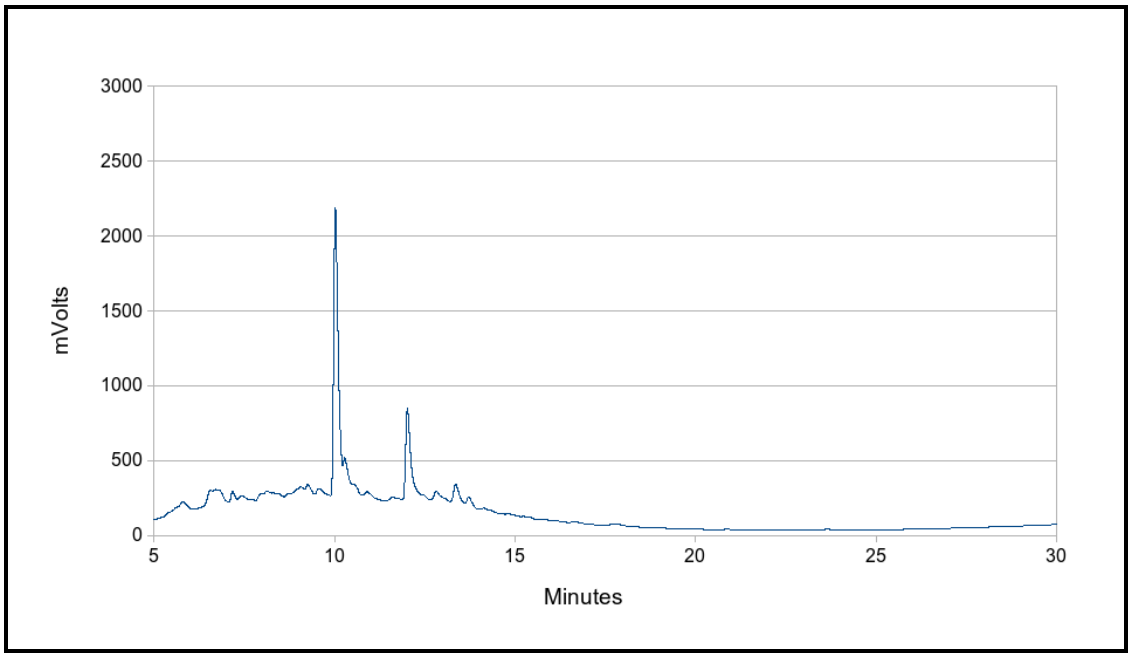

Figure 4.14: *HPLC trace for P. fluorescens ΔH-6d strain with batC expressed in trans.*

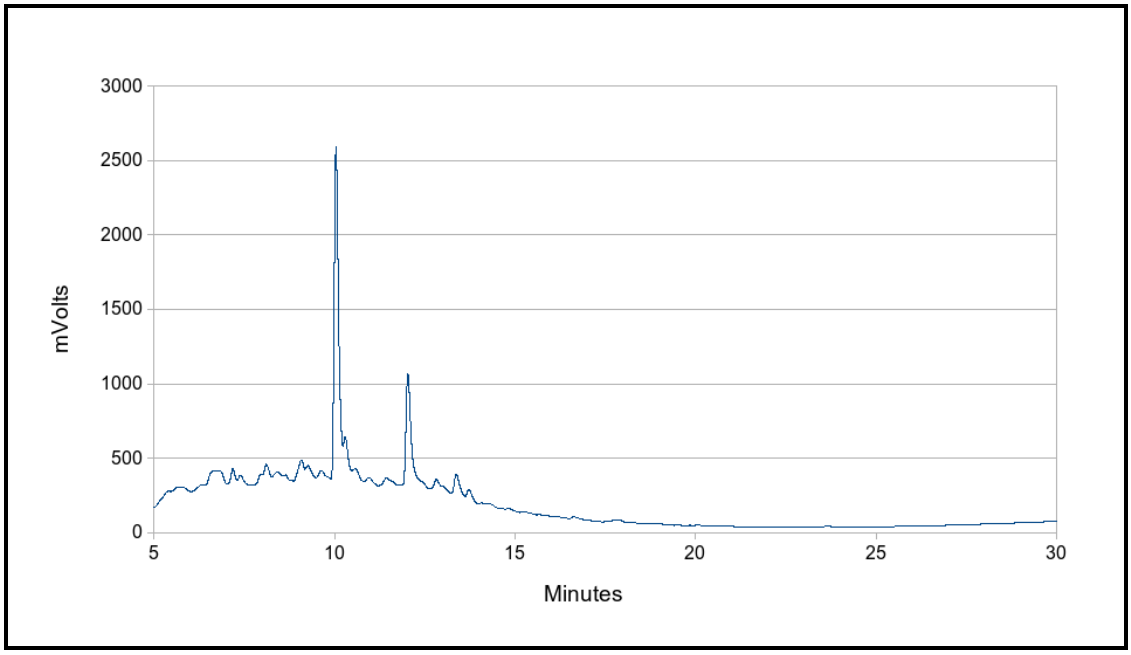

Figure 4.15: *HPLC trace for P. fluorescens ΔH-6d strain with batC L to M mutant expressed in trans.*

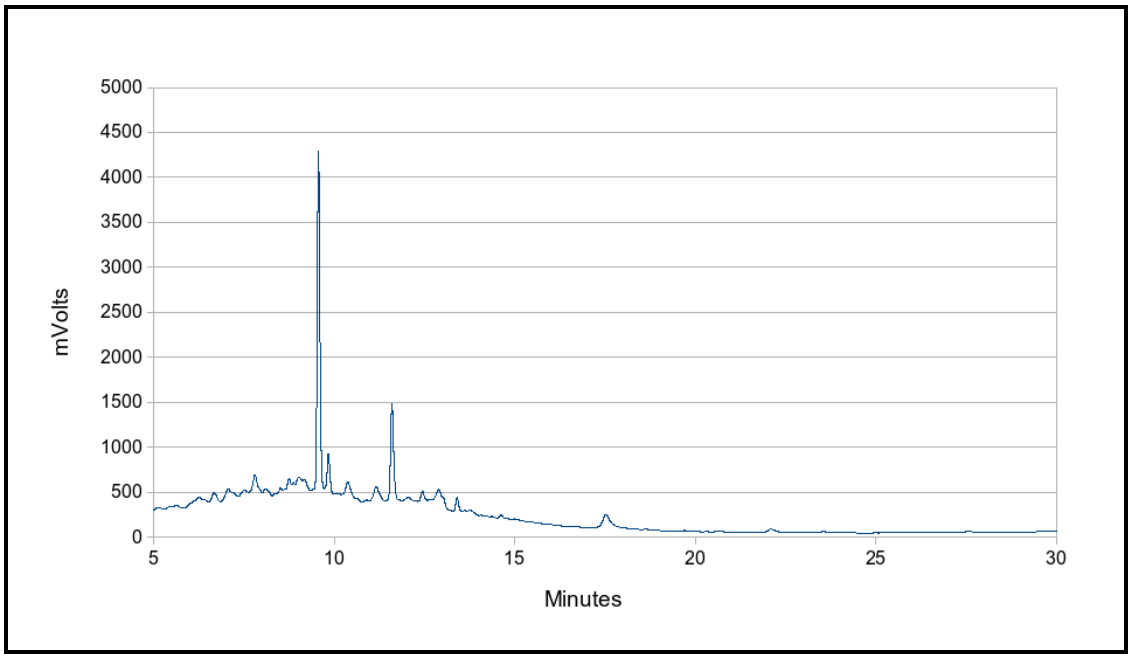

Figure 4.16: *HPLC trace for P. fluorescens Δ4-1a strain*

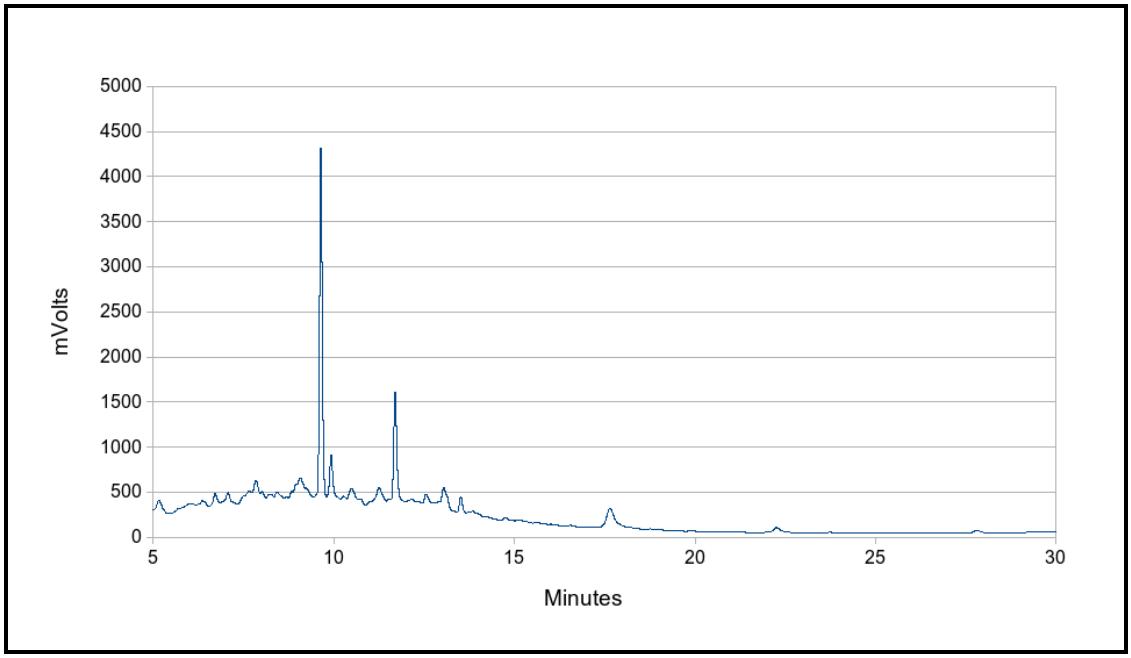

Figure 4.17: *HPLC trace for P. fluorescens Δ4-1a strain with blank pJH10 plasmid.*

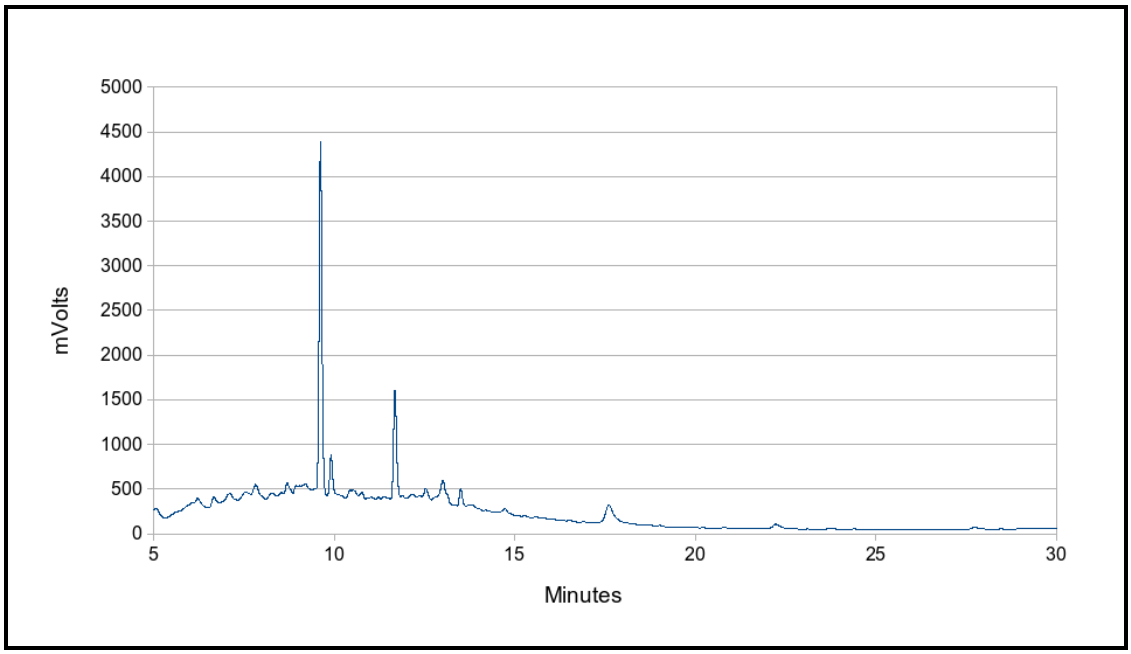

Figure 4.18: *HPLC trace for P. fluorescens Δ4-1a strain with batC expressed in trans.*

4.2.5 Molecular dynamics simulation of ACP-mupA3a+MupH complex

Molecular dynamics simulations were carried out on the ACP-mupA3a + MupH complex obtained from HADDOCK (see section 3.2.6). The starting conformation was setup to mimic the condensation stage in the HMG-CoA reaction mechanism (see Figure 3.1), with monic acid attached to the phosphopantetheine arm of ACP-mupA3a and acetyl attached to C115 of MupH, to mimic the β-branching step. Three independent simulations were run for 50 ns each. Coordinate files were extracted at every 4 ns from the trajectory and residues in ACP-mupA3a that were within 5 Å distance of M219 (MupH) were selected. Residue R30 and L32 were found to be highlighted in all the frames whereas Y62 and P65 were highlighted in most of the frames (Figure 4.19).

Based on this observation it was proposed that either these positions or their neighbours were responsible for the ACP recognition specificity associated with position M219 of MupH and its homologues. To test this hypothesis and plan mutagenesis experiments Dr. Anthony Haines segregated the β-branching ACPs from well characterised clusters into two groups based on whether their cognate HCS protein carried a methionine or a leucine. Figure 4.20 shows the alignment of the two groups. Figure 4.21 shows sequence logos built from these alignments.

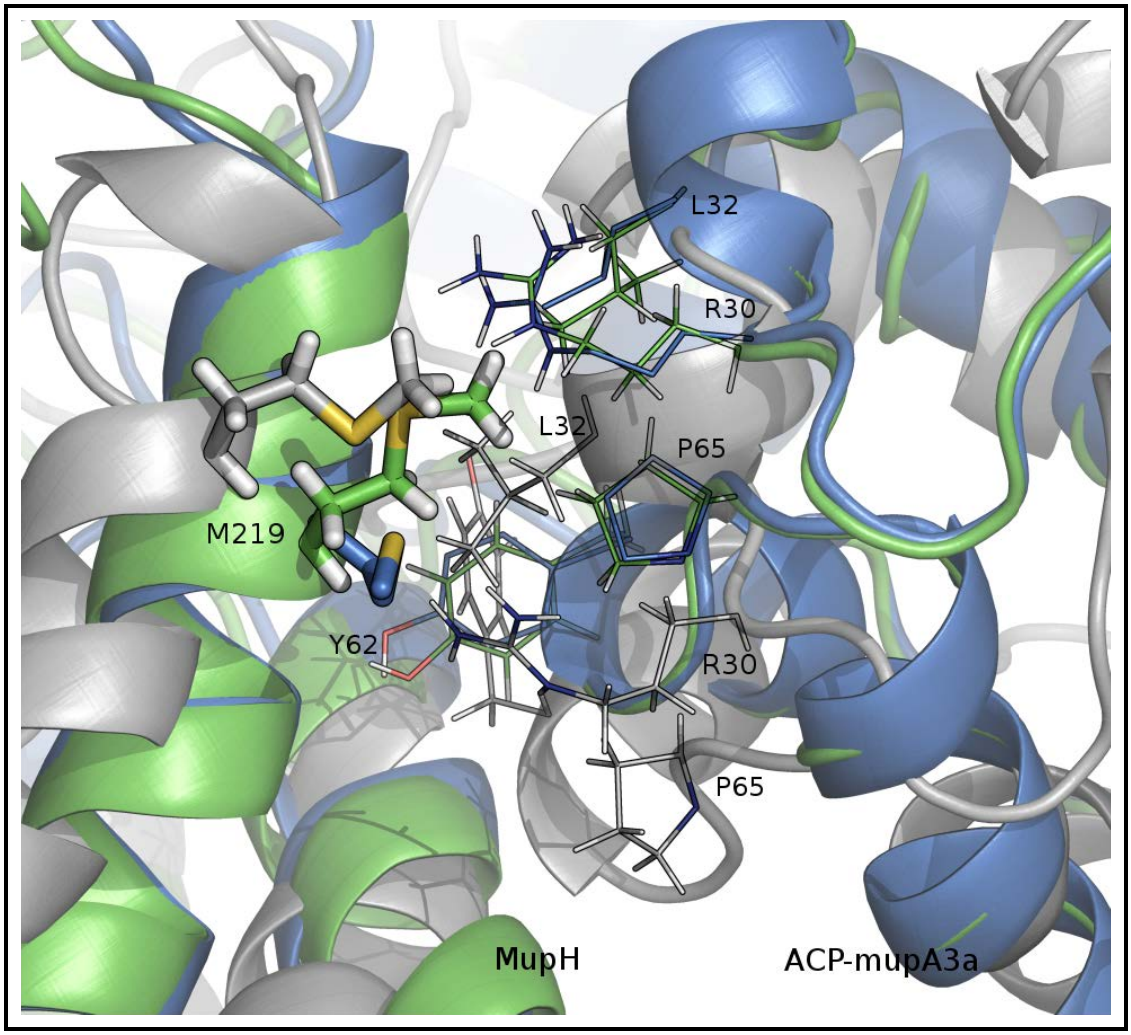

Figure 4.19: *ACP-mupA3a + MupH complex interface refined by molecular dynamics simulation. Blue: complex I of the ACP-mupA3a + MupH docked complex; Green: first frame (10 ps) of the ACP-mupA3a + MupH simulation; Grey: last frame (50 ns) of of the ACP-mupA3a + MupH simulation. R30 and L32 were highlighted within 5 Åof M219 in all the frames whereas Y62 and P65 were mostly highlighted.*

Neither the R30 or the L32 position equivalent or their neighbouring positions were differentially conserved in one set compared to the other.

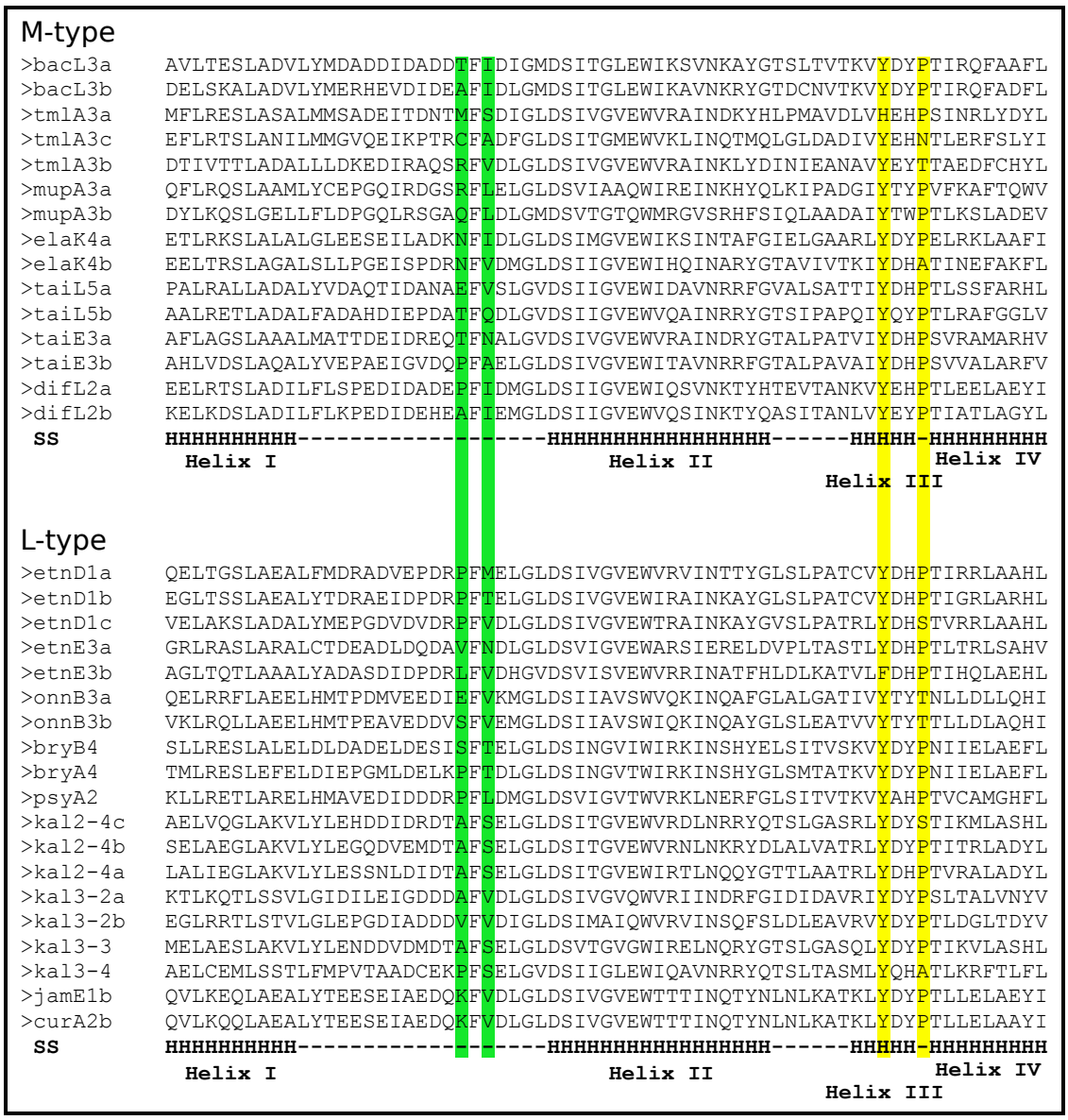

Figure 4.20: *Sequence alignment of the β-branching ACPs segregated into two groups based on their cognate HCS protein. Green/Yellow: residue position which was always/mostly highlighted with 5 Å of M219 in all the frames of molecular dynamics simulation respectively.*

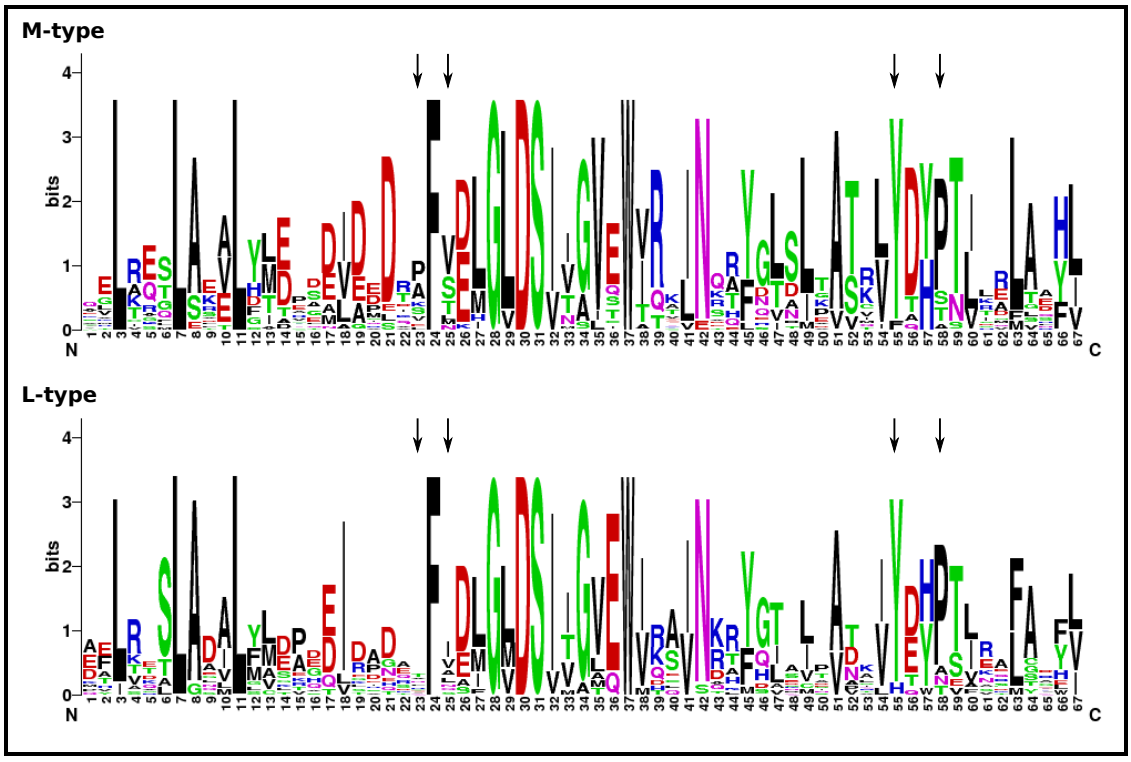

Figure 4.21: *Sequence logo built on the alignment of the β-branching ACPs segregated into two groups based on their cognate HCS protein. Arrows shows the residue positions which were always/mostly highlighted with 5 Å of M219 in all the frames of molecular dynamics simulation.*

4.3 Discussion

These experiments were designed to gain a deeper understanding of the interaction of β-branching ACPs with the HCS proteins by replacing the mupirocin β-branching ACPs with one from the kalimantacin cluster. It was proposed in the previous chapter that β-branching is a phenomenon that is dependent on the recognition specificity of a subclass of ACPs by proteins in the HCS cassette and it was demonstrated that this subclass required a conserved tryptophan in the core of the ACP, six residues downstream of the catalytic serine. This tryptophan packing in the core of the ACP appears to allow helix III to be presented at an angle which facilitates its interaction with MupH. It was also found that a residue at the position 219 on MupH alternates between a methionine and a leucine in the MupH orthologues. Moreover MupH and TmlH, which complement a ΔMupH strain, have methionine, whereas BatC, which doesn't complement ΔMupH, has leucine. Mutating this leucine in BatC into methionine does allow complementation of ΔMupH. This gain of function experiment, along with other mutations, supported the computationally

predicted ACP-MupH complex and suggested the possibility of engineering β-branching into different positions in PKS pathways. It also suggested that there exists an ACP-HCS subtype pairwise specificity. Pairwise specificity is indeed observed in the myxovirescin system, where the two HCS cassettes interact with different cognate ACPs, without apparent cross talk.

The next obvious question was to test whether the β-branching ACPs from the kalimantacin cluster can complement the β-branching ACPs in the mupirocin cluster. It was hypothesised that, as *batC* was very poor at complementing Δ*mupH* mutant in the mupirocin cluster, ACP-K24a will not complement β-branching ACPs in the mupirocin cluster. However, either by expressing *batC in trans* or upon mutating ACP-K24a to favourably interact with MupH there should be an increase in the production of pseudomonic acids. To test this hypothesis, β-branching ACP-mupA3a and ACP-mupA3ab in the *P. fluorescens* Δ*acp4* and Δ*mupH* strains respectively were replaced with ACP-K24a from the kalimantacin cluster. Suicide mutagenesis lead to the formation of the *P. fluorescens* ΔH-6d strain which lacked *mupH* and *acp-mupA3ab* on the chromosome but had *acp-K24a* incorporated into the chromosome, and the *P. fluorescens* Δ4-1a strain which had *mupH* and *acp-K24a* on the chromosome, but no *acp-mupA3ab*. Bioassay and HPLC analysis was carried out by expressing *mupH*, *batC* or *batC* L218M mutant *in trans* in the ΔH-6d strain and BatC *in trans* in the Δ4-1a strain.

Bioassay results showed the Δ4-1a strain had a diameter for the clearance zone that was 50% of WT, this observation was consistent with the original hypothesis that there would be either no or an unfavourable interaction between ACP-K24a and MupH. However, a similar drop was also observed in the ΔH-6d strain which does not contain chromosomal *mupH* in the HCS cassette. It was also thought that *in trans* expression of *batC* in Δ4-1a strain might increase the antibiotic production as BatC will compete with MupH for ACP-K24a and would make a more favourable interaction. However, the clearance zone was found to have almost half the diameter of that of the Δ4-1a strain without *batC* expressed *in trans*, both with and without IPTG induction. ΔH-6d with *in trans* expression of *mupH* should have given a similar clearance zone to that of the Δ4-1a strain as the only difference was that *mupH* is on the pJH10 plasmid in the former and on the chromosome in the latter. However, the clearance zone was much smaller

without IPTG and slightly smaller with IPTG, as compared to the Δ4-1a strain. ΔH-6d strain with *in trans* expression of BatC, which was thought to have a favourable interaction with ACP-K24a, did not have any results that were different from ΔH-6d with MupH *in trans*. Similar clearance zones were also observed for ΔH-6d with BatC L218M expressed *in trans*. Thus, because of the mixed results in the bioassay, it was difficult to conclude whether the hypothesis stands true or not. HPLC performed on the control samples showed the characteristic peak of pseudomonic acid A at an average time of 20:21 min. However, no peaks for pseudomonic acids were detected in any of the ΔH-6d samples. No observable peak was detected for pseudomonic acids in the Δ4-1a strain but all the samples had a small peak at around 17:57 min retention time. The samples were sent to our collaborators in Bristol for the structural elucidation of the metabolite produced at this previously unknown retention time.

It was hypothesised that owing to the proposed pairwise specificity between the branching ACPs and their cognate HCS proteins there would be an incompatibility between ACP-K24a and MupH, however, ACP-K24a should be able to perform well with the *batC* expressed *in trans* in the mupirocin cluster. Previous experiments showed that BatC L218M mutation was capable of functioning well with the HCS proteins in the mupirocin cluster, which suggests that the incompatibility may not be because of the BatC but because of the ACP-K24a which was unable to interact efficiently with the other domains in the mupirocin cluster. There could be many possibilities that caused ACP-K24a not to allow production of pseudomonic acids in the mupirocin cluster. It is possible that the ACP-K24a did interact with MupH or its orthologues but failed to interact with the other HCS cassette proteins. It is also possible that the ACP-K24a failed to interact with its cognate KS and hence it couldn't transfer the extender unit for the Claisen condensation. It could also be possible that the ACP-K24a was successful in interacting with the KS as well as with the HCS cassette protein thus producing a β-branch but that it failed to pass on the product to MmpB. If it fails to pass on the β-branched monic acid moiety to MmpB then it is possible that some of the methylated-monic acid would leak out of the ACP and that would explain the small peak observed in the Δ4-1a samples. Monic acid being less hydrophobic would elute earlier as compared to the pseudomonic acid A. However, this leak

would also be observed in the ΔH-6d strain with MupH expressed *in trans*. The efficiency of MupH being expressed *in trans* or on the chromosome may be in question. Thus, it would only be conclusive once the structure of the metabolite produced at the retention time around 17:57 min is solved by our collaborators.

An alternative method for investigating pairwise specificity might be to make changes in the ACP-mupA3ab β-branching ACPs native to the mupirocin cluster, to allow them to function efficiently with the *batC* when expressed *in trans*. In earlier experiments the BatC L218M mutant showed successful complementation which implies its suitability in the mupirocin cluster. Therefore, to identify residues on the ACPs that might be responsible for determining the specificity of the interaction with BatC, molecular dynamics simulations were carried out on the ACP-mupA3a + MupH complex (as described in Section 2.3.2.7), to refine the structure and identify ACP residues interacting with M219. MD simulation highlighted four positions which were always or mostly present within 5 Å distance of the M219. These four positions R30, L32, Y62 and P65 in ACP-mupA3a were proposed to Dr. Anthony Haines in order to plan mutagenesis experiments. In order to find out to what residue type these positions should be mutated, Dr. Haines built two separate sequence alignments of the β-branching ACPs from well studied clusters, one alignment for each of the two types, MupH homologues which carry an M219 and the ones which carry an L at the equivalent position. The sequence alignment of these two groups of ACPs did not reveal any information about the difference between them, there being no obvious correlation between ACP residues predicted to interact with L/M219 and the presence of L or M..

Assuming that ACPs can be sub grouped according to their cognate HCS protein type, a more sophisticated method such as hidden Markov models could be used to cluster the ACPs. As mentioned in the previous chapter HMMs were successful in clustering the β-branching ACPs from the standard ACPs, constructing HMM models for the two above mentioned sequence alignments and scoring them against each other may reveal the ACP subgroups. Utilizing the technique mentioned in section 3.2.1.1 it would be possible to identify the minimum number of changes required to shift an ACP from one subtype to another, which would help to

design mutagenesis experiments.

Chapter 5

On the dynamics of acyl carrier protein

5.1 Introduction

Acyl carrier proteins play an important role in shuttling substrates and products in fatty acid and polyketide biosynthesis. Inspite of being expressed in a minuscule amount (e.g. ≈0.25% of all the soluble proteins in an *E. coli* cell), the dynamic nature of these proteins helps them to interact with several different core and auxiliary domains in the FAS and PKS machinery. An ACP is a four helix bundle with three helices I, II and IV running parallel to each other, while helix III runs perpendicular to the others. A holo ACP is an ACP charged with phosphopantetheine, a prosthetic linker derived from coenzyme A, which is transferred to the ACP by ACP synthases. The Holo ACP then tethers an acyl group to the terminal sulfhydryl of its phosphopantetheine via a thioester linkage. Experiments and molecular dynamics studies on FAS ACPs have shown the sequestering of acyl substrates within the hydrophobic core formed by the four helix bundle. However, it is not known if there is a similar mechanism in PKS ACPs.

Chan *et al.* (2008) conducted molecular dynamics simulations on apo, holo, and acyl forms of *E. coli* FAS ACP which suggested a mechanism for acyl chain binding inside the ACP core. Simulations were performed using an experimentally determined holo FAS ACP structure (PDB ID 1L0I) with a butyryl molecule trapped inside the core of a holo ACP. Several different lengths of acyl chain were constructed by extending the butyryl molecule by two carbons at a time up to eighteen carbons. Simulations were conducted starting with the acyl chains in solvent exposed

as well as in buried states and solvent exposed acyl chains up to ten carbons long easily found their way into the hydrophobic core of the ACP, stably residing in the core till the end of a 50 ns simulation. However, longer acyl chains found it difficult to find there way into the ACP core, and were not consistently stable in the core till the end of simulations. Chan *et al.* (2008) proposed an eight carbon acyl chain as the optimal for stable accommodation by this FAS ACP (PDB ID 1L0I). They also found a novel binding pocket in which the acyl chain was directed towards helix I as opposed to the binding pocket in the solved structure. Simulations also highlighted residues T39 and E60 as stabilizing the position of the phosphopantetheine linker through hydrogen bonds at the opening of the hydrophobic tunnel.

Here, molecular dynamics (MD) simulations of the apo, holo, and acyl forms of ACP-mupA3a and the acyl form of ACP-mupA2a from the mupirocin cluster are presented in order to investigate the dynamic behaviour of the PKS ACPs upon phosphopantetheine and acyl chain attachment. The main property of interest was to detect the formation of a cavity or tunnel that could sequester a PKS substrate, similar to that seen in the FAS ACP. MD simulation trajectories were analysed to calculate the hydrogen bonds formed by the phosphopantetheine and by the substrates, in the holo and acyl forms respectively, with the protein surface and the solvent. MD trajectories were also analysed to calculate the solvent accessible surface area (SASA) of the phosphopantetheine and the substrates in the holo and acyl forms respectively. To test whether the PKS ACPs become more like FAS ACPs in their holo and acyl forms, root mean square deviations (RMSD) of the PKS ACPs from the reference FAS ACP (PDB ID 1LOI) were calculated.

5.2 Results

5.2.1 Molecular dynamics simulation setup and parameter determination

In order to carry out molecular dynamics simulations of the apo, holo, and acyl forms of ACP-mupA3a and the acyl form of ACP-mupA2a, the first step was to generate the starting structures and assign the force field parameters.

For ACP-mupA3a two structures were used, a wild type (WT) and a W44L mutant. For the

apo ACP-mupA3a WT, the NMR determined structure (PDB ID 2L22) was used as the starting conformation. For the apo ACP-mupA3a W44L, the same ACP-mupA3a structure was used but with the point mutation created by PyMol. To create the holo ACP structures, the catalytic S38 of their respective apo structures was extended with a phosphopantetheine. Two acyl forms of ACP-mupA3a WT were created, the first structure consists of the cognate substrate of the ACP-mupA3a (as shown in Figure 5.1) built directly onto the phosphopantetheine and was referred to as "Acyl ACP-mupA3a WT". The second acyl form consists of a fourteen carbon saturated chain built directly onto the phosphopantetheine of the holo ACP-mupA3a WT and was referred to as "Acyl 14C ACP-mupA3a". One acyl form of ACP-mupA3a W44L was also created by extending the phosphopantetheine of the holo ACP-mupA3a W44L with the cognate substrate of ACP-mupA3a and was referred to as "Acyl ACP-mupA3a W44L".

For ACP-mupA2a, as there was no experimentally determined structure available, a homology model was generated using as the template the ACP (PDB ID 2LIU) from the CurA module of the curacin system in *Lyngbya Majuscula* (as described in Section 2.3.1.2). To create the acyl form of ACP-mupA2 the catalytic serine of the modelled structure was extended with the phosphopantetheine followed by its cognate substrate (as shown in Figure 5.1) and was referred to as "Acyl ACP-mupA2".

Following the structure preparation the force field parameters were assigned to the atoms. Parameters from the AMBER99SB-ILDN force field were used for the protein atoms and GAFF parameters were used for the phosphopantetheine and the acyl chains with charges calculated using the RED server (Vanquelef *et al.* 2011). A detailed description of the molecular dynamics simulations protocol and the procedure used to determine the parameters for the ligands is given in Section 2.3.2 and Section 2.3.2.1 respectively.

Multiple independent simulations were carried out for 50 ns for each of the holo and acyl forms, and one of the simulations that showed a transition into the binding mode was extended to 200 ns. One simulation for the Acyl ACP-mupA3a WT was further extended to 1 μs. Simulation for the apo ACP-mupA3a WT was also extended to 1 μs as a control. Table 5.1 summarizes the different simulations setup with reference to the methods sections for details.

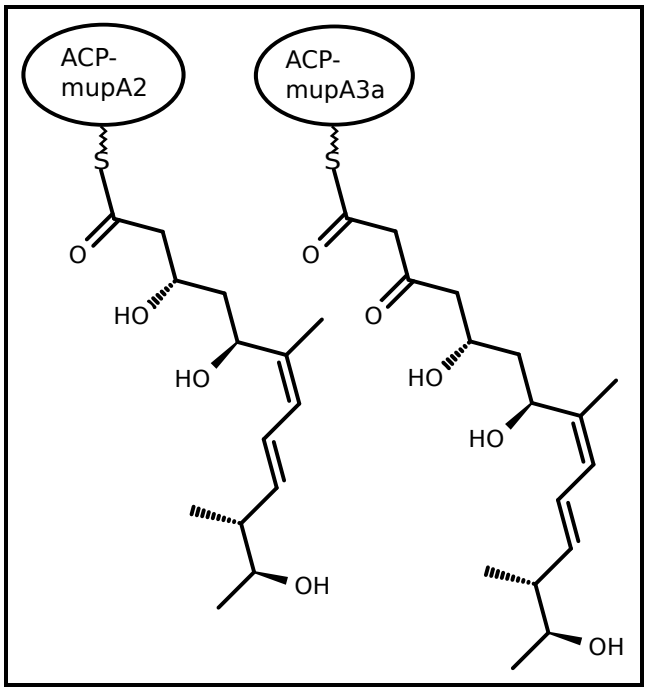

Figure 5.1: *Cognate substrates for ACP-mupA2 and ACP-mupA3a in the mupirocin pathway.*

Table 5.1: *ACP simulation setup summary*

Structure	**Ligand**	**Simulation**	**Method Section**
Apo ACP-mupA3a WT	-	1 X 1 μs	2.3.2.2 B
Apo ACP-mupA3a W44L	-	1 X 200 ns	2.3.2.2 B
Holo ACP-mupA3a WT	phosphopantetheine	4 X 50 ns 1 X 200 ns	2.3.2.3
Holo ACP-mupA3a W44L	phosphopantetheine	4 X 50 ns 1 X 200 ns	2.3.2.3
Acyl ACP-mupA3a WT	ACP-mupA3a cognate substrate	4 X 50 ns 1 X 1 μs	2.3.2.4
Acyl ACP-mupA3a W44L	ACP-mupA3a cognate substrate	4 X 50 ns 1 X 200 ns	2.3.2.4
Acyl 14C ACP-mupA3a	fourteen carbon saturated chain	2 X 50 ns 1 X 200 ns	2.3.2.5
Acyl ACP-mupA2	ACP-mupA2 cognate substrate	2 X 50 ns 1 X 200 ns	2.3.2.6

5.2.2 ACP backbone dynamics over time

All the simulations were stable, as indicated by the RMSD and RMSF calculations. The RMSD values of the backbone atoms from the reference starting structures remained below 2.5 Å for all the simulations and plateaued within the first 50 ns. The root mean square fluctuation (RMSF) calculated for the backbone atoms, averaged per residue (as described in Section 2.3.2.9), showed a higher fluctuation in the mutant structures as compared to the wild type. This observation is consistent with the MD simulations mentioned in the Chapter 3, Section 3.2.2.1. However, upon attaching the cognate substrate of ACP-mupA3a, 200 ns and μs simulations of the acyl ACP-mupA3a WT showed more fluctuation than the 200 ns simulations of acyl ACP-mupA3a W44L (Figure 5.2). Acyl 14C ACP-mupA3a also showed more fluctuation than the acyl ACP-mupA3a W44L. However, the residues showing most fluctuation were different in the acyl ACP-mupA3a WT and the acyl 14C ACP-mupA3a. Interestingly, fluctuations in the simulation of acyl ACP-mupA3a WT across 1 μs remained similar to the fluctuations observed in an independent 200 ns long simulation of the acyl ACP-mupA3a W44L. Most of the differences in fluctuations between the simulations of the wild type and the mutant ACPs were in the region of residues from 34 to 44 (around the catalytic serine) and 52 to 68 (loop II and helix III) (Figure 5.2).

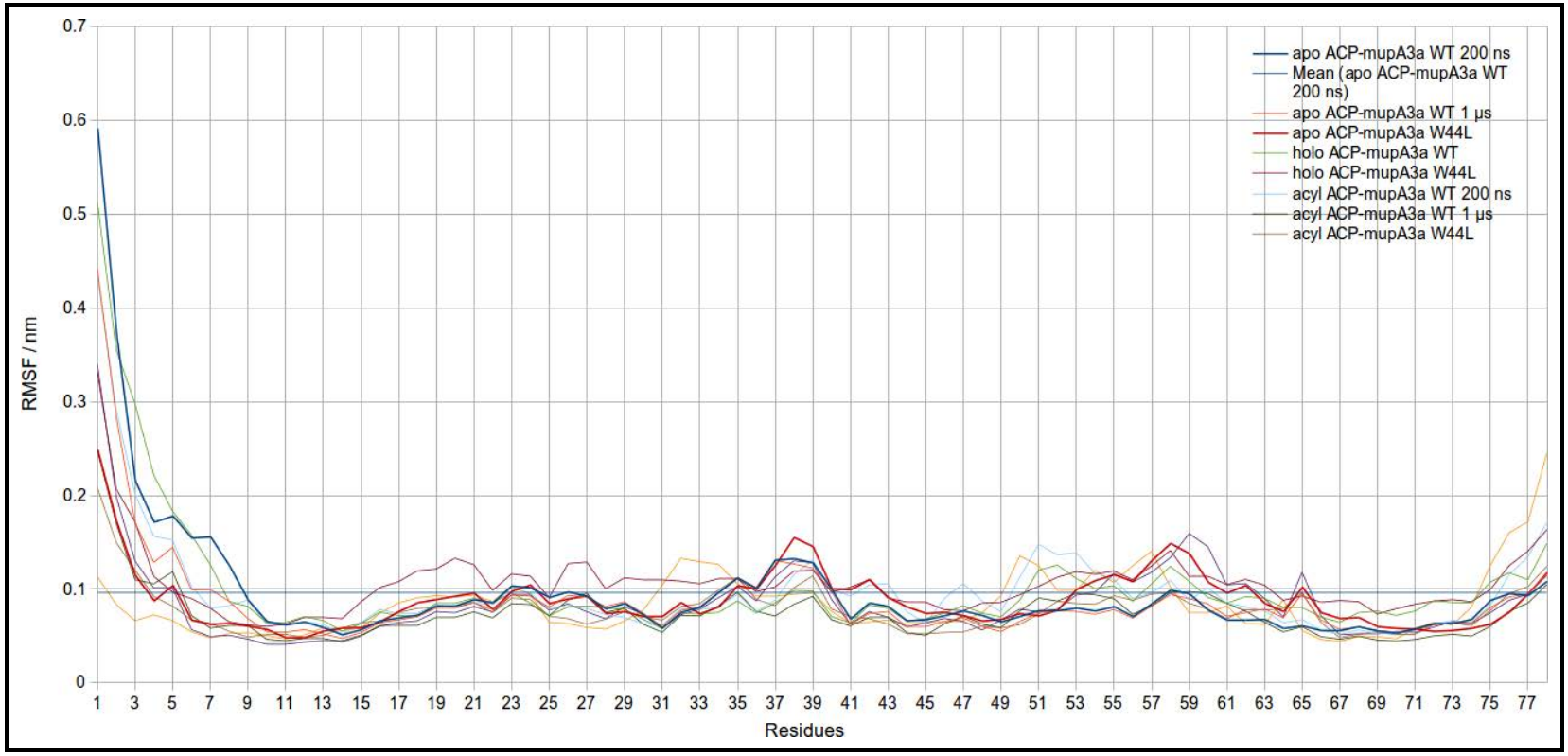

Figure 5.2: *Root mean square fluctuation (RMSF) of the backbone atoms per residue. Thick blue line represents the apo ACP-mupA3a WT and the thick red like represents apo ACP-mupA3a W44L. Straight blue line represents the mean of the apo ACP-mupA3a WT RMSF values. Simulation lengths are 200 ns unless otherwise specified. Apo ACP-mupA3a WT 200 ns and acyl ACP-mupA3a WT 200 ns are the first 200 ns of the 1 µs simulations.*

5.2.3 Formation and change in cavity volume in PKS ACPs over time

After verifying the quality of the simulations, the first macroscopic property of interest was to detect the formation of a cavity similar to that seen by Chan *et al.* (2008). The cavity volume calculation was performed by using a third party GROMACS plugin *trj_cavity* as described in Section 2.3.2.11. Monitoring the change in the cavity volume over time in ACP-mupA3a and ACP-mupA2a ACPs revealed the formation of larger cavities in the holo and acyl forms. Table 5.2 shows the mean and modal cavity volumes formed by the different structures over time, the apo form having the lowest volume with a mean value of 62.856 $Å^3$ in the 1μs long simulation. The holo as well as the different acyl forms in both ACP-mupA3a (wild and mutant types) and ACP-mupA2a showed higher mean volumes than the apo ACP-mupA3a structures. Thus, suggesting the role of the prosthetic linker as well as the acyl group attachment in the induction of a larger cavity. It was also observed, that the mean cavity volume in holo ACP-mupA3a W44L and acyl ACP-mupA3a W44L structures, were lower than their wild type counter parts. The mean cavity volume of the apo ACP-mupA3a W44L was found to be larger than the apo ACP-mupA3a WT structure however, the modal cavity volume was smaller in the mutant than the wild type. This observation suggests, the mutation at the core of the ACP might have changed the packing of the helices, thus affecting the cavity formation at the surface. Although there is an induction of cavity formation and a change in the volume during the simulations of holo and acyl states, in both cases the cavity is a solvent exposed groove (see example in Figure 5.4 A & B) rather than a buried tunnel (see example in Figure 5.3 A). Figures from C.1 to C.10 in the Appendix III shows graphs of the change in the cavity volume (non zero time frames) over time, along with the running average over 500 frames and a mean value line.

The largest cavity (avoiding any cavity with spill overs) recorded among all the structures was 266 $Å^3$ in acyl ACP-mupA3a WT (Figure 5.4). Here, spill overs means that the probe couldn't find protein in five directions and hence ran off the intended cavity space until it finds the protein. This usually happens with flatter and surface exposed cavities. The parameters used should detect a cavity in as many time frames in the simulation as possible including the possibility of detecting slightly flatter surface cavities without producing spill overs. However,

Table 5.2: *Average values for cavity volume, hydrogen bonds and solvent accessible surface.*

Simulation	**Mean & modal cavity volume ($Å^3$)**		**Mean number of hydrogen bonds with PPT**				**Mean number of hydrogen bonds with acyl chain**				**Mean of SASA (nm^2)**	
	WT	**W44L**	**WT**		**W44L**		**WT**		**W44L**		**WT**	**W44L**
			Protein	**Solvent**	**Protein**	**Solvent**	**Protein**	**Solvent**	**Protein**	**Solvent**		
Apo ACP-mupA3a (200ns)	64.951 / 52.728	75.408 / 50.531	-	-	-	-	-	-	-	-	-	-
Apo ACP-mupA3a (1 μs)	62.856 / 52.728	-	-	-	-	-	-	-	-	-	-	-
Holo ACP-mupA3a	107.303 / 109.85	95.976 / 92.274	1.180	11.612	2.376	10.473	-	-	-	-	5.876	5.275
Acyl ACP-mupA3a (200ns)	107.452 / 103.259	99.163 / 57.122	0.415	10.341	0.445	9.346	2.618	5.719	1.295	6.825	4.943	4.920
Acyl ACP-mupA3a (1μs)	82.463 / 72.501	-	0.376	10.012	-	-	2.221	5.493	-	-	4.901	-
Acyl 14C ACP-mupA3a	94.186 / 96.668	-	0.574	10.649	-	-	-	-	-	-	4.288	-
Acyl ACP-mupA2a	101.973 / 92.274	-	1.430	7.504	-	-	1.060	5.748	-	-	4.433	-

finding perfect parameters to detect a cavity in all the time frames whilst avoiding any spill overs was difficult. On visual inspection of 10000 time frames for 100ns of apo ACP-mupA3a WT simulation, showed fewer instances of spill overs with the chosen parameters compared to the other parameters tried, as detailed in Section 2.3.2.11. Table 5.3 lists the largest as well as the modal cavity volumes (for comparison) recorded for all the structures. Using the same cavity detection parameters, the volume for the FAS ACP (PDB ID 1L0I) structure was found to be 253 $Å^3$. Comparing this FAS cavity size from the crystal structure with the largest recorded in the PKS acyl ACP-mupA3a WT, the difference is $\approx$13 $Å^3$ (around 5 %). However, the shape of the FAS ACP looks more like a narrow deep tunnel as compared to the shallow but wide surface exposed channel of ACP-mupA3a. Figure 5.3 (A) shows a space filled drawing of the cavity volume detected in the FAS ACP and Figure 5.3 (B) shows the orientation of the butyryl acyl group attached to the phosphopantetheine inside the FAS ACP hydrophobic tunnel. Comparing largest and modal cavities from the simulations of other PKS ACP structures also revealed shallow but wide surface exposed channels (see Figures from C.11 to C.17 in the Appendix III). Not having a deep tunnel shaped cavity means that the ligand was partially accessible to the solvent throughout the ACP-mupA3a simulations. This observation was supported by measuring the change in the solvent accessible surface area (SASA) of the ligands (phosphopantetheine and the acyl groups) over time. Table 5.2 shows the mean of the SASA for different ligands which was found to range in between 4 to 6 nm^2 (see Figures from C.40 to C.46 in the Appendix III).

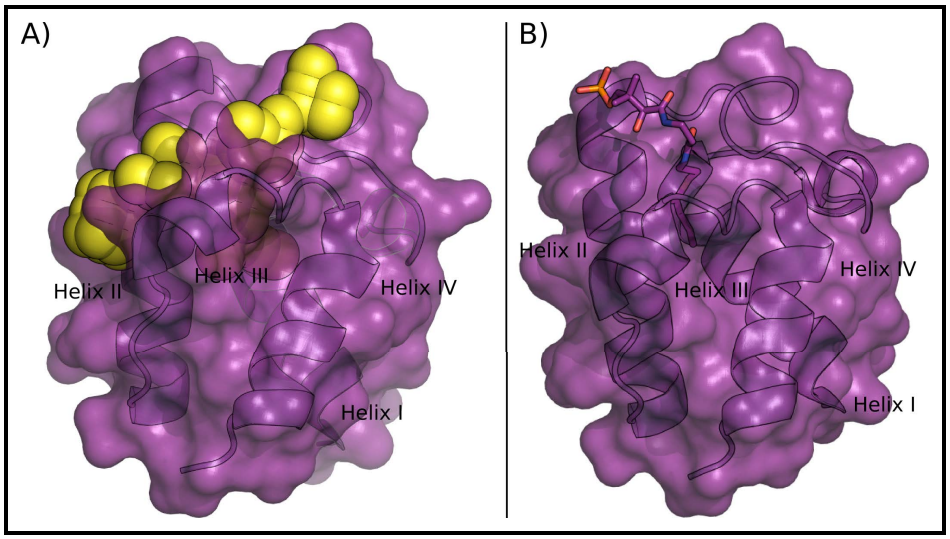

Figure 5.3: *Cavity volume detected in the reference FAS ACP (PDB ID 1L0I) structure. (A) space filled (yellow spheres) drawing of the cavity detected (B) orientation of the butyryl acyl group attached to the phosphopantetheine inside the FAS ACP hydrophobic tunnel. Helices are shown as cartoon drawing and numbered.*

5.2.4 Change in RMSD of PKS ACPs from FAS ACP over time

In watching movies of the simulations it seemed that holo and acyl forms of the PKS ACPs become more like FAS ACPs in structure. It seemed that the structure might be responding to the presence of the ligands to produce a cavity. To quantify whether there was a relationship between cavity volume and similarity to FAS structure the correlation was calculated between the cavity volume at every ns and the RMSD of the backbone atoms of the PKS ACPs from the FAS ACP reference structure (Table 5.4). A weak negative correlation was observed in all the structures except in the acyl ACP-mupA2a i.e. as the volume increases the similarity to FAS increases, albeit only slightly. The correlations were slightly stronger in the ACP-mupA3a W44L structures with the strongest in holo ACP-mupA3a W44L simulation. Comparing the Figures 5.5 and 5.6 shows a high rise in the RMSD between 120 and 160 ns which strongly correlates with the sudden dip in the cavity volume in the same time frame. Figures from C.18 to C.27 in the Appendix III show the change in the RMSD of the backbone atoms of each PKS ACP from the FAS ACP over time. Taken individually these correlation coefficients might seem insignificant but the fact that they are all negative does point to a small but significant effect.

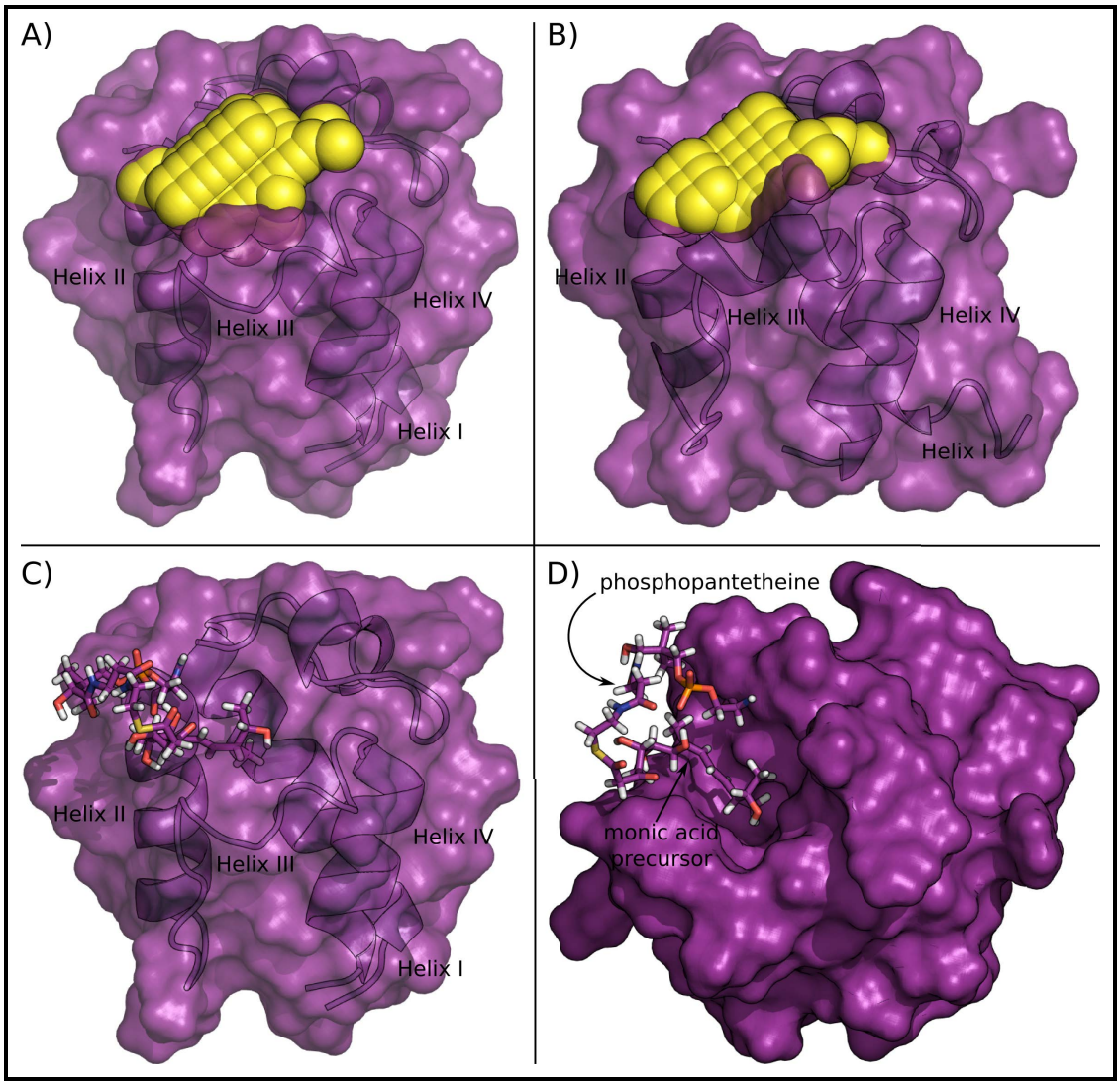

Figure 5.4: *Largest cavity detected in the acyl ACP-mupA3a WT. (A & B) space filled (yellow spheres) drawing of the largest and the modal cavity volume respectively. (C & D) orientation of the ligand in the cavity formed in two different views.*

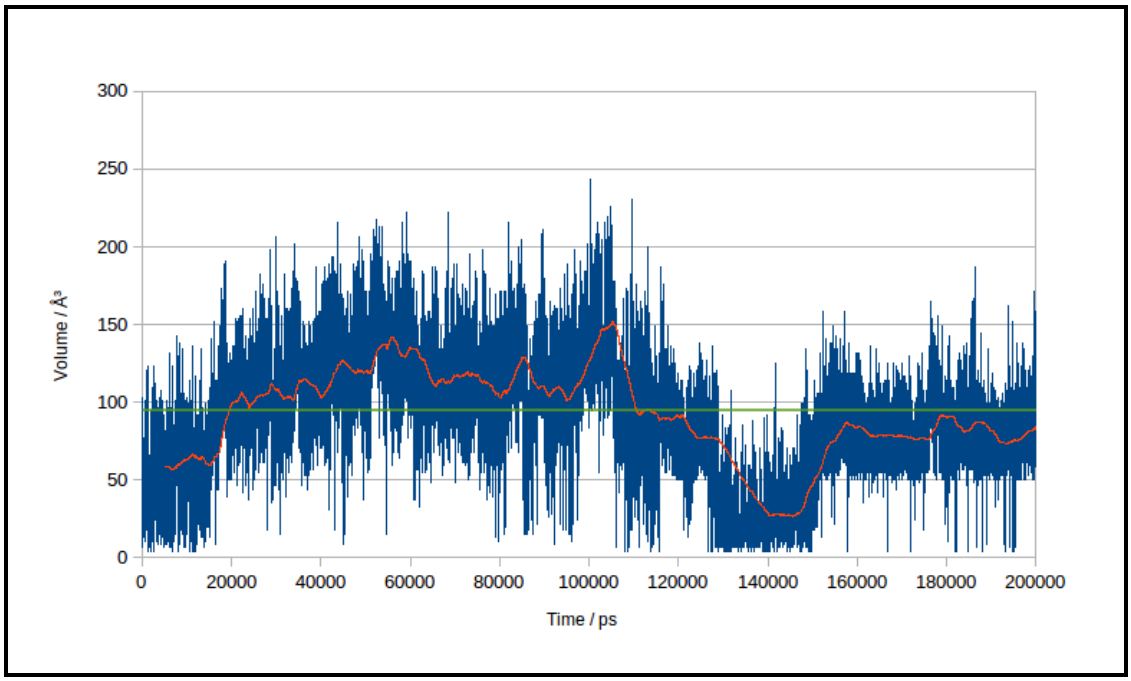

Figure 5.5: *Formation and change in cavity volume over time in the holo ACP-mupA3a W44L. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

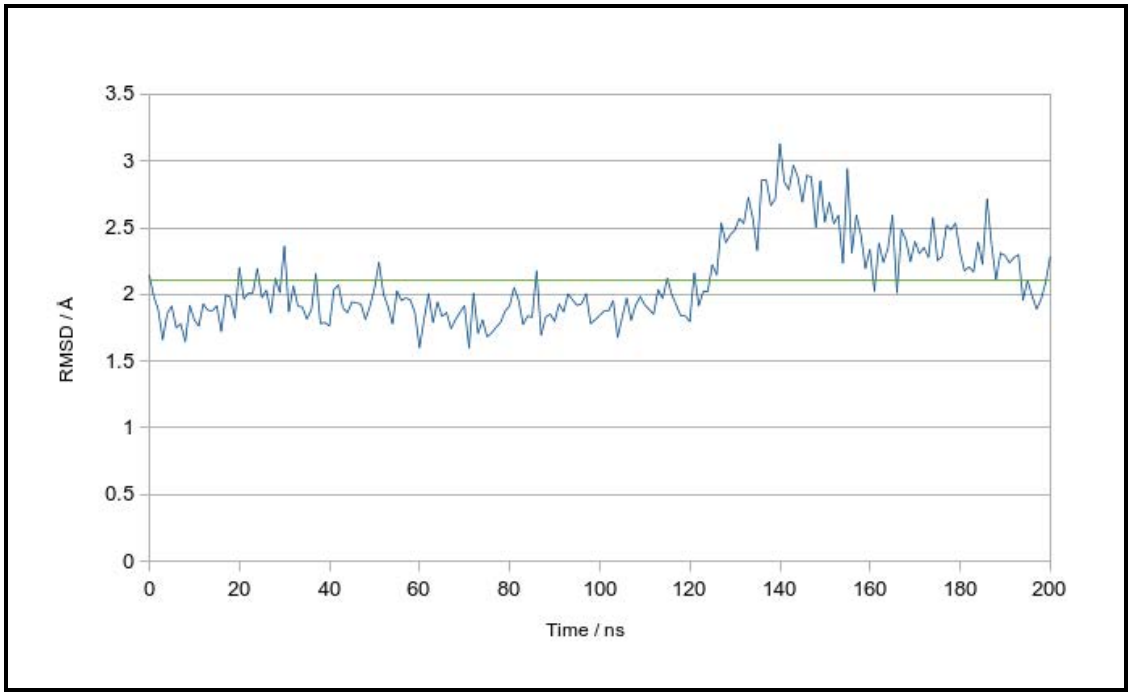

Figure 5.6: *RMSD between FAS ACP and the holo ACP-mupA3a W44L over time. Green line represents the mean.*

Table 5.3: *Largest and modal cavity volume for each simulation*

Simulation	WT		W44L	
	Largest cavity volume	**Modal volume**	**Largest cavity volume**	**Modal volume**
Apo ACP-mupA3a (200ns)	151.593	52.728	173.563	50.531
Apo ACP-mupA3a (1μs)	151.593	52.728	-	-
Holo ACP-mupA3a	228.488	109.85	243.867	92.274
Acyl ACP-mupA3a (200ns)	265.837	103.259	243.867	57.122
Acyl ACP-mupA3a (1μs)	265.837	72.501	-	-
Acyl 14C ACP-mupA3a	195.533	96.668	-	-
Acyl ACP-mupA2a	248.261	92.274	-	-

5.2.5 Hydrogen bonding between the phosphopantetheine, acyl groups and protein/solvent

In watching movies of the simulations, the phosphopantetheine and acyl chains were seen to interact with the protein. However, as the cavity detected during the simulation was shallow and solvent exposed the ligands seemed to be in partial interaction with the solvent as well. To detect the interaction of the ligands with the protein and solvent hydrogen bonds were measured between the ligands and the protein/solvent throughout the simulation. Table 5.2 shows the mean number of hydrogen bonds per frame formed with the protein and solvent (columns 3 and 4 respectively) by the phosphopantetheine and the acyl groups. It was observed that the phosphopantetheine constantly makes hydrogen bonds with the solvent, with a very slight drop in the average number when an acyl group is attached, as compared to without substrate. No significant correlation was found between the cavity volume and the hydrogen bonds formed between the ligand and the protein. However, a negative correlation was observed between the hydrogen bonds formed by the ligand with the protein and the solvent. This negative correlation showed a general trend of a rise of hydrogen bonds made by the ligand with the protein and fall of hydrogen bonds made by the ligand with the solvent over time (see Figures from C.28 to

C.39 in Appendix III). A general trend where for every hydrogen bond formed between ligand and protein one between ligand an solvent is lost would be expected.

5.2.6 Structural and sequence comparison of the *E. coli* FAS ACP and ACP-mupA3a

In order to understand what might be preventing ACP-mupA3a from forming a deep tunnel shaped cavity, FAS ACP (PB ID 1L0I) and ACP-mupA3a structures were compared by structural and sequence alignments. Figure 5.7 shows the superimposed structures of the reference FAS ACP (PDB ID 1L0I) and apo ACP-mupA3a WT. Two noticeable differences in the residue positions were observed which might be associated (amongst others) with the tunnel shaped cavity found in the FAS ACPs. Residue 59 on helix III of the FAS ACP is an alanine, but is equivalent to I61 in the ACP-mupA3a. In Figure 5.7 (B) the bulky isoleucine can be seen as a hindrance in the path leading to the deep tunnel. Similarly A34 in the FAS ACP is equivalent to L36 in ACP-mupA3a, this position corresponds to the residue between G and D in the GXDS motif and is usually a bulky residue in PKS ACPs.

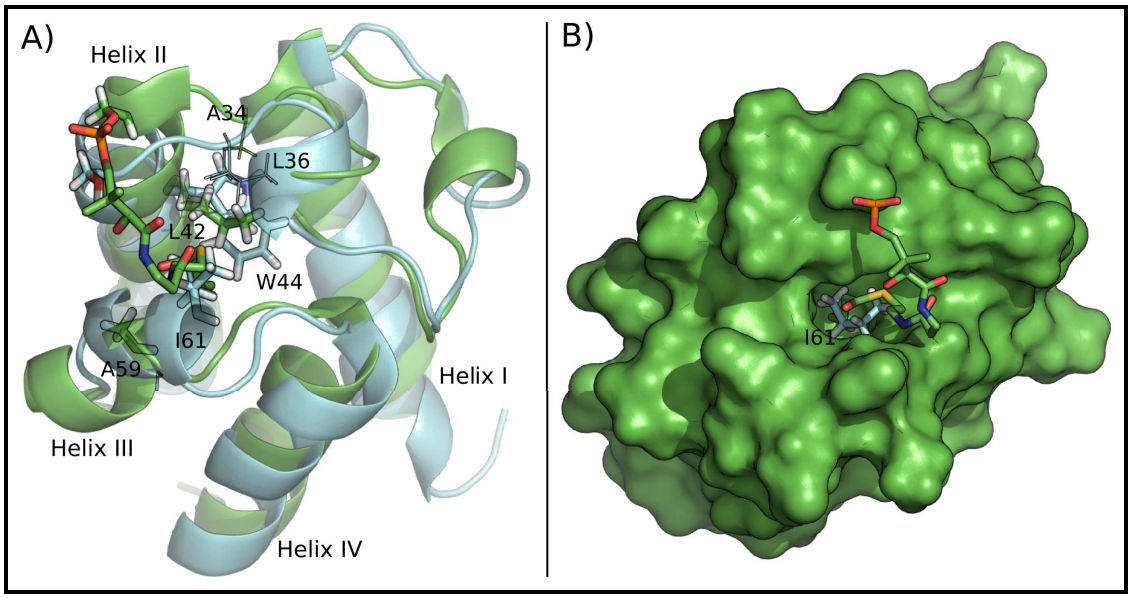

Figure 5.7: *Structural comparison of an FAS ACP (PDB ID 1L0I; green) and an apo ACP-mupA3a WT (cyan). (A) positions A59/I61, L42/W44 and a butyryl molecule attached to the phosphopantetheine in the FAS ACP are drawn as sticks and position A34/L36 as lines. (B) surface drawing of the FAS ACP (green) highlighting the orientation of the ligand and opening of the tunnel shaped cavity I61 (cyan sticks) can be seen as a blockage in the tunnel.*

Table 5.4: Correlation between cavity volume and RMSD / hydrogen bonds.

Simulation	Correlation between cavity volume and RMSD from the FAS ACP		Correlation between cavity volume and hydrogen bonds with the protein		Correlation between the hydrogen bonds formed by the ligand with protein and solvent	
	Wild	W44L	Wild	W44L	Wild	W44L
Apo ACP-mupA3a (200ns)	-0.241	-0.309	-	-	-	-
Apo ACP-mupA3a (1μs)	-0.241	-	-	-	-	-
Holo ACP-mupA3a	-0.225	-0.601	0.041	0.021	-0.433	-0.222
Acyl ACP-mupA3a (200ns)	-0.111	-0.435	0.164	0.015	-0.653	-0.323
Acyl ACP-mupA3a (1μs)	-0.245	-	0.203	-	-0.429	-
Acyl 14C ACP-mupA3a	-0.538	-	-	-	-	-
Acyl ACP-muA2a	0.143	-	0.012	-	-0.207	-

In order to understand the sequence conservation of these positions in the FAS and PKS ACPs two sets of similarity searches were carried out using PHI BLAST against the NCBI's non redundant protein database. The first search used the sequence of FAS ACP (PDB ID 1L0I) as the query along with the motif GADS, a highly conserved motif in the FAS ACPs. This search matched 1055 unique sequences with 95% query coverage or greater (sequences with large insertions were removed from the alignment). These 1055 unique sequences represented species from 473 genera with an average sequence identity of 65%. Figure 5.8 shows the sequence logo built on the multiple sequence alignment of the 1055 unique FAS ACP sequences. A59 of helix III was found to be absolutely conserved in all the sequences, along with the two highly conserved glutamic acid residues immediately before and one highly conserved glutamic acid residue immediately after position 59 (i.e. an EEAE motif). Searching the FAS ACP sequence against the NCBI's non redundant protein database with the EEAE motif yielded 472 unique sequences with at least 95% query coverage. Figure 5.9 shows the sequence logo built using the multiple sequence alignment of the 472 unique FAS ACP sequences with the EEAE motif. It was interesting to observe that by searching with the GADS motif A59 was found absolutely conserved and while searching with the EEAE motif A34 was found absolutely conserved as well. Comparing the two sequence logos one can say that the FAS ACPs are highly conserved with an average sequence identity of 65% and 76% when searched with the GADS and EEAE motifs respectively. The FAS ACP sequence was also searched against the PDB database with the GADS and EEAE motifs, which matched 11 and 8 FAS ACP structures respectively with the similar sequence conservation pattern as shown in sequence logos in Figures 5.8 and 5.9.

Searching with the *E. coli* FAS ACP sequence (PDB ID 1L0I) without any additional motif constraint against NCBI's non redundant protein database matched 2078 unique sequences with 95% or greater query coverage. Alignment of these sequences revealed three broad groups: group one contains the GADS motif (1280 sequences), group two contains a GLDS motif instead of the GADS motif (257 sequences) and group three has neither a GADS or a GLDS motif (541 sequences). The third group also lacked the conservation of G in the GXDS motif. Figures from C.47 to C.49 in Appendix III show the sequence logos built on the above mentioned three

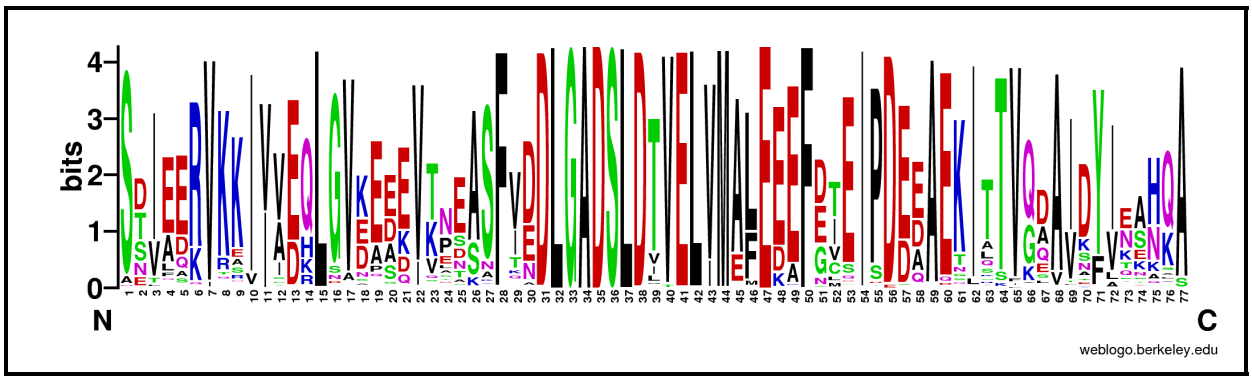

Figure 5.8: *Sequence logo built on 1055 unique FAS ACP sequences containing a GADS motif (position 34-36).*

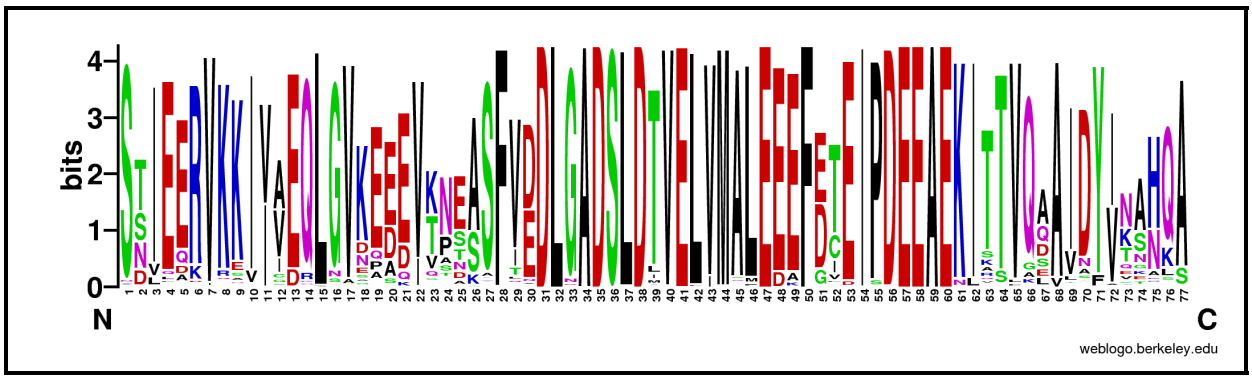

Figure 5.9: *Sequence logo built on 472 unique FAS ACP sequences containing a EEAE motif (position 57-60).*

sets of sequences. The sequences which carry a GADS motif show a high conservation of alanine at the 59th position of the FAS ACP structure. The sequences which carry a GLDS motif have a position 59 with predominantly A, L and R but also other residues. The alignment/logo of FAS ACP sequences which did not carry either of these motifs indicates poor conservation at most of the positions.

FAS ACP sequences carrying a GLDS motif are more like PKS ACPs with a bulky residue at the X position in the GXDS motif and another bulky reside at the equivalent position of I61 in ACP-mupA3a, equivalent to A59 in the FAS ACP structure discussed in the previous paragraph. Figures 5.10 and 5.11 show the sequence logos built on the ACPs from 15 well characterised PKS clusters grouped into ACPs associated with β-branching and standard ACPs as used in Chapter 3. Both the sequence logos show the presence of a bulky residue at the X position in the GXDS motif, which co-occurs with the bulky reside at the equivalent position of I61 (ACP-mupA3a). These observations suggest that having a less bulky residue at the position A59/I61 is correlated with a sequence being an FAS ACP and presumably allows the formation of a deep tunnel shaped cavity in the FAS ACPs. Owing to a high level of sequence conservation in the FAS ACPs this structural feature would be shared widely.

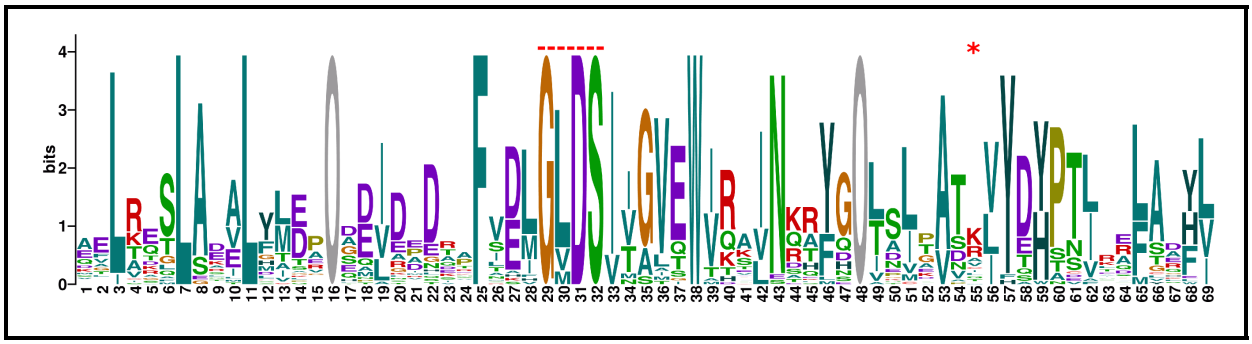

Figure 5.10: *Sequence l ogo built on the β-branching ACP sequences from 15 well characterized polyketide synthase clusters. - - - - - indicates the GXDS position and * indicates the position 55 equivalent to A59 in the FAS ACPs or I61 in MupA.*

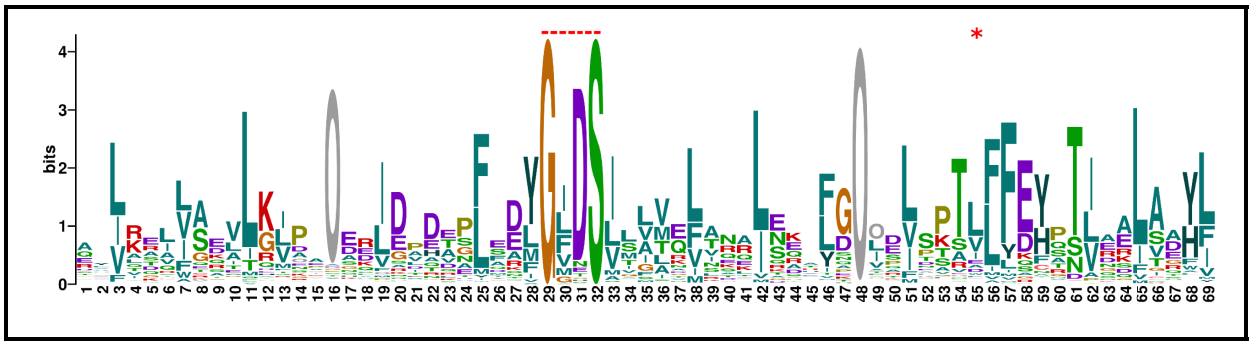

Figure 5.11: *Sequence logo built on the standard ACP sequences from 15 well characterized polyketide synthase clusters.* - - - - - *indicates the GXDS position and* * *indicates the position 55 equivalent to A59 in the FAS ACPs.*

5.3 Discussion

In order to investigate the possible formation of a cavity and the sequestering of phosphopantetheine and acyl chains in the ACP core, a series of molecular dynamics simulations were performed on the apo, holo and acyl forms of the ACPs from the MmpA module of the mupirocin cluster (ACP-mupA3a and ACP-mupA2a). Experimentally determined ACP structures and molecular dynamics studies on FAS ACPs have shown the formation of a hydrophobic tunnel within the ACP core, which sequesters the growing saturated fatty acid chain, protecting it from the outer hydrophilic environment, but this had not been seen for PKS ACPs. Chan *et al.* (2008) have shown that an optimal acyl chain length from eight carbons up to ten carbons can be stably accommodated by the *E. coli* FAS ACP. In this chapter, the acyl chains were of 12C and 14C backbone atoms long. Three independent simulations of 50 ns each were set up for all the structures namely apo ACP-mupA3a (WT and W44L), holo ACP-mupA3a (WT and W44L), acyl ACP-mupA3a (WT and W44L), acyl 14C ACP-mupA3a and acyl ACP-mupA2 were setup. The starting structures for the holo as well as the acyl structures had solvent exposed ligands attached to the catalytic serine. Chan *et al.* (2008) ran 50 ns simulations for the structures with solvent exposed acyl chains, the time scale was enough for acyl chains of four, six, eight, ten and sixteen carbons in length to find their way into the cavity. Acyl chains with twelve, fourteen and eighteen carbons failed to find the cavity even in extended simulations.

Here, three independent 50 ns simulations per holo and acyl structure seem to be a good

starting point for sampling enough phase space for the ligands to find their way into the proposed cavity. However, as there was no experimental evidence of cavity formation and sequestering of acyl chains in the PKS systems, there was no clear idea of how long a simulation should be to see the sequestering of the ligand, if indeed sequestering occurred. Therefore, one of the multiple 50 ns simulations that showed a transition of the acyl chain into the proposed cavity space was extended upto 200 ns for apo and holo ACPs, with one simulation of the wild type ACP-mupA3a charged with its cognate substrate simulated for 1μs. The simulation of the apo ACP-mupA3a WT was also extended to 1 μs as a control.

The root mean square deviation (RMSD) of the backbone atoms of the simulated structures from their reference starting structure remained under 2.5 Å, suggesting that the structures in apo, holo or acyl forms did not undergo any extreme unfolding events. This observation was crucial to test the stability of the holo and acyl forms since the solved ACP-mupA3a structure did not have the coordinates for the phosphopantetheine or for any of the acyl chains. The topology information for the phosphopantetheine and the acyl chains were built manually and the bond, angle, dihedral and van der Waals forcefield parameters were derived from the General Amber Force Field (GAFF), with partial charges derived by the RESP method, since these parameters were not in the GROMACS suite. The GAFF parameters and RESP charges were designed to be compatible with the AMBER99SB-ILDN forcefield.

Chan *et al.* (2008) utilized the GROMOS96 43a2 forcefield for their ACP simulation and most of the atom types in the acyl chain and parameters for the for the atom types which were not present in the parent GROMOS96 43a2 forcefield database were generated through GAUSSIAN quantum mechanics software. Observations from their simulations were similar to what was observed here. RMSF values showed a general trend of more fluctuation in the mutant structures as compared to their wild type counter parts, these observations are consistent with the MD simulations mentioned in Chapter 3, Section 3.2.2.1. Most of the fluctuation was observed from residues 34 to 40 (around the catalytic S38) and 52 to 68 (loop II and helix III).

A cavity was detected in more time frames during the holo and ACP simulations as compared to the apo ACP-mupA3a structures. The cavity volumes detected for the holo as well as

the acyl forms were also larger in size as compared to the apo structures. These observations suggested that the attachment of phosphopantetheine and acyl chains not only induced the formation of the cavity but that the attachment also influenced the size of the cavity formed. Upon utilizing the same parameters for the cavity detection the reference FAS ACP (PDB ID 1L0I) structure recorded a volume of 253 $Å^3$ which was comparable to the largest cavity detected for acyl ACP-mupA3a WT. However, the shape of the cavity for the FAS ACP was found to be like a deep tunnel as compared to the wide, shallow and surface exposed cavity in the PKS ACPs. A similar trend for wide, shallow and surface exposed cavities, as those seen in the ACP-mupA3a WT was seen in the other simulations of PKS ACPs performed here.

Chan *et al.* (2008) found that the average cavity volume continuously increases in size according to the size of the ligand for upto ten carbons and but for chains between 10 and 18 carbon long remained unchanged. Here the acyl chains are twelve and fourteen backbone carbons long and the cavity volumes seen in these simulations do not seem to vary with chain length. Interestingly the average size of the cavity induced by the phosphopantetheine alone was also of the same size as that induced by the fourteen backbone carbon long acyl form. Visual inspection of the MD simulation trajectory revealed that the phosphopantetheine occupies the same cavity space as the substrate occupies in the acyl ACP-mupA3a simulations, in the latter case the phosphopantetheine is in solvent. This observation may be because the size of the phosphopantetheine backbone is similar to the size of the ACP-mupA3a cognate substrate backbone (i.e. 14C). and because the shape of the cavity is surface exposed the highly polar phosphopantetheine can hydrogen bond with the solvent while still residing in the cavity.

The shallow surface exposed cavities also meant that no part of the acyl chains was buried inside the ACP but rather the acyl chains were in constant interaction with the solvent. This hypothesis was based on the thinking that the polyketide molecules usually have some polar groups which can hydrogen bond with the solvent and do not necessarily require a hydrophobic casing to segregate them from the solvent. This does not equally apply to the FAS ACPs because the acyl chains in fatty acid synthesis are usually saturated and non polar in nature (except the β-carbon processed intermediates) and would experience a constant repulsion from the solvent.

It would rather be energetically favourable for a highly hydrophobic chain to envelope itself within a hydrophobic core. In principle a hydrophobic chain without such sequestration inside the ACP, could also attach itself to other hydrophobic entities in the cell making it not available to the FAS proteins.

The hypothesis that the acyl chains were in constant interaction with the solvent was also supported by the observation that the SASA calculated for the phosphopantetheine and the acyl chains (once the ligands have associated themselves with the protein) over time was constant, avoiding any further dips which would have suggested a transition from a less to more buried state. Chan *et al.* (2008) showed a very low solvent accessible surface area (SASA) for the acyl chains upto eight carbons long, which increases rapidly for an acyl chain of ten to eighteen carbons long, as the FAS ACP could not shield completely a carbon chain more than ten carbons long. The lowest SASA value recorded for an acyl chain in the PKS ACPs was for acyl 14C ACP-mupA3a, which was ≈43 $Å^2$, ≈9.5 $Å^2$ more than the SASA value calculated for the 14C acyl chain by Chan *et al.* (2008). Not surprisingly the highest SASA value calculated was for the phosphopantetheine, which was ≈16 $Å^2$ more than the SASA for the 14C saturated acyl chain. The SASA calculations were also supported by the number of hydrogen bonds made between the ligands (phosphopantetheine and the acyl chains) and the protein/solvent. Phosphopantetheine as well as the acyl chains were constantly making hydrogen bonds with the solvent as well as the protein, although the number of H-bond to solvent did not correlate with the size of the cavity induced.

During the smulations, as the PKS ACPs become more like the FAS ACP structure the volume of the cavity increased, as evidenced by the RMSD between the PKS and FAS ACPs decreasing as the cavity volume increased (negative correlation) in all the structures except one. However, as the cavity formed in the PKS structures is wide, shallow and surface exposed, as compared to deep tunnel in the FAS ACP, it is clear that only part of the structure was changing to be like an FAS ACP. A stronger correlation between the FAS and W44L mutant ACP structures suggested that having a less bulky residue in the PKS ACP core might allow the PKS ACPs structures to move closer to the reference FAS ACP.

Comparing the PKS apo ACP-mupA3a WT structure with the reference FAS ACP, revealed I61 in the ACP-mupA3a (which is an A59 in the FAS ACP) blocking the entry of the tunnel. Multiple sequence alignment of FAS and PKS ACPs revealed that position I61/A59 is highly conserved with an alanine in the FAS ACP and a bulkier residue in the PKS ACPs. This position also seems to have co-evolved with the X of the GXDS motif (active site serine) in the PKS and FAS ACPs, with a bulky residue at the position X in the PKS ACPs and an alanine in the FAS ACPs. It would be interesting to mutate position I61 and/or X in the GXDS motif to alanine in the ACP-mupA3a and possibly monitor a deeper transition of the acyl chain through NMR.

The observations here support the hypothesis that the PKS ACPs do form a cavity upon the attachment of the phosphopantetheine and acyl chains. However, whether different chain lengths influence the size of the cavity was not tested. The cavity formed does not form a deep tunnel that envelopes the acyl chain and thus shields it from the solvent, but is solvent exposed which enables the polar groups on the acyl chain to hydrogen bond with the solvent. It was also seen that I61 in the PKS ACPs is bulkier than A59 in the FAS ACPs and prohibits the formation of a deep tunnel. In the absence of any experimental data in this regard another set of MD simulations in a different PKS system for example ACP from module 2 of DEBS system (PDB ID 2JU1) and ACP from curacin system (PDB ID 2LIU) could be performed to verify a similar mechanism of acyl chain sequestering as observed here. It would also be interesting to test whether an FAS ACP would envelope a polyketide molecule into its hydrophobic core in lieu of a fatty acid acyl chain. These studies might explain the dynamic mechanism through which small proteins like an ACP recognize/interact with various domains in a complex PKS machinery. The information gathered from such studies not only helps to gain a deeper understanding of protein structure and function in general but would also help to re-engineer PKS pathways for the production of novel compounds.

Chapter 6

Ligand specificity and dynamics in the *mup* cluster domains

6.1 Introduction

This chapter focusses on understanding two observations made in two different unrelated studies in the *mup* cluster. The first study builds on work conducted in Prof. Thomas' group by Dr. Joanne Hothersall in order to identify the tailoring enzyme responsible for the synthesis of the 6-hydroxyl of mupirocin. According to the position of the 6-hydroxyl (Figure 1.26) in the mupirocin molecule it was hypothesized that the enzyme responsible must be acting after MmpD. Dr. Hothersall proposed MupA may be the likely candidate and therefore in order to confirm her prediction she created two different mutants Δ*mupA* and *mupA* G127A, D134A. The point mutations were chosen by comparison of MupA with other oxygenases, e.g. PedJ, to identify the active site residues. MupA belongs to the reduced flavin mononucleotide ($FMNH_2$)-dependent oxygenase family of proteins that are responsible for the addition of a hydroxyl group to the substrates in many different metabolic pathways (El-Sayed *et al.* 2003). Bioassay and HPLC showed that the deletion strain had no biological activity and wasn't producing pseudomonic acid A; and that the *mupA* G127A, D134A mutant strain was essentially like wild type *Pseudomonas fluorescens* NCIMB 10586. The strains were sent to our collaborators in Bristol for analysis by HPLC, LC-MS and proton NMR, which confirmed these results. They

found that the deletion strain produced mupiric acid but not mupirocin H (Wu *et al.* 2008). Mupiric acid is thought to be formed by the release of an intermediate from module 4 of MmpD. Since no intermediates longer than mupiric acid were found this fits with the idea that MupA is acting after MmpD. However, it doesn't confirm that it is the 6-hydroxylase.

This observation raised a new question that, if MupA is responsible for 6-hydroxylation and the deletion strain therefore produces a product with no 6-hydroxyl (α-hydroxyl) and does not progress beyond MmpD, what prohibits the non hydroxylated molecule from proceeding further in the synthesis pathway. It was hypothesized that the KS-mupA2 in the second module of the MmpA subunit might be specific for the α-hydroxylated intermediate and that in the absence of a α-hydroxyl it does not catalyse Claisen condensation. To determine what in the KS-mupA2 might recognise the α-hydroxyl, a docking study was performed in which the modelled KS-mupA2 dimer was docked with the expected cognate mupirocin intermediate. The docking was performed to reproduce the decarboxylation stage of the Claisen condensation with the substrate attached to the catalytic cysteine in the KS active site and the extender molecule (malonate) attached to phosphopantetheine of the ACP, positioned ready for decarboxylation and subsequent elongation of the substrate. The cognate ACP-mupA2 was docked attached to the phosphopantetheine with distance restraints to ensure that the phosphopantetheine was stretched along the active site tunnel. Key residues for hydroxyl recognition as suggested by docking and sequence analysis were investigated by mutagenesis experiments carried out by Miss Yousra Alsamarraie from Prof. Thomas' group.

The second study discussed here arose from observations during the molecular dynamics simulations of an ACP-mupA3a:MupH complex described in Chapter 4, Section 4.2.5. The 50 ns long simulations of the complex revealed a movement in two loop regions covering the MupH active site. These loop movements suggested that they might be assisting in the accommodation of the ligand upon ACP-mupA3a/b docking to MupH. In order to understand this phenomenon, the variation in distance between the two loops was measured during the ACP-mupA3a:MupH simulations as well as simulations of the, ACP-mupA3a:MupH dimer, MupH wild type monomer and MupH monomer C115 acetylated. The idea behind simulating the

ACP-mupA3a:MupH dimer was to see if the dimeric state of MupH (if it exists as a dimer) would affect the loop movements. The MupH monomer acetylated at C115 was simulated to see if the acetylation of the catalytic cysteine in the MupH alone would also trigger the same loop movements. Simulation of the non-acetylated wild type MupH monomer was used as a control.

6.2 Results

6.2.1 A loop at the KS dimer interface appears to be responsible for the substrate specificity

In order to understand what might be responsible for KS-mupA2 not to accepting the acyl chain that lacks the α-hydroxyl (6-hydroxyl in the final mupirocin molecule), the expected natural substrate of KS-mupA2 was docked into a homology model of the KS-mupA2 dimer structure (as described in Section 2.3.4.2). The KS-mupA2 dimer along with the docking domains was modelled using the X-ray structure of the KS-AT dimer from DEBS module 3 (PDB ID 2QO3, (Tang *et al.* 2007)), as described in Section 2.3.1.3. Docking was performed to reproduce the decarboxylation stage of the Claisen condensation reaction mechanism described in Section 1.2.4.3. KS dimer crystal structures show that the active site is made from atoms from both the KS subunits, which was thus also seen in our MupH model and MupH residues within a 5 Å radius of the docked α-hydroxyl included a motif D128, N129, Y130, K131 on the loop of the subunit not covalently carrying the substrate.

Figure 6.1 shows the overall docked complex with the lowest energy. Figure 6.2 shows the closeup view of the loop with the residues within 5 Å radius of the α-hydroxyl highlighted. Figure 6.2 also shows the backbone NH atoms of residues C158 and A403 which were used to restrain the acyl chain close to the catalytic cysteine (C158) in order to mimic the oxyanion hole during acyl chain transfer. Residues F219, H293 and H333 were used to restrain malonate within the active site in order to mimic the decarboxylation step, since these are implicated in decarboxylation as discussed in Section 1.2.4.3. The catalytic serine S38 of ACP-mupA2 was

restrained to be within 2 Å of the phosphate of the phosphopantetheine.

A multiple sequence alignment (Figure 6.3) of the KS sequences from the mupirocin and thiomarinol clusters revealed that the DNYK motif was conserved in KSs of the second module of MmpA and its equivalent TmpA but not in the other KSs. On the whole much less conservation was found among the other positions in the loop connecting an α-helix at the N-terminus and a β-strand at the C-terminus (Figure 6.3). Based on this observations it was hypothesized that if this loop were swapped with the loops from KS-mupA1 or KS-mupA3, which do not require a substrate with an α-hydroxyl, then the KS mutant might allow the pathway to proceed further. Figure 6.4 shows the loop region of interest on the KS-mupA2 and the sections of KS-mupA1 and KS-mupA3 to replace it. Miss Yousra Alsamarraie carried out suicide mutagenesis to swap the loops and HPLC to detect antibiotic production.

At the time of writing this thesis Miss Y. Alsamrarraie was able to replace the loop from KS-mupA2 with the loop of KS-mupA1 in *P. fluorescens* NCIMB 10586 wild type and *P. fluorescens* Δ*mupA* strains. The blue line in Figures 6.5 and 6.6 show the HPLC trace for the *P. fluorescens* Δ*mupA* strain which was used as the control, there is no peak at 21 mins, which would correspond to pseudomonic acid A production. The red line in Figure 6.5 shows the HPLC trace for the KS-mupA2 loop replaced with the loop from KS-mupA1 expressed in the *P. fluorescens* Δ*mupA* strain. The trace shows a small peak at around 22:30 mins, which suggests that the substrate was processed by KS-mupA2 however, the metabolite produced still needs to be characterised. The red line in Figure 6.6 shows the HPLC trace for the KS-mupA2 loop replaced with the loop from KS-mupA1 expressed in *P. fluorescens* NCIMB 10586 wild type strain. This trace also shows a small peak at around 22 mins, suggesting that the KS-mupA1 loop is tolerant to both the substrates with and without α-hydroxyl. These samples have been sent to our collaborators in Bristol for LC-MS analysis to determine the structures of the metabolite produced.

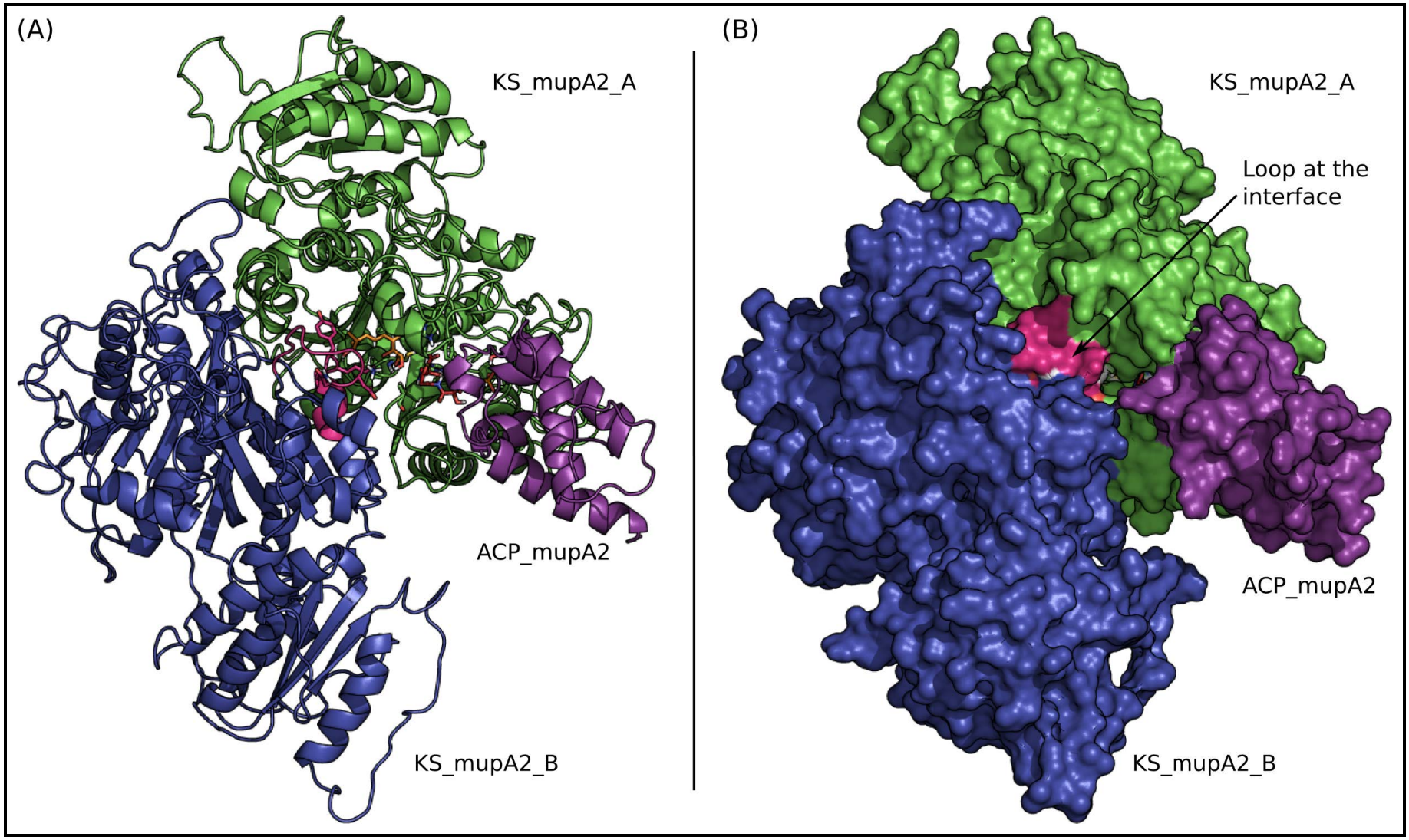

Figure 6.1: *Cartoon and surface drawings of the KS-mupA2 dimer with ACP-mupA2 docked. Chains A and B of the KS-mupA2 dimer are coloured green and blue respectively with their interface loop coloured pink, ACP-mupA2 is coloured purple. (A) cartoon rendering with phosphopantetheine and polyketide intermediate shown as sticks surrounded by key catalytic residues a zoomed in view of this can be seen in Figure 6.2 (B) surface rendering.*

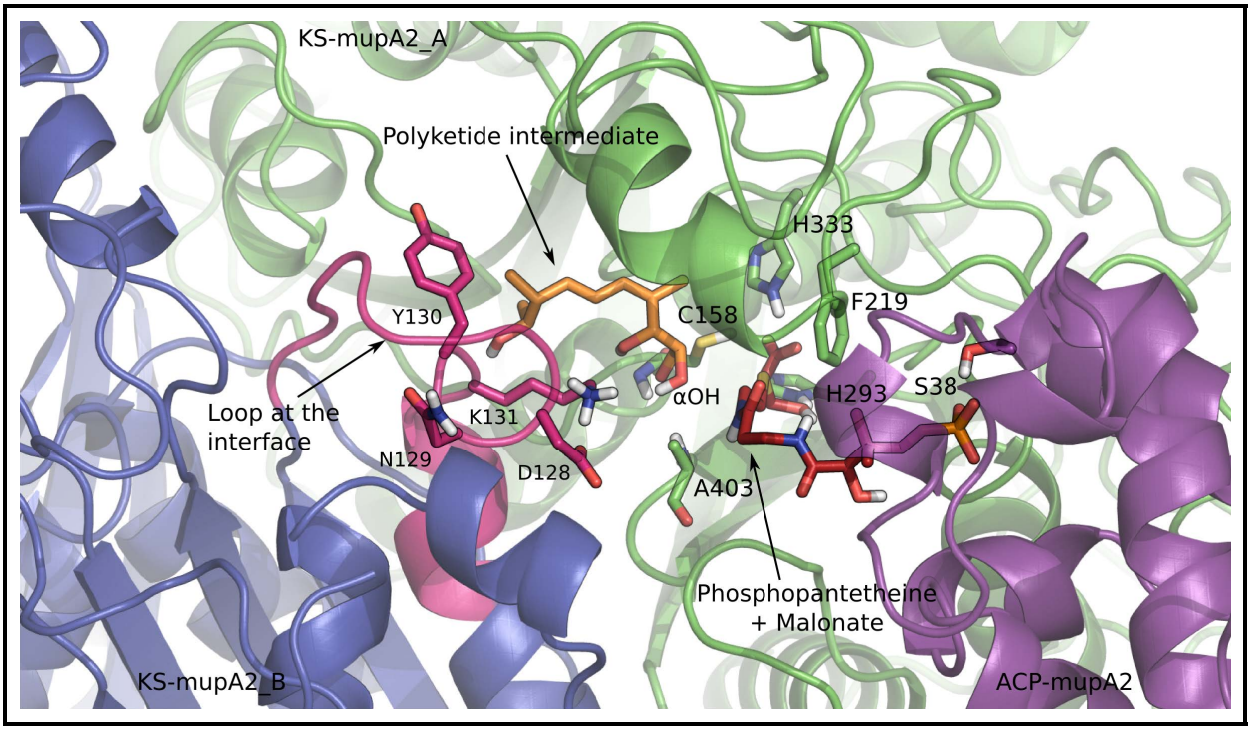

Figure 6.2: *Close up view of the KS-mupA2 dimer and ACP-mupA2 docking interface. This configuration of substrates represents the decarboxylation/condensation stage of the KS reaction mechanism. Chains A and B of the KS-mupA2 dimer are coloured as green and blue respectively with their interface loop coloured pink, ACP-mupA2 is coloured purple. The DNYK motif is shown as sticks. The malonate extender unit attached to the phosphopantetheine was docked inside the KS-mupA2 active site with three distance restraints requiring it to be 3 Å from each of F219, H293 and H333. The distance between phosphopantetheine and S38 of ACP-mupA2 was restrained to be 2 Å. The natural substrate of KS-mupA2 was docked inside the active site with three distance restraints of 1.8 Å to the sulphur of C158, and 2.93 Å each to the backbone NH of C185 and A403. The substrate is bound to chain A of the KS whereas the D128NYK motif is predicted to recognise the substrate is on Chain B.*

```
KS-mupA1     DWKGSAT--------GVFIAAERNEY--HLNLLQAQIDPGEGLDQAASMLANRVSHFYDL
KS-mupA2     SLEQEKV--------GVFVGVSKAGH-DNY-------------KDSFFSIANRVSYRFGF
KS-mupA2SS   ----EEE--------EEEE--------------------------HHHHHHHHHHH----
KS-mupA3     GLAASARHEDAEGMVGVFVGVTYEEY-QLYGAQQTAEGRPLVLSMSPSSIANRVSFVNGF
KS-mupD1     ALAGSRT--------GVFVAAFNYDYKQLLESAGLPIDAHHSTGNAAAVIANRISHFYDL
KS-mupD2     QRRAEV---------SVYVGCEQGDYDQLFDDMPPPQSFW---GNAPSIVPARIAYYLDL
KS-mupD3     ALRRT----------GVYVGVMYQEY-QLFGAEQTLLGRPMALSGSSASIANRVSWSLGL
KS-mupD4     ATEGQPV--------GVYVGASASDYRSLFAEAAPAQAFW---GNASSIIPARIAYHLDL
KS-tmpA1     SLRGSNT--------GVFVGFERNEY--LLNLIESGHDTGESLHQSDSMIANNISYFFDF
KS-tmpA2     ELESDTV--------GVFVGISKAGF-DNY-------------KDSYFSAANRISYRFNF
KS-tmpA3     ALAQSTMIDKTPGVVGVYVGVTYQEY-QLYGAQETMLGNPMVLSMSPSSIANRVSYFCDF
KS-tmpD1     TLSGQNV--------GVYIGAFNFDYKELLEKHERPIEAYQSTGTANAIIANRISHFYDF
KS-tmpD2     SIDGNAV--------SIYVGCEQGDYERLFADSPLPQSFW---GNAPSIIPARLSYHLNL
KS-tmpD3     SKLGDHV--------GVFVGVMYSEY-QLYGAQQGVLGTPISLGGSAASIANRVSFSLNF
KS-tmpD4     PEFGENA--------GVYVGCNSGDYKNILSDEAPAQAFW---GNAGSITPARLSYYLNL
```

Figure 6.3: *Portion of the multiple sequence alignment of the KS domains from mupirocin (MmpA/D) and thiomarinol (TmpA/D) clusters.*

```
KS-mupA2   ---VGVFVGVSKAGHDNY----------KDSFFSIANRVSYRFGFTGPSLPVDTACS
KS-mupA1   ---TGVFIAAERNEYHLNLLQAQIDPGE-GLDQAASMLANRVSHFYDLRGPSERIDAMCA
KS-mupA3   EGMVGVFVGVTYEEYQLYGAQQTAEGRPLVLSMSPSSIANRVSFVNGFHGPSMAIDAMCA
```

Figure 6.4: *Region of the loop on the KS-mupA2 which was proposed to be swapped by the loops from KS-mupA1 and KS-mupA3.*

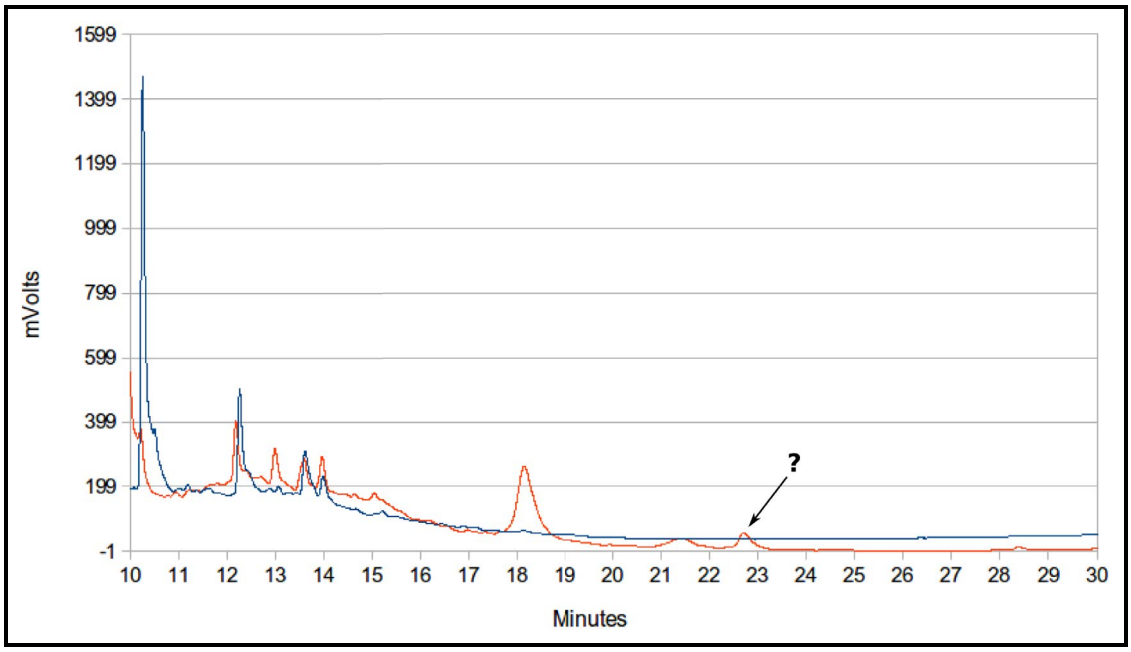

Figure 6.5: *HPLC trace for the KS-mupA2 loop replaced with the loop from KS-mupA1 in P. fluorescens ΔmupA strain, shown in red. The blue line represents the HPLC trace for P. fluorescens ΔmupA strain used as control.*

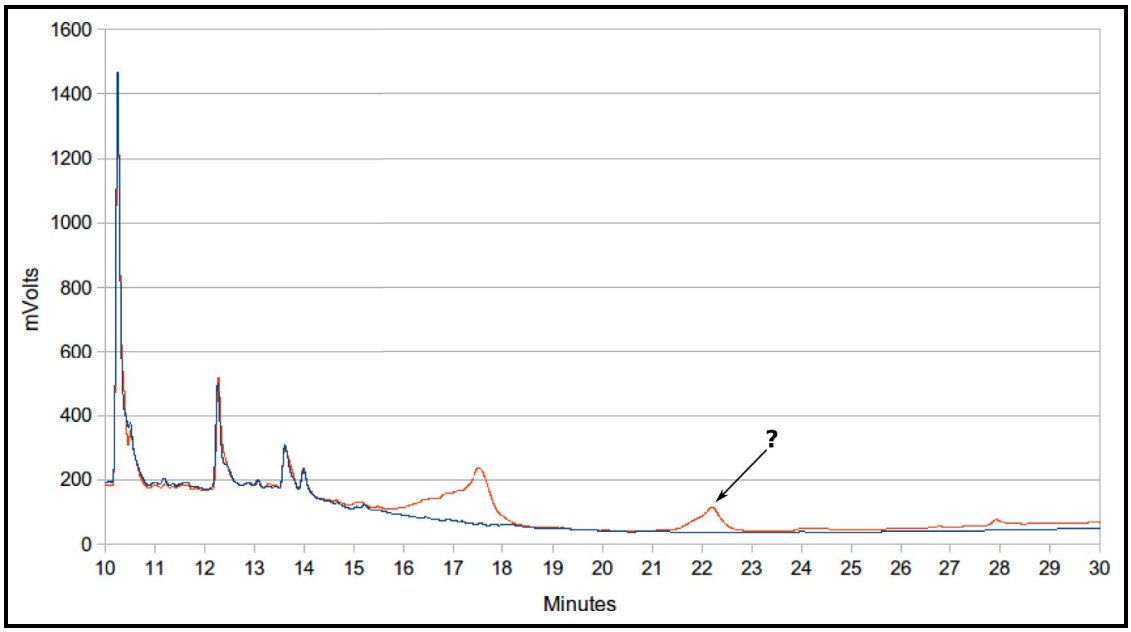

Figure 6.6: *HPLC trace for the KS-mupA2 loop replaced with the loop from KS-mupA1 in P. fluorescens NCIMB 10586 wild type strain, shown in red. The blue line represents the HPLC trace for P. fluorescens ΔmupA strain used as control.*

6.2.2 Movement of MupH surface loops may have a role in ligand binding

Three independent molecular dynamics simulations of the ACP-mupA3a:MupH complex with the bound substrate, each of 50 ns length, revealed large movements in two loop regions causing a separation between the loops which would otherwise cover the MupH active site tunnel, loop I and II in Figure 6.7. To further explore this observation, each of the three simulations were extended to 100 ns, as described in Section 2.3.2.7. Out of the two loops, loop II showed greater movement (Figure 6.8) and upon visualizing the movie for the simulation of the ACP-mupA2a:MupH complex the loop movement appears to aid the accommodation of the ligand in the active site by widening the active site tunnel (Figure 6.7). This observation also lead to the hypothesis that such a loop movement in MupH may be required to let a large ligand, like monic acid attached to a phosphopantetheine, to enter inside its active site. Plotting the average of the distances between P207 in loop II and each of L150, M151 and I152 in loop I, for simulation replicate 1 (Figure 6.9), as well as the average distances between D208 and S209 in loop II and L150, M151 and I152 in loop I revealed that the loops widen up to 3 nm during the first 30 ns and then narrow down to around 1 nm the remaining 70 ns. Replicate 2 did not show

any huge fluctuation in the distance and replicate 3 showed only moderate fluctuation on two different occasions along the trajectory (Figure 6.10 and 6.11). The procedure to measure the distances between the loops is described in the methods Section 2.3.2.13. Table 6.1 summarizes the different simulations setup for the analysis of loop movements in MupH.

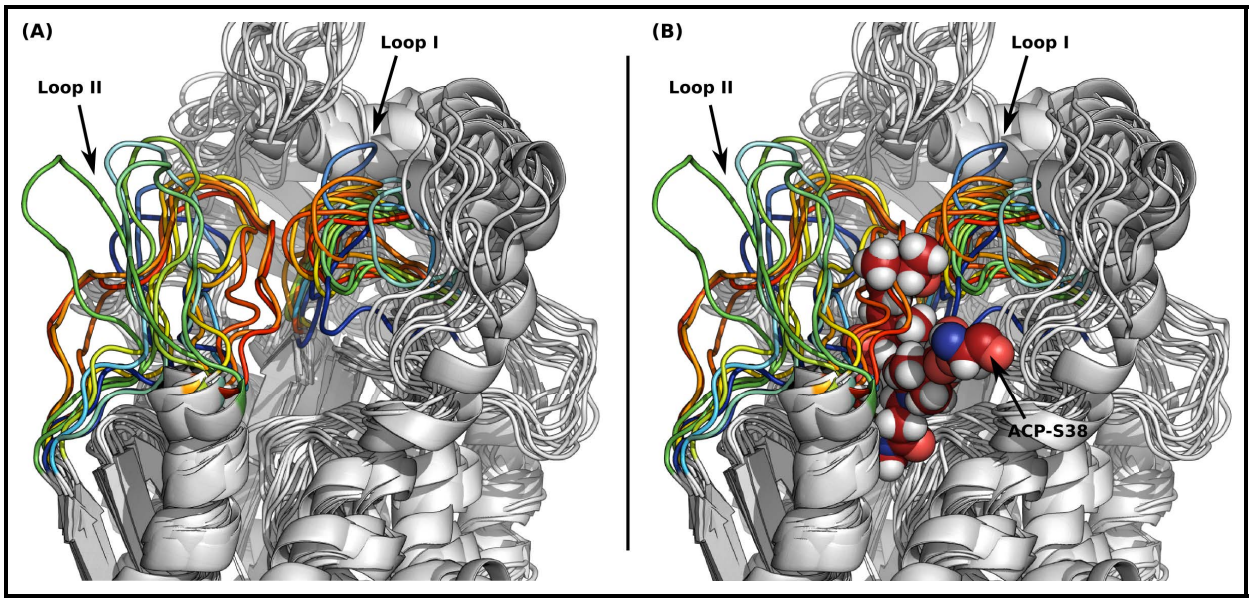

Figure 6.7: *Simulation of ACP-mupA3a:MupH monomer with substrate shows movements of loops over the MupH active site. (A) without and (B) with cognate substrate of MupH attached to the side chain of ACP-mupA3a S38 rendered as spheres. Snapshots at every 5 ns of the 50 ns molecular dynamics simulation revealed the movement in the loop I (residue 147-171) and II (residue 198-214) which may assist in the binding of the ligand. Loop II showed greater movement than loop I. Different coloured loops represent different snapshots.*

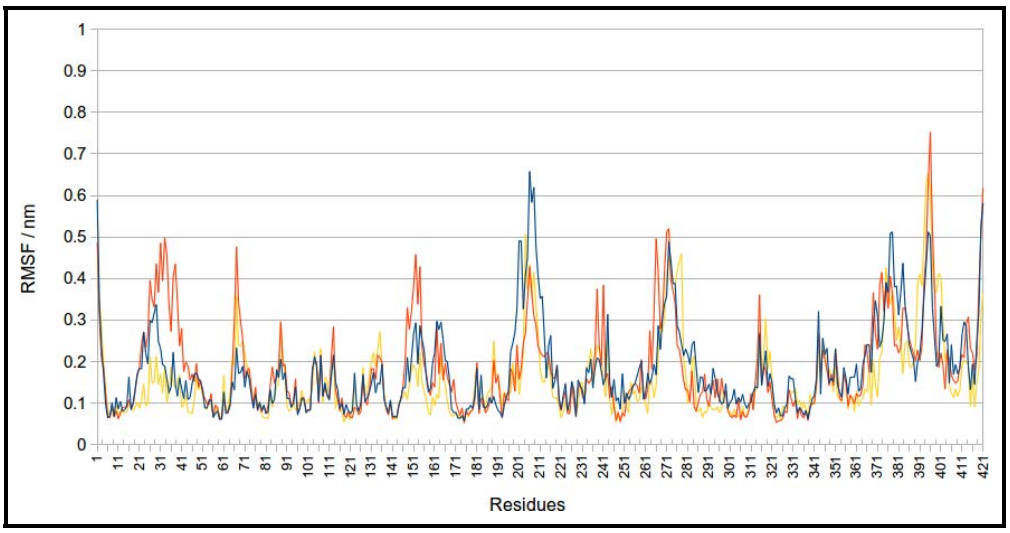

Figure 6.8: *MupH atomic RMSF for the ACP-mupA3a:MupH monomer simulation. RMSF values for the residues in loop II (residue 198-214) can be seen larger than the residues on the loop I (residue 147-171). Blue, red and yellow lines represents the three replicates respectively.*

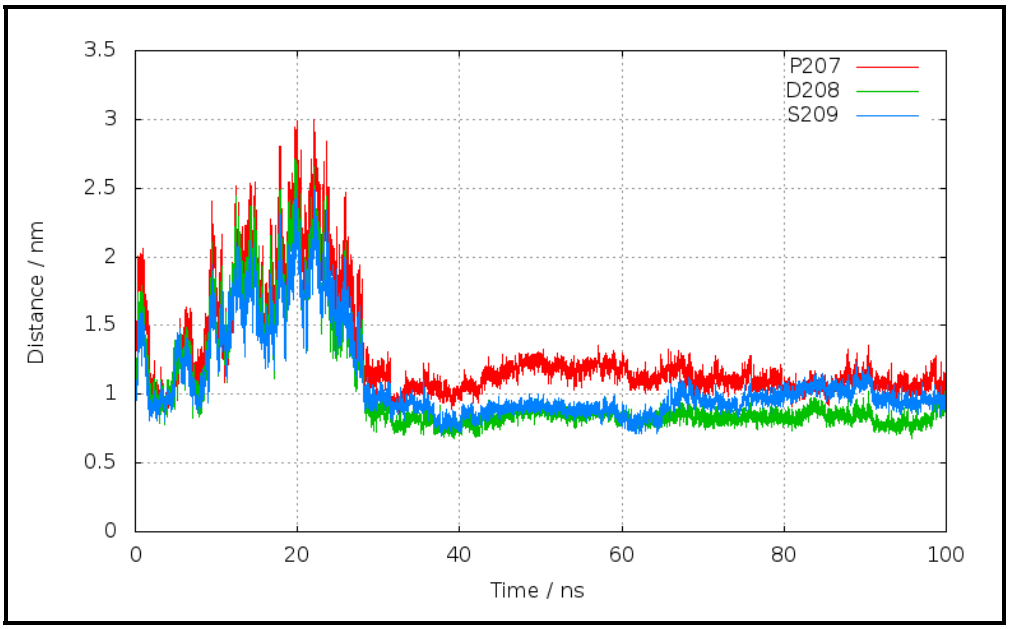

Figure 6.9: *Distances between residues in loop I and loop II for the ACP-mupA3a:MupH monomer complex simulation replicate 1. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I. The distance between the loops rise upto 3nm within 30 ns and then stabilized around 1nm.*

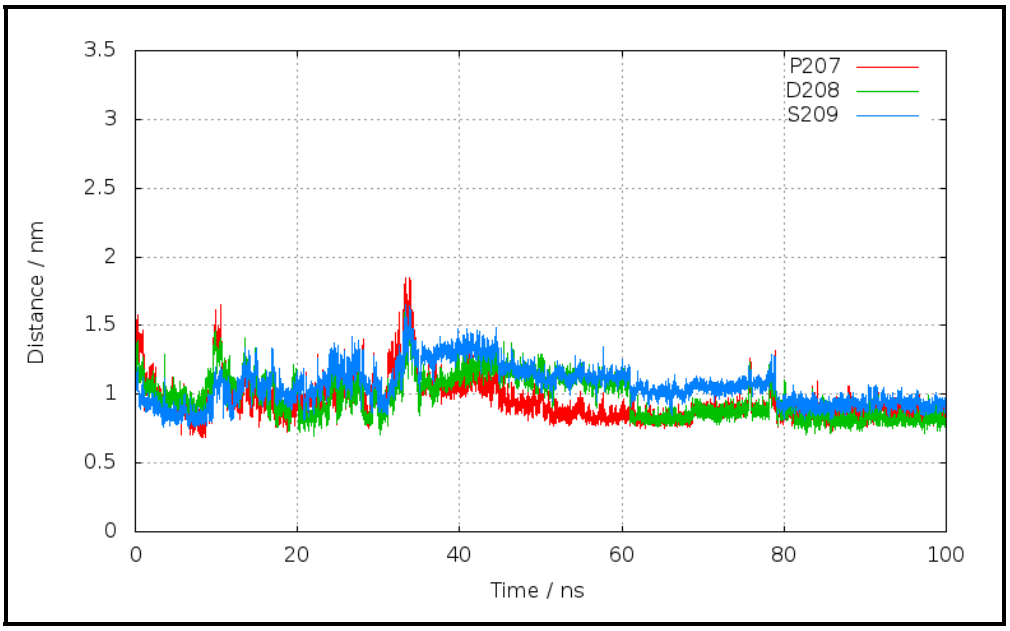

Figure 6.10: *Distances between residues in loop I and loop II for the ACP-mupA3a:MupH monomer complex simulation replicate 2. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

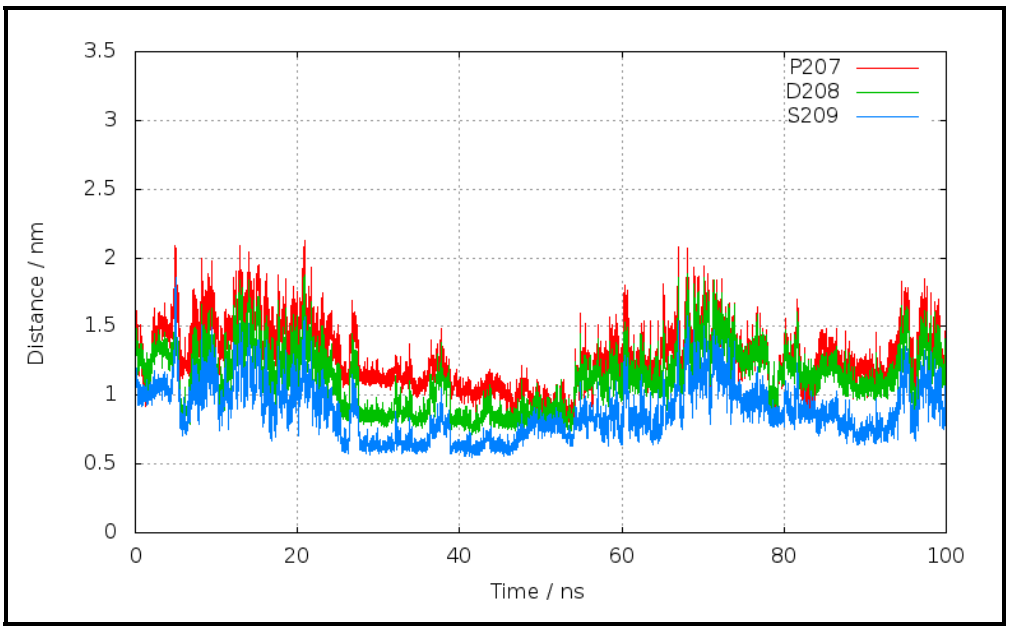

Figure 6.11: *Distances between residues in loop I and loop II for the ACP-mupA3a:MupH monomer complex simulation replicate 3. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

Table 6.1: *MupH simulation setup summary*

Structure	**Ligand/modification**	**Simulation**	**Methods Section**
ACP-mupA3a:MupH monomer	MupH C115 acetylated ACP-mupA3a cognate substrate	3 X 100 ns	2.3.2.7
ACP-mupA3a:MupH dimer	MupH C115 acetylated ACP-mupA3a cognate substrate	3 X 50 ns	2.3.2.7
MupH monomer WT	-	3 X 50 ns	2.3.2.8
MupH acetylated C115	C115 acetylated	3 X 50 ns	2.3.2.8

A similar simulation with three independent replicates of 50 ns each, was also carried out with MupH in the dimeric form together with ACP-mupA3a docked to one of the monomers of MupH (as described in Section 2.3.2.7). It is quite likely that MupH exists as a dimer since other HMG-CoA orthologues exist as dimers and PIER analysis of a MupH model suggested an interface consistent with that seen in the crystal structure (Chapter 3, Section 3.2.6.2). The RMSF calculated for all atoms averaged per residue of MupH in the simulation of the ACP-mupA3a:MupH dimer complex showed lower values for the loop II than loop I (Figure 6.12). The distance measured between the two loops in the simulation of the ACP-mupA3a:MupH dimer complex also showed slightly lowered values for the largest distance reached between the two loops as compared to the ACP-mupA3a:MupH monomer simulation. The distance between the two loops remained constant throughout 50 ns in all the three replicates averaging around 1 nm and 1.5 nm with a rise up to 2 nm in two out of three replicates (Figures from 6.13 to 6.15). Figure 6.15 shows replicate 3 for the ACP-mupA3a:MupH dimer simulation in which the distance between the loops reached and remained at 2 nm and stayed plateaued for around 15 ns. These observations support the hypothesis that the dimer formation hinders the loop II movement, both in terms of its fluctuation on its position as well as its relative distance from loop I. This could be because loop II, being close to the dimer interface, interacts with the residues from the other monomer and hence stabilizes its position. Simulating the MupH dimer for a longer time scale might allow better sampling of the large ACP-mupA3a:MupH dimer

complex which may not be enough within the 50 ns of simulations here. Due to lack of time it was not possible to run longer simulation on this complex.

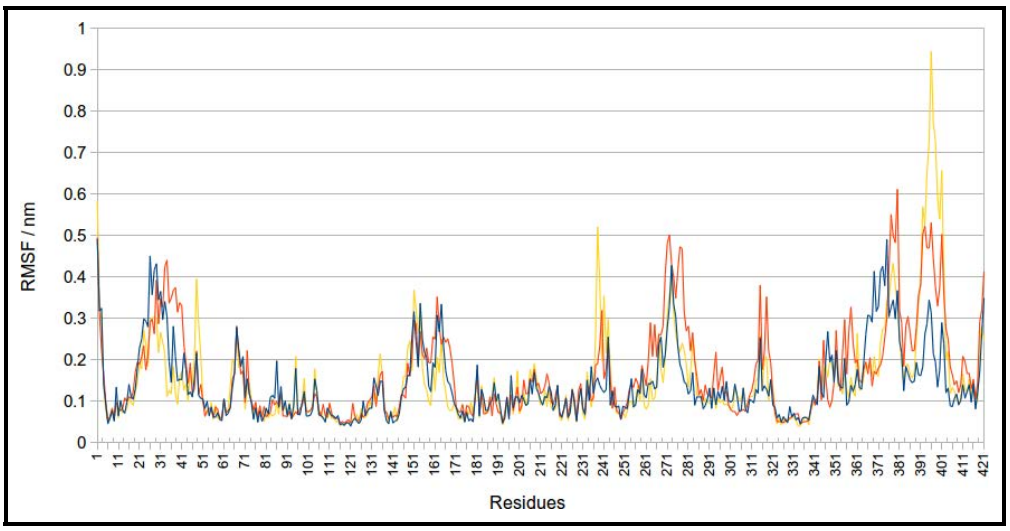

Figure 6.12: *ACP-mupA3a:MupH dimer complex, RMSF of all atoms averaged per residue of MupH. RMSF values for the residues in loop II (residue 198-214) can be seen smaller than the residues on the loop I (residue 147-171). Blue, red and yellow lines represents the three replicates respectively.*

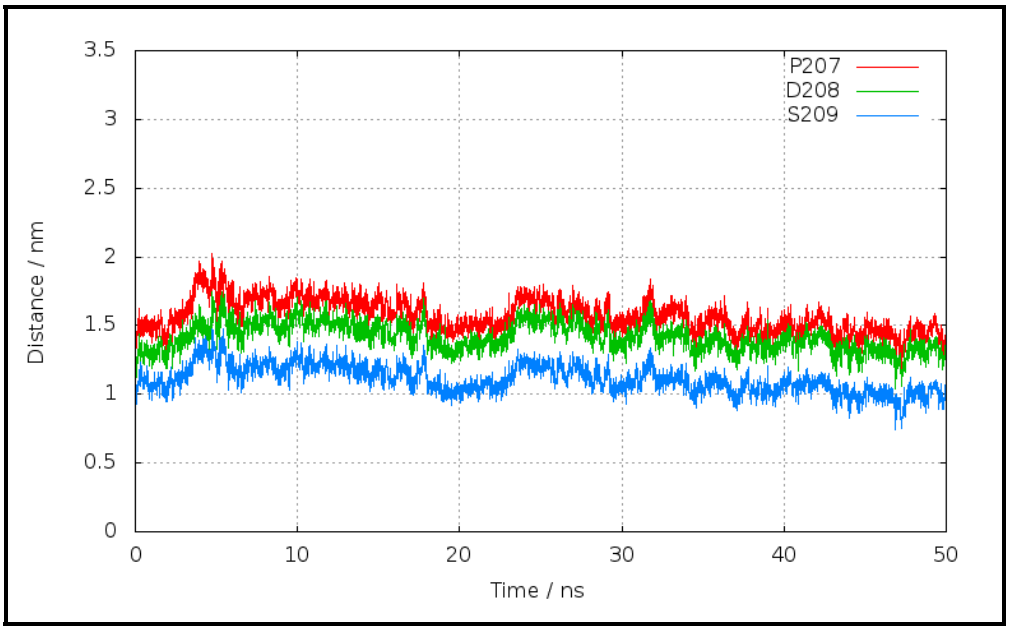

Figure 6.13: *ACP-mupA3a:MupH dimer complex simulation replicate 1, distance measured between the loop I and II over the time of 50ns. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

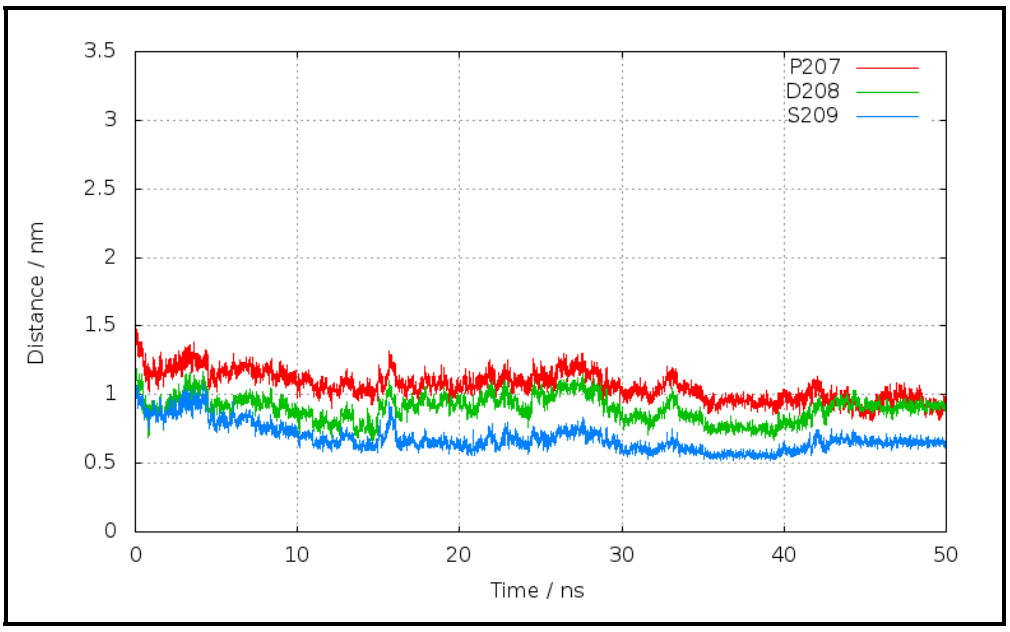

Figure 6.14: *ACP-mupA3a:MupH dimer complex simulation replicate 2, distance measured between the loop I and II over the time of 50ns. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

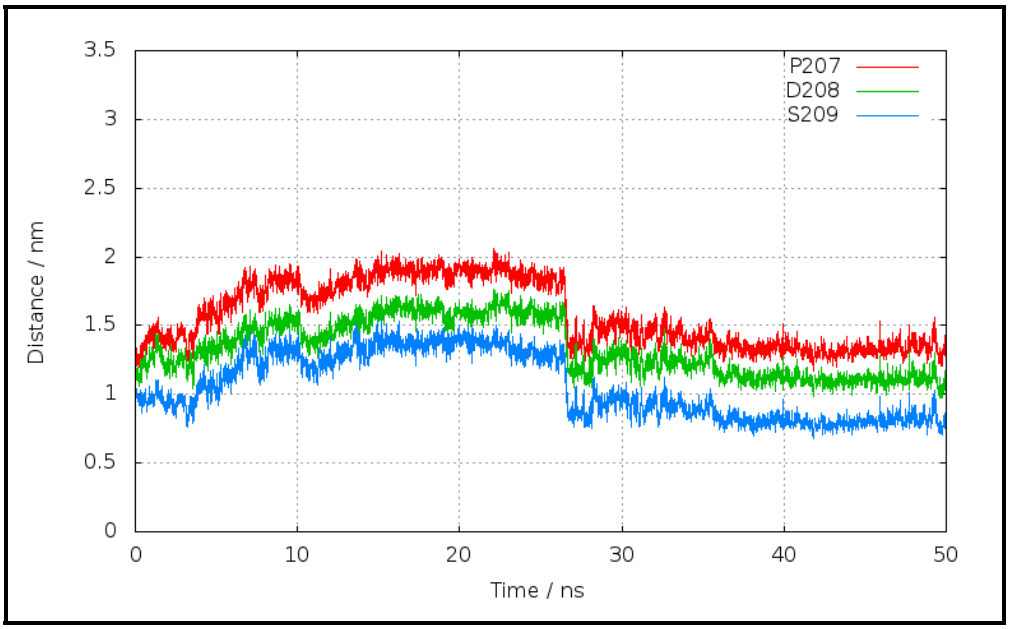

Figure 6.15: *ACP-mupA3a:MupH dimer complex simulation replicate 3, distance measured between the loop I and II over the time of 50ns. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

One hypothesis that the loop movement may be triggered by the acetylation of the catalytic cysteine in MupH, that would signal MupH to open up the loops to let the ligand in. To test this hypothesis three independent 50 ns simulations were setup for the MupH monomer with C115 acetylated (as described in Section 2.3.2.8). RMSFs calculated for all atoms averaged per residue of MupH do not seem to be very different between loop I and II (Figure 6.16). Figures 6.17 to 6.18 show the distances measured between the loops for the MupH monomer with C115 acetylated simulations. Distances measured in the three replicates remained below 2 nm averaging around 1 nm. These loop movements do not seem to be different from the movements seen in the ACP docked simulations.

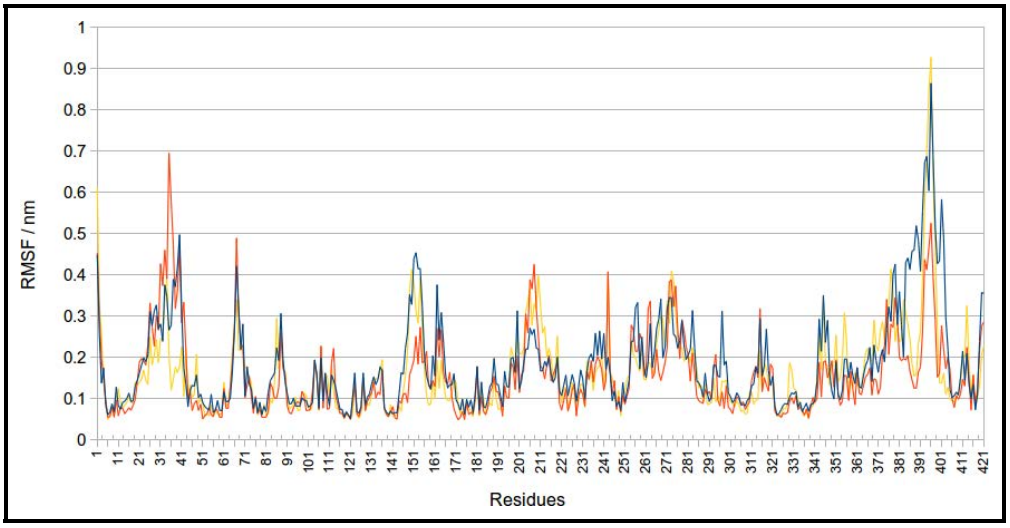

Figure 6.16: *MupH monomer with C115 acetylated simulation, RMSF of all atoms averaged per residue of MupH. RMSF values for the residues in loop II (residue 198-214) and loop I (residue 147-171) does not seem to be very different from each other. Blue, red and yellow lines represents the three replicates respectively.*

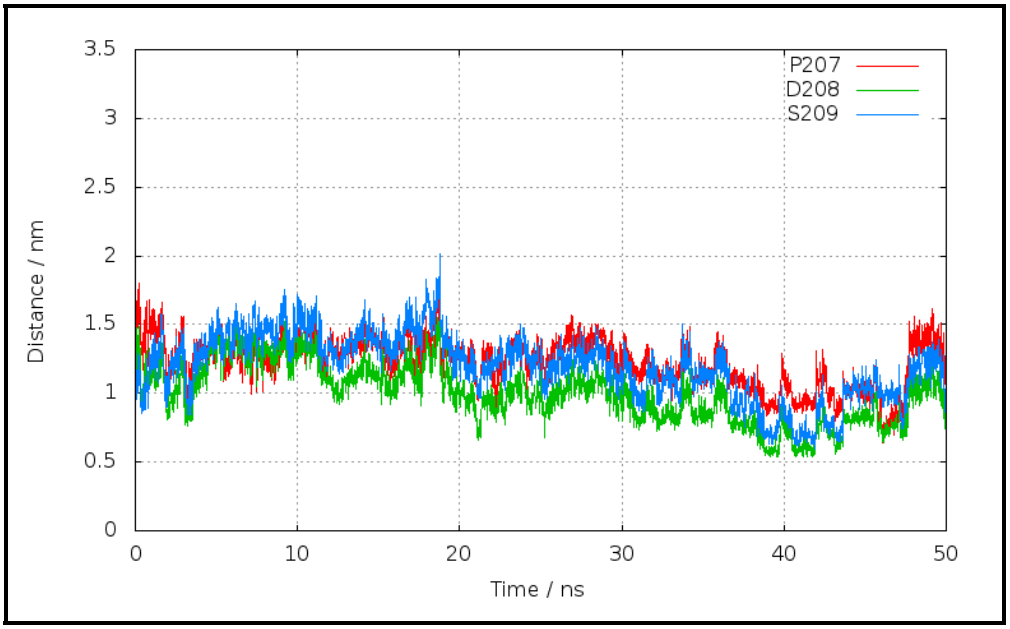

Figure 6.17: *MupH monomer with C115 acetylated simulation replicate 1, distance measured between the loop I and II over the time of 50ns. CYA was the acytylated cysteine. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

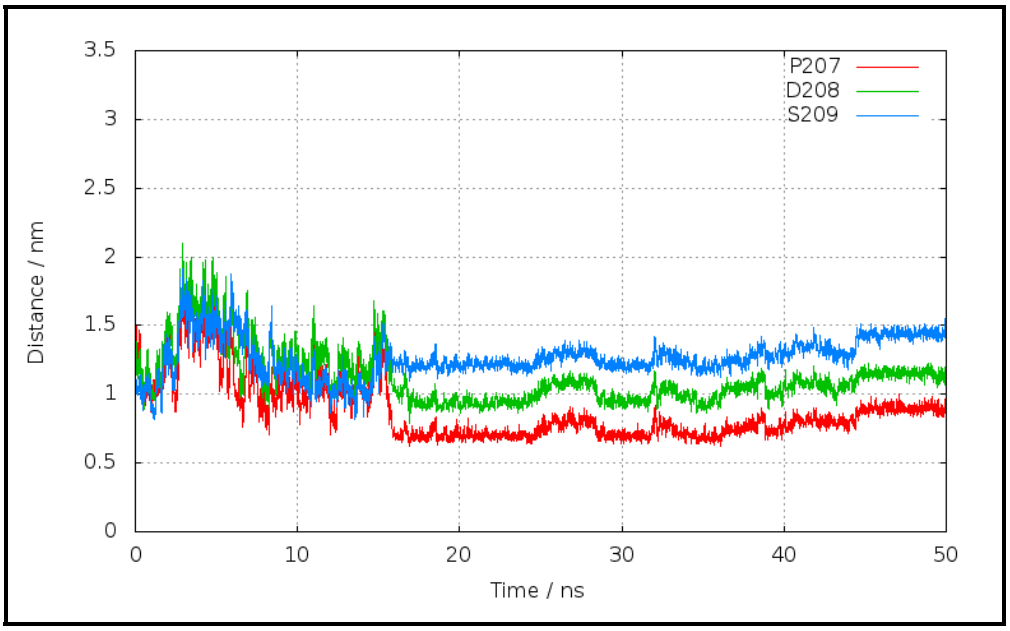

Figure 6.18: *MupH monomer with C115 acetylated simulation replicate 2, distance measured between the loop I and II over the time of 50ns. CYA was the acytylated cysteine. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

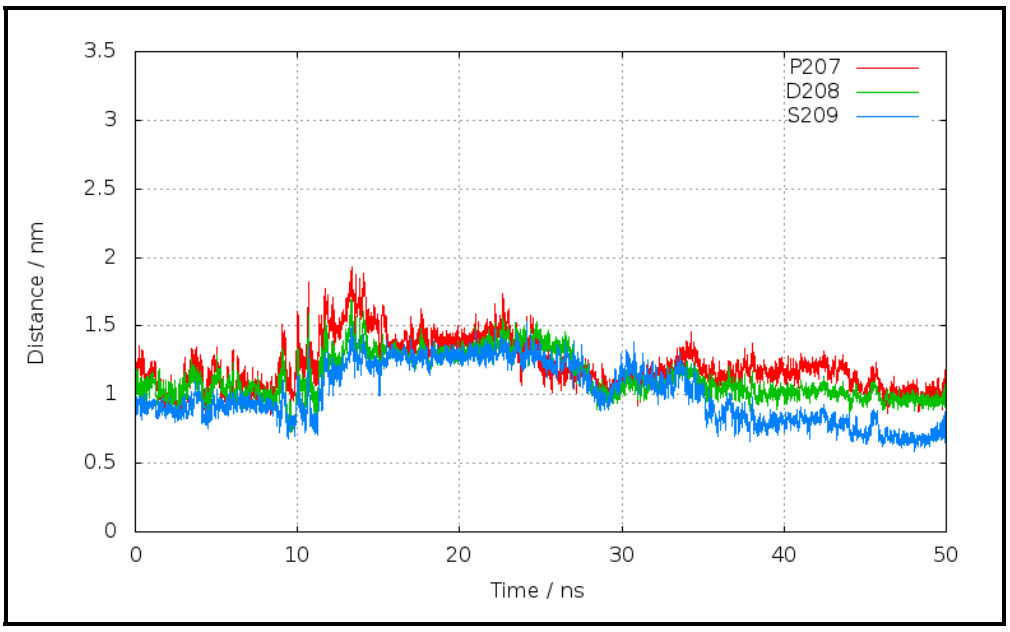

Figure 6.19: *MupH monomer with C115 acetylated simulation replicate 3, distance measured between the loop I and II over the time of 50ns. CYA was the acytylated cysteine. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

The next obvious question was to see if this loop movement can be observed in the non-acetylated wild type MupH monomer, without being docked to ACP or with a ligand in its active site. If the loop movement is triggered by the docking of ACP or the presence of the ligand then there shouldn't be any huge fluctuation in the distances between the loops throughout the simulation. In order to test this hypothesis three independent simulations were run for a non-acetylated wild type MupH monomer for 50 ns each. RMSFs calculated for all atoms averaged per residue of MupH showed larger fluctuation in loop II than loop I, similar to the ACP-mupA3a:MupH monomer simulation (Figure 6.20). Figure 6.22 shows the replicate 2 of the non-acetylated wild type MupH monomer simulation and it can be seen that the distance between the loops could fluctuate beyond 3 nm which was even more than with the bound ligand. Figures 6.21 to 6.23 show the distances measured for all the three replicates. These observations suggest that maybe this loop movement is intrinsic to MupH monomer proteins and it does not require binding of ACP or ligand to trigger it. However, it would also be interesting to see if this inter loop distance is consistent in the wild type ACP free MupH dimer as well. Due to lack of time it was not possible to run the non-acetylated wild type MupH dimer or MupH dimer with C115 acetylated simulations.

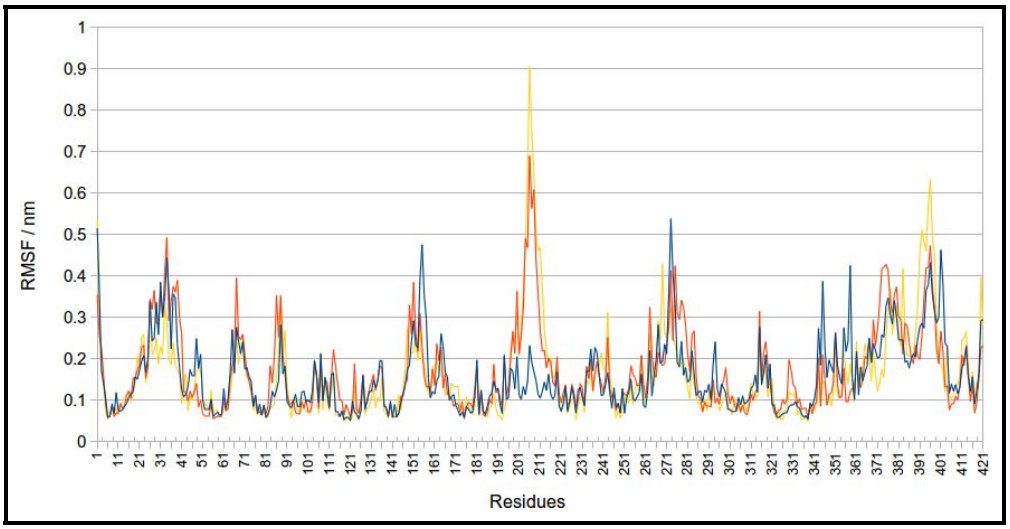

Figure 6.20: *Non-acetylated wild type MupH monomer simulation, RMSF of all atoms averaged per residue of MupH. RMSF values for the residues in loop II (residue 198-214) can be seen larger than the residues on the loop I (residue 147-171). Blue, red and yellow lines represents the three replicates respectively.*

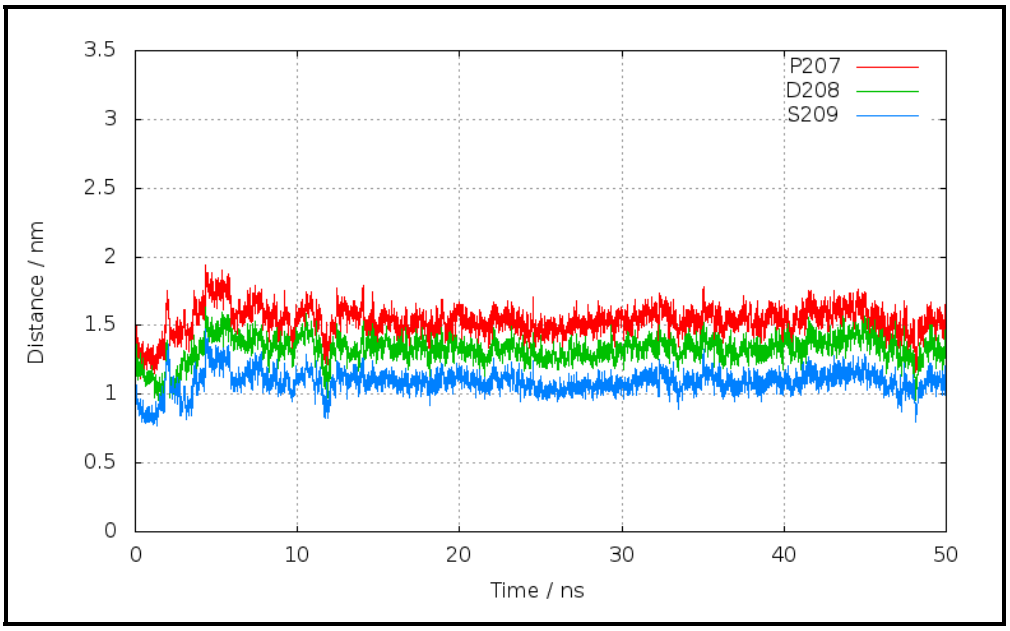

Figure 6.21: *Non-acetylated wild type MupH monomer simulation replicate 1, distance measured between the loop I and II over the time of 50ns. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

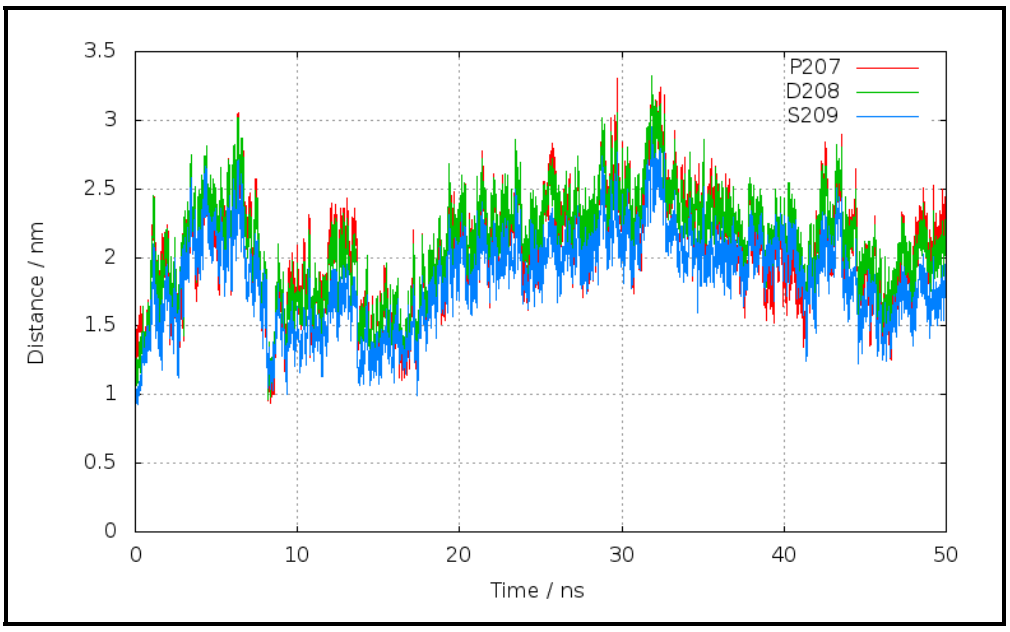

Figure 6.22: *Non-acetylated wild type MupH monomer simulation replicate 2, distance measured between the loop I and II over the time of 50ns. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

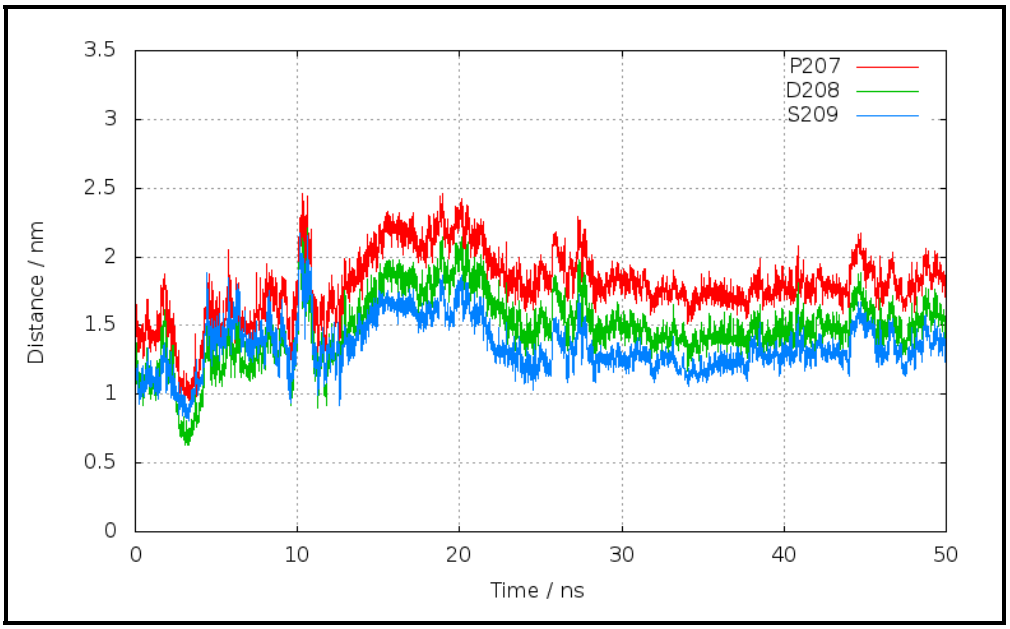

Figure 6.23: *Non-acetylated wild type MupH monomer simulation replicate 3, distance measured between the loop I and II over the time of 50ns. Distances were measured between the three Cα on the loop I (L150, M151 and I152) and three Cα on the loop II (P207, D208 and S209). Red, green and blue lines represents the average distance of P207, D208 and S209 respectively from the three residues on the loop I.*

6.3 Discussion

6.3.1 A loop at the KS dimer interface appears to be responsible for the substrate specificity

To find out what in the mup cluster is responsible for the addition of the 6-hydroxyl (α-hydroxyl), Dr. Joanne Hothersall from Prof. Thomas group carried out mutagenesis to delete *mupA*, a tailoring gene. According to the position of occurrence of the 6-hydroxyl in the mupirocin (Figure 1.26) it was hypothesized that MupA must be acting after MmpD. She found that the deletion strain produced mupiric acid but not mupirocin H (Wu *et al.* 2008). It was inferred that this was formed by the release of an intermediate from module 4 of MmpD. Since no intermediates longer than mupiric acid were found this fits with the idea that MupA is acting after MmpD and if it is acting as a hydrolase, as hypothesised, then it suggests the non-hydroxylated monic acid intermediate cannot be recognised by the first extender domain of MmpA. Supposing that MupA was responsible for the addition of the 6-hydroxyl and that a lack of it would produce an

intermediate without an α-hydroxyl, then what might be stopping KS-mupA2 from recognizing the non α-hydroxylated substrate?

Previous studies on the *cis*-AT DEBS systems have shown that the KSs were able to be acylated by unnatural substrates but were not able to carry out the elongation step, suggesting that the pathway stalled at the KS with the clogged intermediate (Watanabe *et al.* 2003). In a recent study, Busch *et al.* (2013) have shown a gate keeping mechanism in iterative KSs from the *cis*-AT aureothin system. However, studies conducted by Menzella *et al.* (2005) showed PKSs accepting unnatural substrates in a *cis*-AT system. Thus the substrate recognition mechanisms in the *cis*-AT KSs are still not very well understood.

In the *trans*-AT PKSs, Nguyen *et al.* (2008) showed a strong correlation between the KS sequences and their preferred substrates. Based on this association of different substrates with different KS clades Jenner *et al.* (2013) determined the basis of substrate specificity of a *trans*-AT KS associated with accepting β-branched substrates. Using a novel mass spectrometry method they identified a position in the *trans*-AT BaeL KS5 (bacillaene cluster) which is responsible for accepting only β-branched substrates.

To investigate the hypothesis that KS-mupA2 is specific for an α-hydroxylated precursor of monic acid the expected α-hydroxylate substrate was docked into a model of the KS-mupA2 dimer structure. A malonate molecule attached to the phosphopantetheine was also docked, while keeping the substrate C158 interaction, to mimic the decarboxylation stage of the Claisen condensation, with the phosphopantetheine attached to a modelled ACP-mupA2. The final docked conformation mimics the system ready to carry out the carbon-carbon bond formation and the acyl chain transfer to the phosphopantetheinylated ACP. Docking results revealed a motif, DNYK, within 5 Å of the α-hydroxyl, in a loop contributed by the opposite subunit at the dimer interface. This loop connects an α-helix at its N-terminus to a β-strand at its C-terminus. The loop is similar to its counterpart in the thiomarinol TmpA module, whose synthetic pathway also produces a product hydroxylated at the same position, with TmpA and MmpA both having the conserved DNYK motif, but this loop is not conserved among the other KSs from the mupirocin and thiomarinol clusters. It was hypothesized that if this loop were to

be swapped with the loops from KS-mupA1 or KS-mupA3 then the KS which does not process a substrate with an α-hydroxyl might allow the pathway to proceed further.

There seems to be no similar DNYK motif present in the structures available in the PDB, determined by using the PDBe Motif webserver. Limiting the search to a motif in a helical context found only the Vaccinia Virus H7 protein (PDB ID 4W5X) containing this motif, however no ligand information was present. Looking into the structure of Vaccinia Virus H7 protein, the DNYK motif lies in a loop with the D of the motif (D56) at the C-terminus of a helix. Vaccinia Virus H7 protein has a unique fold with no sequence homology outside poxvirus family. The Vaccinia Virus H7 protein shares some similarity in secondary structure and surface properties with the PX domains, which highlights surface residues K108, R109 and K112 forming a basic patch that are found to be important for binding phosphoinositides (Kolli *et al.* 2015). However, the DNYK motif is structurally far away (≈25 Å) from this basic patch and seems to have no role in the binding of phosphoinositides. Moreover, the loop points to solvent with no obvious interaction partners in the crystal lattice and in no way looks like a binding site. Limiting the search finding a DNYK motif in a loop context did not find any matching structure. Searching the motif without limiting the search for any secondary structure information found 31 structures with only a rabbit IGG fc fragment (PDB ID 2VUO) associated with ligand binding information. In the rabbit IGG fc fragment structure the backbone atoms of D389 are in van der Waals contact with an azide ion and the backbone nitrogen of N390 is making a hydrogen bond with the azide ion. The NZ of K392 is in van der Waals contact with the carbon of a formic acid molecule. Although this structure shows the DNYK motif to be in contact with the bound ligands, neither of the ligands carries an α-hydroxyl and the motif is on a β-strand rather than on a loop or helix. Searching for DLYK, DLFK and DLLK variants did not yield any results different results from the DNYK motif search.

To test the importance of this loop in recognizing the substrate, Miss. Y. Alsamrarraie from Prof. Thomas group inserted the key loop from KS-mupA1 into KS-mupA2 in a *P. fluorescens* Δ*mupA* strain and in the wild type *P. fluorescens* NCIMB 10586. HPLC traces revealed a peak for both the mutant strains which was not detected in the parent *P. fluorescens* Δ*mupA* strain.

This implies that the pathway has produced a full length substrate, and thus must have been successfully processed by the KS-mupA2 mutants. Thus, the chimera of KS-mupA2 with KS-mupA1 loop processed the substrates from WT and ΔmupA strains, presumed to be with and without α-hydroxyl. This could mean that the wild type KS-mupA2 serves as a checkpoint, which only allows an α-hydroxylated substrate to pass through and stalls the pathway in case of an α-dehydroxylated substrate. The structure of the metabolites produced are still to be confirmed from our collaborators at Bristol but is likely to be more hydrophobic than pseudomonic acid since it has a longer retention on the column, which is consistent with 6-dehydroxylated pseudomonic acid A. Once it is confirmed that the pathway has produced a molecule which lacks 6-hydroxyl further point mutations can be done to pin point the residue(s) responsible for the α-hydroxyl substrate specificity. Target point mutations may also help not only to modulate the specificity of KS-mupA2 towards accepting an non α-hydroxylated substrate but might also increase the amount of the metabolite produced as compared to swapping the whole loop. These insights would eventually lead towards successfully re-engineering *trans*-AT KSs for accepting non natural substrates for the production of novel compounds, and may also apply to *cis*-AT systems but further experiments are required.

6.3.2 Movement of MupH surface loops may have a role in ligand binding

A large movement in two of the surface loops over the active site of MupH was seen in the simulations of an ACP-mupA3a:MupH complex (Chapter 4, Section 4.2.5) with the ACP-mupA3a cognate substrate attached to the ACP and an acetyl molecule covalently bound to the catalytic C115 in MupH. This observation lead to the hypothesis that this large movement of MupH surface loops at the opening of the MupH active site might be assisting in accommodating the ligand inside the MupH active site. To test the hypothesis that the movement may be triggered either by the acetyl transfer to the catalytic C115 in MupH (see HMG-CoA reaction mechanism in Section 3.1) or due the ACP-mupA3a docking to the MupH, three independent simulations were performed for each of the systems: ACP-mupA3a:MupH monomer, ACP-mupA3a:MupH dimer, MupH monomer with C115 acetylated and wild type MupH monomer structure.

The simulations of the MupH monomer structure showed a general trend of larger movement in loop II, both in terms of its over all fluctuation (RMSF) and the distance between loop II and loop I, however, these fluctuations decreased in the simulations with a dimeric MupH. This could be because loop II, being close to the dimer interface, would interact with the residues from the other monomer, which might restrict its movement. It was also seen that the loops in the MupH monomer with ligand docked in the active site tend to have a large distance between them at the beginning of the simulation which decreases eventually and then remains at a roughly constant value of 10 Å. One interpretation is that the MupH monomer structures may allow easy access to the active site until the substrate is bound. Whether these observations are biologically relevant or not is difficult to say with the available data. Multiple or longer simulation runs with monomeric and dimeric states might provide more evidence.

Determining experimentally that MupH exists as a monomer or dimer may provide some insight into the simulations or may lead to further experiments. In my previous analysis in Chapter 3, Section 3.2.6.2 real value evolutionary trace and PIER analysis showed a strong signal on the MupH surface corresponding to the dimer interface in the HMG-CoA homologue structure, hence suggesting that MupH is also a dimer. To test this experimentally a simple technique such as electrospray ionization mass spectrometry (ESI-MS) may be utilized. Since, ESI-MS is a 'soft ionization'technique which produces very few fragments, it would be straight forward to determine the mass of the whole MupH or MupH dimer (preserving the noncovlent interactions (Huang *et al.* 1993)). ESI-MS can also be utilized to quantify protein-protein association constant (Boeri Erba *et al.* 2011).

Chapter 7

General discussion

7.1 Overview

The aim of the thesis was to explore details of the mupirocin biosynthesis pathway, focussing in particular on aspects of the function of the MmpA complex and its *in trans* interactions. Generalizing some of the principles found should aid in the redesign of existing PKS, for the synthesis of novel compounds, with particular interest in medicinal compounds such as new antibiotics or potential anticancer agents.

In order to investigate the mechanism of MmpA this thesis focused mainly on exploiting already existing molecular modelling methods and proposes some protocols which can be used for similar purposes. Along with utilizing various pre existing molecular modelling methods the projects in this thesis heavily relied on Perl scripts written during the course of the project for various purposes for example HMM analysis and integrating GAFF parameters into GROMACS software. The computational protocols used along with the experimental validation of the predictions underlines the importance of molecular modelling in modern biological research, allowing the prediction of key residues in MmpA and its interacting partner MupH that would have been expensive and very time consuming to discover solely by experiments.

At the inception of this thesis the major outstanding question was the β-branching mechanism in the mupirocin as well as other polyketide biosynthesis pathway. β-branching which is the addition of a methyl branch at the β position was hypothesised to be caused by the con-

certed action of five proteins collectively called the HCS cassette. Out of the five proteins an HMG-CoA homologue (MupH in the mupirocin system) protein is the first to interact with an ACP carrying the substrate to be β-branched. It was not understood what MupH recognizes on the C-terminal ACPs of MmpA. There are 16 ACPs associated with the PKSs and 5 discrete ACPs associated with the tailoring proteins in the mupirocin cluster. Out of the 21 ACPs how does MupH recognizes its partner ACP? According to the position of the β-branch in the monic acid moiety it was easy to deduce that the β-branching happens at the end of the third module of MmpA. However, it does not tell us how does MupH exclude all the 21 possible partner APCs and only choose the two tandem ACPs in the MmpA.

Sequence analysis carried out by Dr. Tony Haines discovered the presence of a conserved tryptophan six residues downstream of the catalytic cysteine in the ACPs associated with the β-branching but never conserved in the non-branching ACPs. Structure determination of the ACP-mupA3ab didomain by Dr. Matt Crump indicated that the tryptophan is buried at the core of the ACPs surrounded by the other highly conserve hydrophobic residues. Cross complementation experiments in Prof. Thomas' group showed that the branching ACPs from the thiomarinol system complements branching ACPs in the mupirocin system but that the non-branching ACPs do not. These experiments established that the branching ACPs differ from the non-branching ACPs and probably the structural variations between the two helps MupH in recognizing the correct ACP. However, a mechanistic view of what is happening at the interface of the two proteins when they interact was missing. Computational studies which included docking, molecular dynamics, hidden Markov model analysis, and sequence and structure based interface prediction carried out in this regard proposed the importance of the tryptophan at the core of the protein in maintaining the correct packaging of the core. This correct packaging would lead to the correct orientation of the helix III which forms the interface of the ACP with MupH. One residue mutation on the helix III, Y62F/A showed the importance of helix III at the interface.

Although the predicted ACP:MupH complex laid down the general rules for the β-branching ACP and HCS interaction there are more residues involved at the interface which define pair

wise specificity between the ACP:HCS pairs in the different systems. Cross complementation experiments showed that BatC (the MupH homologue from kalimantacin system) failed to complement MupH in the mupirocin cluster. A sequence analysis revealed one residue near the ACP:MupH interface, that differs between the TmlH and MupH, which complements Δ*mupH in trans* and BatC that does not. The L to M mutation was found to lead to gain of function with BatC L219M complements Δ*mupH*.

Another way to resume the failed complementation of Δ*mupH* by BatC might be to swap the mupirocin ACPs for the cognate ACPs of BatC from the kalimantacin system. Suicide mutagenesis experiments were conducted to express the β-branching ACPs from the kalimantacin cluster in the mupirocin cluster. No pseudomonic acid production was detected in the HPLC traces for the strains expressing kalimantacin ACPs and the wild type MupH. This observation was in line with the initial hypothesis. However, upon expressing kalimantacin ACPs with the wild type BatC there was also no pseudomonic acid production detected. However, a new peak was detected which may correspond to a new metabolite being produced similar to the pseudomonic acids and experiments are on going to characterise it.

An alternative method for investigating pairwise specificity might be to make changes in the ACP-mupA3ab β-branching ACPs native to the mupirocin cluster, to allow them to function efficiently with the *batC* when expressed *in trans*, as previous experiments have shown the suitability of BatC in the mupirocin cluster. Molecular dynamics simulation of the ACP-mupA3a:MupH complex helped to refine the docking interface proposing residues at the interface which might play role in improving ACP-mupA3a's ability to interact well with the BatC. However, comparing the sequences of ACPs from the clusters which are likely to complement MupH and the ones which do not, there were no obvious patterns that identified putative key residues.

The docking experiments of ACP-mupA3a and MupH have shown that in order to reach the ligand in the correct orientation for the β-branching almost all of the phosphopantetheine along with the 16C monic acid precursor needs to get completely inside the MupH active site. It has always been intriguing on how a large substrate like this is accommodated in the MupH

active site. The docking analysis of rigid static structures does not tell us if there is an involvement of a dynamic element during the process. Molecular dynamics simulations on the ACP-mupA3a:MupH complex revealed a large movement in two loops at the surface of MupH at the opening of the MupH active site. Upon visualising the simulation trajectories it seems that these loop movement assist in the accommodation of the ligand in the active site. The distance between the mobile loops was found to be larger in the MupH monomer structures as compared to the MupH dimer structure. HMG-CoA orthologues exists as a dimer therefore its quite likely that MupH also exists as a dimer however, an experimental validation will be required to support this hypothesis. These large loop movements were not previously reported either through experiments or computational methods.

Dynamic change in the structure upon ligand binding was also identified in the ACPs. FAS ACPs are shown to sequester the acyl chains within their hydrophobic core through experiments as well as molecular dynamics simulations. However, no such observations had been made in the PKS ACPs. In order to explore this phenomenon in the PKS ACPs different apo, holo, and acyl forms of ACPs from the mupirocin cluster were simulated in explicit solvent for upto 1 μs. The molecular dynamics simulations revealed that the PKS ACPs do form a cavity upon the attachment of the phosphopantetheine and acyl chains. The length of the acyl chain might also influence the size of the cavity. The cavity formed does not form a deep tunnel as in the FAS ACPs but, is rather a shallow, solvent exposed, surface groove, enabling the polar groups on the acyl chain to hydrogen bond with the solvent whilst shielding the hydrophobic parts of the polyketide. It was also seen that a bulky residue at the proposed tunnel opening in the PKS ACPs prohibits the formation of a deep tunnel as opposed to the smaller residue at the equivalent position in the FAS ACPs. These dynamic behaviours were seen for the first time in the PKS ACPs in this work.

In re-engineering polyketides there are two major problems, the recognition specificity between the protein domains and the recognition specificity of the substrate. In the present work I worked on analysing the KS substrate specificity for an α-hydroxylated polyketide substrate by molecular docking of the cognate substrate of KS-mupA2 with the KS-mupA2 homo dimer.

Dr. Joanne Hothersall knocked out MupA, which is thought to hydroxylate pseudomonic acid at the 6-OH position. For Δ*mupA* the longest intermediate found was mupric acid which is produced by MmpD. This observation led to the hypothesis that an un α-hydroxylated substrate cannot be recognized by the KS-mupA2, the first condensing domain in the MmpA, resulting in the pathway shutdown. Docking the expected α-hydroxylated cognate substrate to the KS-mupA2 homo dimer revealed a loop in the opposite monomer at the dimer interface interacting with the α-OH. A sequence alignment of this loop from KS-mupA2 with all the other KSs from the mupirocin and thiomarinol cluster showed a similar loop in the equivalent KS from the thiomarinol system but no conservation was found with the other KSs. With this observation it was hypothesised that replacing this loop in KS-mupA2 with the equivalent loops from the KSs preceding and following it, KS-mupA1 and KS-mupA3, for which the cognate substrates does not have an α-OH, might allow the pathway to proceed further. By the time of writing this thesis Miss Yousra Alsamarraie from Prof. Thomas group was able to replace the KS-mupA2 loop with KS-mupA1 loop. The HPLC trace showed no peak for pseudomonic acid A or B but a new peak with larger retention time suggested that the pathway has produced a full length substrate, that is slightly more hydrophobic than pseudomonic acid A, consistent with dehydroxylated form of the product. More detailed structural analysis of the product is awaited, but the experiments do seem to show that these loops are important for specificity.

7.2 Conclusions

The body of the work here shows the effectiveness of a range of modelling and bioinformatics techniques for the analysis of the protein-protein, domain-domain and protein-substrate interaction involved in polyketide synthesis. Combined with experiments performed by me and others elucidated key details of control of substrate flow at the start of the MmpA subunit from mupirocin system and of the β-methylation step at the end of the MmpA subunit. The methods used here and the results obtained should be applicable to other systems. In the future this may lead to experimental and computational tools to design and synthesis novel compounds de novo.

List of References

Akey, David L., Jamie R. Razelun, Jason Tehranisa, David H. Sherman, William H. Gerwick, and Janet L. Smith (2010). "Crystal Structures of Dehydratase Domains from the Curacin Polyketide Biosynthetic Pathway". *Structure* 18.1, pp. 94–105 (cit. on p. 41).

Alanis, Alfonso J (2005). "Resistance to antibiotics: are we in the post-antibiotic era?" *Archives of medical research* 36.6, pp. 697–705 (cit. on pp. 1, 2).

Alekseyev, Viktor Y, Corey W Liu, David E Cane, Joseph D Puglisi, and Chaitan Khosla (2007). "Solution structure and proposed domain domain recognition interface of an acyl carrier protein domain from a modular polyketide synthase." *Protein science : a publication of the Protein Society* 16.10, pp. 2093–107 (cit. on p. 26).

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman (1990). "Basic local alignment search tool." *Journal of molecular biology* 215.3, pp. 403–10 (cit. on p. 303).

Ames, B. D., C. Nguyen, J. Bruegger, P. Smith, W. Xu, S. Ma, E. Wong, S. Wong, X. Xie, J. W.-H. Li, J. C. Vederas, Y. Tang, and S.-C. Tsai (2012). "Crystal structure and biochemical studies of the trans-acting polyketide enoyl reductase LovC from lovastatin biosynthesis". *Proceedings of the National Academy of Sciences* 109, pp. 11144–11149 (cit. on pp. 18, 42–44).

Anand, Swadha, M V R Prasad, Gitanjali Yadav, Narendra Kumar, Jyoti Shehara, Md Zeeshan Ansari, and Debasisa Mohanty (2010). "SBSPKS: structure based sequence analysis of polyketide synthases." *Nucleic acids research* 38.Web Server issue, W487–96 (cit. on pp. 66, 306).

Ansari, Mohd Zeeshan, Jyoti Sharma, Rajesh S Gokhale, and Debasisa Mohanty (2008). "In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites." *BMC bioinformatics* 9, p. 454 (cit. on pp. 303, 308).

Aparicio, Jesús F., Patrick Caffrey, Andrew F a Marsden, James Staunton, and Peter F. Leadlay (1994). "Limited proteolysis and active-site studies of the first multienzyme component of the erythromycin-producing polyketide synthase." *Journal of Biological Chemistry* 269.11, pp. 8524–8528 (cit. on pp. 12, 13).

Arthur, Christopher J, Anna Szafranska, Simon E Evans, Stuart C Findlow, Steven G Burston, Philip Owen, Ian Clark-Lewis, Thomas J Simpson, John Crosby, and Matthew P Crump (2005). "Self-malonylation is an intrinsic property of a chemically synthesized type II polyketide synthase acyl carrier protein." *Biochemistry* 44.46, pp. 15414–21 (cit. on p. 21).

Austin, Michael B and Joseph P Noel (2003). "The chalcone synthase superfamily of type III polyketide synthases." *Natural product reports* 20.1, pp. 79–110 (cit. on pp. 22, 30).

Bachmann, Brian O and Jacques Ravel (2009). "Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data." *Methods in enzymology*. Vol. 458, pp. 181–217 (cit. on p. 67).

Bahnson, Brian J (2004). "An atomic-resolution mechanism of 3-hydroxy-3-methylglutaryl-CoA synthase." *Proceedings of the National Academy of Sciences of the United States of America* 101.47, pp. 16399–400 (cit. on p. 124).

Baker, N a, D Sept, S Joseph, M J Holst, and J a McCammon (2001). "Electrostatics of nanosystems: application to microtubules and the ribosome." *Proceedings of the National Academy of Sciences of the United States of America* 98.18, pp. 10037–10041 (cit. on p. 145).

Bayly, Christopher I. Ci Christopher I, Piotr Cieplak, Wendy D Cornell, and Peter a Kollman (1993). "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model". *The Journal of Physical ...* 97.40, pp. 10269–10280 (cit. on p. 81).

Bender, Cl, V Rangaswamy, and J Loper (1999). "POLYKETIDE PRODUCTION BY PLANT-ASSOCIATED PSEUDOMONADS." *Annual review of phytopathology* 37, pp. 175–196 (cit. on p. 3).

Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak (1984). "Molecular dynamics with coupling to an external bath". *The Journal of Chemical Physics* 81.8, pp. 3684–3690 (cit. on p. 83).

Birch, AJ, RA Massy-Westropp, and CJ Moye (1955). "Studies in relation to biosynthesis. VII. 2-Hydroxy-6-methylbenzoic acid in Penicillium griseofulvum Dierckx". *Australian Journal of Chemistry* 8.4, p. 539 (cit. on p. 5).

Bisang, C, P F Long, J Cortés, J Westcott, J Crosby, A L Matharu, R J Cox, T J Simpson, J Staunton, and P F Leadlay (1999). "A chain initiation factor common to both modular and aromatic polyketide synthases." *Nature* 401.6752, pp. 502–5 (cit. on p. 21).

Boehringer, Daniel, Nenad Ban, and Marc Leibundgut (2013). "7.5-ÅCryo-Em Structure of the Mycobacterial Fatty Acid Synthase." *Journal of molecular biology* 425.5, pp. 841–9 (cit. on pp. 46, 47, 53).

Boeri Erba, Elisabetta, Konstantin Barylyuk, Yang Yang, and Renato Zenobi (2011). "Quantifying protein-protein interactions within noncovalent complexes using electrospray ionization mass spectrometry." *Analytical chemistry* 83.24, pp. 9251–9 (cit. on p. 224).

Brink, Jacob, Steven J Ludtke, Chao-Yuh Yang, Zei-Wei Gu, Salih J Wakil, and Wah Chiu (2002). "Quaternary structure of human fatty acid synthase by electron cryomicroscopy." *Proceedings of the National Academy of Sciences of the United States of America* 99.1, pp. 138–43 (cit. on p. 46).

Broadhurst, R William, Daniel Nietlispach, Michael P Wheatcroft, Peter F Leadlay, and Kira J Weissman (2003). "The structure of docking domains in modular polyketide synthases." *Chemistry & biology* 10.8, pp. 723–31 (cit. on p. 307).

Busch, Benjamin, Nico Ueberschaar, Swantje Behnken, Yuki Sugimoto, Martina Werneburg, Nelly Traitcheva, Jing He, and Christian Hertweck (2013). "Multifactorial control of iteration events in a modular polyketide assembly line." *Angewandte Chemie (International ed. in English)* 52.20, pp. 5285–9 (cit. on p. 221).

Busche, Alena, Daniel Gottstein, Christopher Hein, Nina Ripin, Irina Pader, Peter Tufar, Eli B Eisman, Liangcai Gu, Christopher T Walsh, David H Sherman, Frank Löhr, Peter Güntert, and Volker Dötsch (2012). "Characterization of molecular interactions between ACP and halogenase domains in the Curacin A polyketide synthase." *ACS chemical biology* 7.2, pp. 378–86 (cit. on pp. 19, 25, 26, 110, 152).

Byers, David M and Huansheng Gong (2007). "Acyl carrier protein: structure-function relationships in a conserved multifunctional protein family." *Biochemistry and cell biology = Biochimie et biologie cellulaire* 85.6, pp. 649–62 (cit. on p. 24).

Caboche, Ségolène, Maude Pupin, Valérie Leclère, Arnaud Fontaine, Philippe Jacques, and Gregory Kucherov (2008). "NORINE: a database of nonribosomal peptides." *Nucleic acids research* 36.Database issue, pp. D326–31 (cit. on p. 302).

Caffrey, Patrick (2003). "Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases." *Chembiochem : a European journal of chemical biology* 4.7, pp. 654–7 (cit. on p. 39).

Campbell, Chantel D. and John C. Vederas (2010). "Biosynthesis of lovastatin and related metabolites formed by fungal iterative PKS enzymes." *Biopolymers* 93.9, pp. 755–63 (cit. on p. 18).

Canutescu, Adrian A, Andrew A Shelenkov, and Roland L Dunbrack (2003). "A graph-theory algorithm for rapid protein side-chain prediction." *Protein science : a publication of the Protein Society* 12.9, pp. 2001–14 (cit. on p. 306).

Challis, G L, J Ravel, and C a Townsend (2000). "Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains." *Chemistry & biology* 7.3, pp. 211–24 (cit. on p. 305).

Challis, Gregory L. (2008). "Mining microbial genomes for new natural products and biosynthetic pathways." *Microbiology (Reading, England)* 154.Pt 6, pp. 1555–69 (cit. on p. 60).

Chan, David I, Thomas Stockner, D Peter Tieleman, and Hans J Vogel (2008). "Molecular dynamics simulations of the Apo-, Holo-, and acyl-forms of Escherichia coli acyl carrier protein." *The Journal of biological chemistry* 283.48, pp. 33620–9 (cit. on pp. 24, 70, 178, 179, 184, 196–199).

Chothia, C and Arthur M Lesk (1986). "The relation between the divergence of sequence and structure in proteins." *The EMBO journal* 5.4, pp. 823–826 (cit. on p. 75).

Cohen, S N, A C Chang, H W Boyer, and R B Helling (1973). "Construction of biologically functional bacterial plasmids in vitro." *Proceedings of the National Academy of Sciences of the United States of America* 70.11, pp. 3240–4 (cit. on p. 98).

Collie, John Norman (1907). "CLXXI.?Derivatives of the multiple keten group". *Journal of the Chemical Society, Transactions* 91, p. 1806 (cit. on p. 5).

Conly, J. M. Jm and Bl L. Johnston (2005). "Where are all the new antibiotics? The new antibiotic paradox." *The Canadian journal of infectious diseases & medical microbiology = Journal canadien des maladies infectieuses et de la microbiologie médicale / AMMI Canada* 16.3, pp. 159–60 (cit. on p. 2).

Cortes, J, S F Haydock, G A Roberts, D J Bevitt, and P F Leadlay (1990). "An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of Saccharopolyspora erythraea." *Nature* 348.6297, pp. 176–178 (cit. on p. 6).

Cortes, J, K E Wiesmann, G A Roberts, M J Brown, J Staunton, and P F Leadlay (1995). "Repositioning of a domain in a modular polyketide synthase to promote specific chain cleavage." *Science (New York, N.Y.)* 268.5216, pp. 1487–9 (cit. on p. 37).

Dahiyat, B I, C a Sarisky, and S L Mayo (1997). "De novo protein design: towards fully automated sequence selection." *Journal of molecular biology* 273.4, pp. 789–796 (cit. on p. 152).

Das, Abhirup and Chaitan Khosla (2009). "Biosynthesis of aromatic polyketides in bacteria." *Accounts of chemical research* 42.5, pp. 631–9 (cit. on pp. 20, 22).

Davies, C, R J Heath, S W White, and C O Rock (2000). "The 1.8 A crystal structure and active-site architecture of beta-ketoacyl-acyl carrier protein synthase III (FabH) from escherichia coli." *Structure (London, England : 1993)* 8.2, pp. 185–95 (cit. on pp. 31, 32).

Davison, Jack, Jonathan Dorival, Hery Rabeharindranto, Hortense Mazon, Benjamin Chagot, Arnaud Gruez, and Kira J Weissman (2014). "Insights into the function of trans-acyl transferase polyketide synthases from the SAXS structure of a complete module". *Chemical Science*, pp. 3081–3095 (cit. on p. 26).

Dolinsky, Todd J, Jens E Nielsen, J Andrew McCammon, and Nathan a Baker (2004). "PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations." *Nucleic acids research* 32.Web Server issue, W665–7 (cit. on p. 145).

Dominguez, Cyril, Rolf Boelens, and Alexandre M J J Bonvin (2003). "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information." *Journal of the American Chemical Society* 125.7, pp. 1731–7 (cit. on p. 94).

Dreier, J, A N Shah, and C Khosla (1999). "Kinetic analysis of the actinorhodin aromatic polyketide synthase." *The Journal of biological chemistry* 274.35, pp. 25108–12 (cit. on p. 21).

Du, P and I Alkorta (1994). "Sequence divergence analysis for the prediction of seven-helix membrane protein structures: I. Comparison with bacteriorhodopsin." *Protein engineering* 7.10, pp. 1221–9 (cit. on p. 90).

Dunn, Briana J, David E Cane, and Chaitan Khosla (2013). "Mechanism and Specificity of an Acyltransferase Domain from a Modular Polyketide Synthase." *Biochemistry* 52.11, pp. 1839–41 (cit. on p. 27).

Dupradeau, François-Yves, Adrien Pigache, Thomas Zaffran, Corentin Savineau, Rodolphe Lelong, Nicolas Grivel, Dimitri Lelong, Wilfried Rosanski, and Piotr Cieplak (2010). "The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building." *Physical chemistry chemical physics : PCCP* 12.28, pp. 7821–7839 (cit. on p. 81).

Dutta, Somnath, Jonathan R Whicher, Douglas A Hansen, Wendi A Hale, Joseph A Chemler, Grady R Congdon, Alison R H Narayan, Kristina Hå kansson, David H Sherman, Janet L Smith, and Georgios Skiniotis (2014). "Structure of a modular polyketide synthase." *Nature* 510.7506, pp. 512–7 (cit. on pp. 57–59).

El-Sayed, A K, J Hothersall, and C M Thomas (2001). "Quorum-sensing-dependent regulation of biosynthesis of the polyketide antibiotic mupirocin in Pseudomonas fluorescens NCIMB 10586." *Microbiology (Reading, England)* 147.Pt 8, pp. 2127–39 (cit. on p. 97).

El-Sayed, A Kassem, Joanne Hothersall, Sian M Cooper, Elton Stephens, Thomas J Simpson, and Christopher M Thomas (2003). "Characterization of the mupirocin biosynthesis gene cluster from Pseudomonas fluorescens NCIMB 10586." *Chemistry & biology* 10.5, pp. 419–30 (cit. on p. 201).

Eswar, Narayanan, Ben Webb, Marc A Marti-Renom, M S Madhusudhan, David Eramian, Min-Yi Shen, Ursula Pieper, and Andrej Sali (2006). "Comparative protein structure modeling using Modeller." *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 5, Unit 5.6 (cit. on pp. 75, 76).

Ferrer, J L, J M Jez, M E Bowman, R a Dixon, and J P Noel (1999). "Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis." *Nature structural biology* 6.8, pp. 775–784 (cit. on pp. 22, 23).

Finn, Robert D, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, and Alex Bateman (2010). "The Pfam protein families database." *Nucleic acids research* 38.Database issue, pp. D211–22 (cit. on p. 308).

Foerstner, Konrad U, Tobias Doerks, Christopher J Creevey, Anja Doerks, and Peer Bork (2008). "A computational screen for type I polyketide synthases in metagenomics shotgun data." *PloS one* 3.10, e3515 (cit. on p. 303).

Fuller, A T, G Mellows, M Woolford, G T Banks, K D Barrow, and E B Chain (1971). "Pseudomonic acid: an antibiotic produced by Pseudomonas fluorescens." *Nature* 234.5329, pp. 416–7 (cit. on p. 62).

Galtier, N, M Gouy, and C Gautier (1996). "SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny." *Computer applications in the biosciences : CABIOS* 12.6, pp. 543–548 (cit. on p. 76).

Gay, Darren C., Glen Gay, Abram J. Axelrod, Matthew Jenner, Christoph Kohlhaas, Annette Kampa, Neil J. Oldham, Jörn Piel, and Adrian T. Keatinge-Clay (2014). "A close look at a ketosynthase from a trans-acyltransferase modular polyketide synthase." *Structure (London, England : 1993)* 22.3, pp. 444–51 (cit. on pp. 27, 50).

Gokhale, Rajesh S., Janice Lau, David E. Cane, and Chaitan Khosla (1998). "Functional orientation of the acyltransferase domain in a module of the erythromycin polyketide synthase". *Biochemistry* 37.97, pp. 2524–2528 (cit. on p. 14).

Gokhale, Rajesh S., Daniel Hunziker, David E. Cane, and Chaitan Khosla (1999). "Mechanism and specificity of the terminal thioesterase domain from the erythromycin polyketide synthase". *Chemistry and Biology* 6, pp. 117–125 (cit. on p. 44).

Gokhale, Rajesh S, Rajan Sankaranarayanan, and Debasisa Mohanty (2007). "Versatility of polyketide synthases in generating metabolic diversity." *Current opinion in structural biology* 17.6, pp. 736–43 (cit. on p. 306).

Gurney, Rachel and Christopher M Thomas (2011). "Mupirocin: biosynthesis, special features and applications of an antibiotic from a gram-negative bacterium." *Applied microbiology and biotechnology* 90.1, pp. 11–21 (cit. on pp. 63, 307).

Ha, Jun Yong, Ji Young Min, Su Kyung Lee, Hyoun Sook Kim, Do Jin Kim, Kyoung Hoon Kim, Hyung Ho Lee, Hye Kyung Kim, Hye-Jin Yoon, and Se Won Suh (2006). "Crystal structure of 2-nitropropane dioxygenase complexed with FMN and substrate. Identification of the catalytic base." *The Journal of biological chemistry* 281.27, pp. 18660–7 (cit. on p. 42).

Haft, D H, B J Loftus, D L Richardson, F Yang, J A Eisen, I T Paulsen, and O White (2001). "TIGRFAMs: a protein family resource for the functional identification of proteins." *Nucleic acids research* 29.1, pp. 41–3 (cit. on p. 304).

Haines, Anthony S, Xu Dong, Zhongshu Song, Rohit Farmer, Christopher Williams, Joanne Hothersall, Eliza PÅoskoÅ, Pakorn Wattana-Amorn, Elton R. Stephens, Erika Yamada, Rachel Gurney, Yuiko Takebayashi, Joleen Masschelein, Russell J. Cox, Rob Lavigne, Christine L. Willis, Thomas J. Simpson, John Crosby, Peter J. Winn, Christopher M. Thomas, and Matthew P. Crump (2013). "A conserved motif flags acyl carrier proteins for β-branching in polyketide synthesis." *Nature chemical biology* september (cit. on pp. 18, 25, 26, 70, 97, 113–115, 123, 144, 146).

Haydock, S F, J F Aparicio, I Molnár, T Schwecke, L E Khaw, A König, A F Marsden, I S Galloway, J Staunton, and P F Leadlay (1995). "Divergent sequence motifs correlated with the substrate specificity of (methyl)malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases." *FEBS letters* 374.2, pp. 246–8 (cit. on pp. 29, 30).

Heath, Richard J. and Charles O. Rock (2002). "The Claisen condensation in biology." *Natural product reports* 19.5, pp. 581–96 (cit. on pp. 31, 32).

Hertweck, Christian (2009). "The biosynthetic logic of polyketide diversity." *Angewandte Chemie (International ed. in English)* 48.26, pp. 4688–716 (cit. on p. 14).

Hertweck, Christian, Andriy Luzhetskyy, Yuri Rebets, and Andreas Bechthold (2007). "Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork." *Natural product reports* 24.1, pp. 162–90 (cit. on p. 19).

Hess, Berk, Se Uppsala, Erik Lindahl, Carsten Kutzner, and David van der Spoel (2008). "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation". *Journal of Chemical Theory and Computation* 4.3, pp. 435–447 (cit. on pp. 80, 82).

Hornak, Viktor, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling (2006). "Comparison of multiple Amber force fields and development of improved protein backbone parameters." *Proteins* 65.3, pp. 712–725 (cit. on p. 76).

Huang, Bingding (2009). "MetaPocket: a meta approach to improve protein ligand binding site prediction." *Omics : a journal of integrative biology* 13.4, pp. 325–30 (cit. on p. 87).

Huang, E C, B N Pramanik, A Tsarbopoulos, P Reichert, A K Ganguly, P P Trotta, T L Nagabhushan, and T R Covey (1993). "Application of electrospray mass spectrometry in probing protein-protein and protein-ligand noncovalent interactions." *Journal of the American Society for Mass Spectrometry* 4.8, pp. 624–30 (cit. on p. 224).

Hughes, J and G Mellows (1978). "Inhibition of isoleucyl-transfer ribonucleic acid synthetase in Escherichia coli by pseudomonic acid." *The Biochemical journal* 176.1, pp. 305–18 (cit. on p. 62).

Humphrey, W, A Dalke, and K Schulten (1996). "VMD: visual molecular dynamics." *Journal of molecular graphics* 14.1, pp. 33–8, 27–8 (cit. on p. 87).

Hunter, Sarah, Philip Jones, Alex Mitchell, Rolf Apweiler, Teresa K Attwood, Alex Bateman, Thomas Bernard, David Binns, Peer Bork, Sarah Burge, Edouard de Castro, Penny Coggill, Matthew Corbett, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, Robert D Finn, Matthew Fraser, Julian Gough, Daniel Haft, Nicolas Hulo, Daniel Kahn, Elizabeth Kelly, Ivica Letunic, David Lonsdale, Rodrigo Lopez, Martin Madera, John Maslen, Craig Mcanulla, Jennifer Mcdowall, Conor Mcmenamin, Huaiyu Mi, Prudence Mutowo-Muellenet, Nicola Mulder, Darren Natale, Christine Orengo, Sebastien Pesseat, Marco Punta, Antony F Quinn, Catherine Rivoire, Amaia Sangrador-vegas, Jeremy D Selengut, Christian J a Sigrist, Maxim Scheremetjew, John Tate, Manjulapramila Thimmajanarthanan, Paul D Thomas, Cathy H Wu, Corin Yeats, Siew-Yit Yong, Edouard De Castro, De Lyon, Penny Coggill, Matthew Corbett, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, Robert D Finn, Matthew Fraser, Julian Gough, Daniel Haft, Nicolas Hulo, Daniel Kahn, Elizabeth Kelly, Ivica Letunic, David Lonsdale, Rodrigo Lopez, Martin Madera, John Maslen, Craig Mcanulla, Jennifer Mcdowall, Conor Mcmenamin, Huaiyu Mi, Prudence Mutowo-Muellenet, Nicola Mulder, Darren Natale, Christine Orengo, Sebastien Pesseat, Marco Punta, Antony F Quinn, Catherine Rivoire, Amaia Sangrador-vegas, Jeremy D Selengut, Christian J a Sigrist, Maxim Scheremetjew, John Tate, Manjulapramila Thimmajanarthanan, Paul D Thomas, Cathy H Wu, and De Lyon (2012). "InterPro in 2011: new developments in the family and domain prediction database." *Nucleic acids research* 40.Database issue, pp. D306–12 (cit. on p. 302).

Hurley, James H, Walter a Baase, and Brian W Matthews (1992). "Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme". *Journal of Molecular Biology* 224.4, pp. 1143–1159 (cit. on p. 152).

Jenke-Kodama, Holger and Elke Dittmann (2009). "Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges." *Natural product reports* 26.7, pp. 874–883 (cit. on p. 67).

Jenner, Matthew, Sarah Frank, Annette Kampa, Christoph Kohlhaas, Petra Pçplau, Geoff S Briggs, Jçrn Jörn Piel, Neil J Oldham, and Petra Pöplau (2013). "Substrate Specificity in

Ketosynthase Domains from trans-AT Polyketide Synthases." *Angewandte Chemie (International ed. in English)* 52.4, pp. 1143–7 (cit. on pp. 35, 221).

Jenni, Simon, Marc Leibundgut, Daniel Boehringer, Christian Frick, Bohdan Mikolásek, and Nenad Ban (2007). "Structure of fungal fatty acid synthase and implications for iterative substrate shuttling." *Science (New York, N.Y.)* 316.5822, pp. 254–61 (cit. on pp. 46, 50, 51).

Jong, Anne de, Auke J. van Heel, Jan Kok, and Oscar P. Kuipers (2010). "BAGEL2: mining for bacteriocins in genomic data." *Nucleic acids research* 38.Web Server issue, W647–51 (cit. on p. 308).

Joosten, Robbie P, Tim A H Beek, Elmar Krieger, Maarten L Hekkelman, Rob W W Hooft, Reinhard Schneider, Chris Sander, Gert Vriend, and Tim a H te Beek (2011). "A series of PDB related databases for everyday needs." *Nucleic acids research* 39.Database issue, pp. 411–419 (cit. on p. 76).

Jorgensen, William L., David S. Maxwell, and Julian Tirado-Rives (1996). "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids". *Journal of the American Chemical Society* 118, pp. 11225–11236 (cit. on p. 79).

Joshi, A K, A Witkowski, and S Smith (1998). "The malonyl/acetyltransferase and beta-ketoacyl synthase domains of the animal fatty acid synthase can cooperate with the acyl carrier protein domain of either subunit." *Biochemistry* 37.8, pp. 2515–23 (cit. on p. 11).

Joshi, Anil K., Andrzej Witkowski, and Stuart Smith (1997). "Mapping of functional interactions between domains of the animal fatty acid synthase by mutant complementation in vitro." *Biochemistry* 36.8, pp. 2316–22 (cit. on p. 11).

Joshi, Anil K, Vangipuram S Rangan, Andrzej Witkowski, and Stuart Smith (2003). "Engineering of an active animal fatty acid synthase dimer with only one competent subunit." *Chemistry & biology* 10.2, pp. 169–73 (cit. on p. 11).

Kakavas, S J, L Katz, and D Stassi (1997). "Identification and characterization of the niddamycin polyketide synthase genes from Streptomyces caelestis." *Journal of bacteriology* 179.23, pp. 7515–22 (cit. on p. 31).

Kao, Camilla M, Rembert Pieper, David E Cane, and Chaitan Khosla (1996). "Evidence for two catalytically independent clusters of active sites in a functional modular polyketide synthase." *Biochemistry* 35.38, pp. 12363–8 (cit. on p. 13).

Kao, Camilla M., Michael McPherson, Robert N. McDaniel, Hong Fu, David E. Cane, and Chaitan Khosla (1998). "Alcohol Stereochemistry in Polyketide Backbones Is Controlled by the β-Ketoreductase Domains of Modular Polyketide Synthases". *Journal of the American Chemical Society* 120.10, pp. 2478–2479 (cit. on p. 37).

Kapur, Shiven, Alice Y Chen, David E Cane, and Chaitan Khosla (2010). "Molecular recognition between ketosynthase and acyl carrier protein domains of the 6-deoxyerythronolide B synthase." *Proceedings of the National Academy of Sciences of the United States of America* 107.51, pp. 22066–71 (cit. on pp. 27, 50, 60).

Kapur, Shiven, Brian Lowry, Satoshi Yuzawa, Sanketha Kenthirapalan, Alice Y. Chen, David E. Cane, and Chaitan Khosla (2012). "Reprogramming a module of the 6-deoxyerythronolide B synthase for iterative chain elongation". *Proceedings of the National Academy of Sciences of the United States of America* 109.11, pp. 4110–4115 (cit. on pp. 2, 27, 60).

Karaca, Ezgi, Adrien S J Melquiond, Sjoerd J De Vries, Panagiotis L Kastritis, Alexandre M J J Bonvin, and Sjoerd J de Vries (2010). "Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server." *Molecular & cellular proteomics : MCP* 9.8, pp. 1784–94 (cit. on p. 94).

Keatinge-clay, Adrian (2008). "Crystal structure of the erythromycin polyketide synthase dehydratase." *Journal of molecular biology* 384.4, pp. 941–53 (cit. on p. 306).

Keatinge-Clay, Adrian T (2007). "A tylosin ketoreductase reveals how chirality is determined in polyketides." *Chemistry & biology* 14.8, pp. 898–908 (cit. on p. 39).

Keatinge-Clay, Adrian T and Robert M Stroud (2006). "The structure of a ketoreductase determines the organization of the beta-carbon processing enzymes of modular polyketide synthases." *Structure (London, England : 1993)* 14.4, pp. 737–48 (cit. on pp. 38, 39, 41, 306).

Keatinge-Clay, Adrian T, Anang A Shelat, David F Savage, Shiou Chuan Tsai, Larry J W Miercke, Joseph D O'Connell, Chaitan Khosla, and Robert M Stroud (2003). "Catalysis, specificity, and ACP docking site of Streptomyces coelicolor malonyl-CoA:ACP transacylase." *Structure (London, England : 1993)* 11.2, pp. 147–54 (cit. on pp. 21, 27).

Keatinge-Clay, Adrian T, David a Maltby, Katalin F Medzihradszky, Chaitan Khosla, and Robert M Stroud (2004). "An antibiotic factory caught in action." *Nature structural & molecular biology* 11.9, pp. 888–93 (cit. on p. 21).

Kennedy, J., K Auclair, S G Kendrew, C Park, J C Vederas, and C R Hutchinson (1999). "Modulation of Polyketide Synthase Activity by Accessory Proteins During Lovastatin Biosynthesis". *Science* 284.5418, pp. 1368–1372 (cit. on p. 18).

Khaldi, Nora, Fayaz T Seifuddin, Geoff Turner, Daniel Haft, William C Nierman, Kenneth H Wolfe, and Natalie D Fedorova (2010). "SMURF: Genomic mapping of fungal secondary metabolite clusters." *Fungal genetics and biology : FG & B* 47.9, pp. 736–41 (cit. on p. 304).

Khosla, Chaitan (2009). "Structures and mechanisms of polyketide synthases." *The Journal of organic chemistry* 74.17, pp. 6416–20 (cit. on p. 306).

Khosla, Chaitan, Rajesh S Gokhale, John R Jacobsen, and David E Cane (1999). "Tolerance and specificity of polyketide synthases." *Annual review of biochemistry* 68, pp. 219–53 (cit. on pp. 7, 26, 32, 37, 44).

Khosla, Chaitan, Yinyan Tang, Alice Y Chen, Nathan A Schnarr, and David E Cane (2007). "Structure and mechanism of the 6-deoxyerythronolide B synthase." *Annual review of biochemistry* 76, pp. 195–221 (cit. on pp. 15, 16, 44, 60, 306).

Kitamoto, T, M Nishigai, T Sasaki, and A Ikai (1988). "Structure of fatty acid synthetase from the Harderian gland of guinea pig. Proteolytic dissection and electron microscopic studies." *Journal of molecular biology* 203.1, pp. 183–95 (cit. on p. 46).

Koglin, Alexander, Mohammad R Mofid, Frank Löhr, Birgit Schäfer, Vladimir V Rogov, Marc-Michael Blum, Tanja Mittag, Mohamed a Marahiel, Frank Bernhard, and Volker Dötsch (2006). "Conformational switches modulate protein interactions in peptide antibiotic synthetases." *Science (New York, N.Y.)* 312.5771, pp. 273–6 (cit. on p. 24).

Kohlhaas, Christoph, Matthew Jenner, Annette Kampa, Geoff S. Briggs, José P. Afonso, Jörn Piel, and Neil J. Oldham (2013). "Amino acid-accepting ketosynthase domain from a trans-AT polyketide synthase exhibits high selectivity for predicted intermediate". *Chemical Science* 4.8, p. 3212 (cit. on p. 36).

Kolli, Swapna, Xiangzhi Meng, Xiang Wu, Djoshkun Shengjuler, Craig E. Cameron, Yan Xiang, and Junpeng Deng (2015). "Structure-Function Analysis of Vaccinia Virus H7 Protein Reveals a Novel Phosphoinositide Binding Fold Essential for Poxvirus Replication". *Journal of Virology* 89.4, pp. 2209–2219 (cit. on p. 222).

Krivov, Georgii G, Maxim V Shapovalov, Roland L Dunbrack, and Roland L Dunbrack Jr (2009). "Improved prediction of protein side-chain conformations with SCWRL4." *Proteins* 77.4, pp. 778–95 (cit. on p. 120).

Kufareva, Irina, Levon Budagyan, Eugene Raush, Maxim Totrov, and Ruben Abagyan (2007). "PIER: protein interface recognition for structural proteomics." *Proteins* 67.2, pp. 400–417 (cit. on pp. 90, 92, 93).

Kumar, S, J A Dorsey, R A Muesing, and J W Porter (1970). "Comparative studies of the pigeon liver fatty acid synthetase complex and its subunits. Kinetics of partial reactions and the number of binding sites for acetyl and malonyl groups." *The Journal of biological chemistry* 245.18, pp. 4732–44 (cit. on p. 9).

Kwon, Seok Joon, Mauricio Mora-Pale, Moo-Yeal Lee, and Jonathan S Dordick (2012). "Expanding nature's small molecule diversity via in vitro biosynthetic pathway engineering." *Current opinion in chemical biology* 16.1-2, pp. 186–95 (cit. on p. 60).

Larkin, M A, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins (2007). "Clustal W and Clustal X version 2.0." *Bioinformatics (Oxford, England)* 23.21, pp. 2947–2948 (cit. on p. 76).

Laskowski, R A, M W MacArthur, D S Moss, and J M Thornton (1993). "PROCHECK: a program to check the stereochemical quality of protein structures". *Journal of Applied Crystallography* 26.2, pp. 283–291 (cit. on p. 76).

Lau, J, H Fu, D E Cane, and C Khosla (1999). "Dissecting the role of acyltransferase domains of modular polyketide synthases in the choice and stereochemical fate of extender units." *Biochemistry* 38.5, pp. 1643–51 (cit. on p. 29).

Leibundgut, Marc, Simon Jenni, Christian Frick, and Nenad Ban (2007). "Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase." *Science (New York, N.Y.)* 316.5822, pp. 288–290 (cit. on p. 51).

Letunic, Ivica, Tobias Doerks, and Peer Bork (2009). "SMART 6: recent updates and new developments." *Nucleic acids research* 37.Database issue, pp. D229–32 (cit. on p. 308).

Letunic, Ivica, Tobias Doerks, and Peer Bork (2012). "SMART 7: recent updates to the protein domain annotation resource." *Nucleic acids research* 40.Database issue, pp. D302–5 (cit. on p. 303).

Li, Michael H T, Peter M U Ung, James Zajkowski, Sylvie Garneau, David H Sherman, and Sylvie Garneau-Tsodikova (2009). "Automated genome mining for natural products." *BMC bioinformatics* 10, p. 185 (cit. on p. 67).

Lichtarge, Olivier, Henry R Bourne, and Fred E Cohen (1996). "An evolutionary trace method defines binding surfaces common to protein families." *Journal of molecular biology* 257.2, pp. 342–58 (cit. on pp. 90, 91, 140).

Lindorff-Larsen, Kresten, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw (2010). "Improved side-chain torsion potentials for the Amber ff99SB protein force field". *Proteins: Structure, Function and Bioinformatics* 78.8, pp. 1950–1958 (cit. on pp. 79, 81, 82).

Liou, Grace F, Janice Lau, David E Cane, and Chaitan Khosla (2003). "Quantitative analysis of loading and extender acyltransferases of modular polyketide synthases." *Biochemistry* 42.1, pp. 200–7 (cit. on p. 29).

Ma, Suzanne M and Yi Tang (2007). "Biochemical characterization of the minimal polyketide synthase domains in the lovastatin nonaketide synthase LovB." *The FEBS journal* 274.11, pp. 2854–64 (cit. on p. 18).

MacKerell, A D, D Bashford, M Bellott, R L Dunbrack, J D Evanseck, M J Field, S Fischer, J Gao, H Guo, S Ha, D Joseph-McCarthy, L Kuchnir, K Kuczera, F T K Lau, C Mattos, S Michnick, T Ngo, D T Nguyen, B Prodhom, W E Reiher, B Roux, M Schlenkrich, J C Smith, R Stote, J Straub, M Watanabe, J Wiorkiewicz-Kuczera, D Yin, and M Karplus (1998). "All-atom empirical potential for molecular modeling and dynamics studies of proteins". *Journal of Physical Chemistry B* 102, pp. 3586–3616 (cit. on p. 79).

Maier, Timm, Simon Jenni, and Nenad Ban (2006). "Architecture of mammalian fatty acid synthase at 4.5 A resolution." *Science (New York, N.Y.)* 311.5765, pp. 1258–62 (cit. on pp. 46, 47).

Maier, Timm, Marc Leibundgut, and Nenad Ban (2008). "The crystal structure of a mammalian fatty acid synthase." *Science (New York, N.Y.)* 321.5894, pp. 1315–22 (cit. on pp. 41, 46, 306).

Malpartida, F and D A Hopwood (1984). "Molecular cloning of the whole biosynthetic pathway of a Streptomyces antibiotic and its expression in a heterologous host". *Nature* 309.5967, pp. 462–464 (cit. on p. 6).

Marchler-bauer, Aron, Shennan Lu, John B Anderson, Farideh Chitsaz, Myra K Derbyshire, Carol DeWeese-Scott, Jessica H Fong, Lewis Y Geer, Renata C Geer, Noreen R Gonzales, Marc Gwadz, David I Hurwitz, John D Jackson, Zhaoxi Ke, Christopher J Lanczycki, Fu Lu, Gabriele H Marchler, Mikhail Mullokandov, Marina V Omelchenko, Cynthia L Robertson, James S Song, Narmada Thanki, Roxanne A Yamashita, Dachuan Zhang, Naigong Zhang, Chanjuan Zheng, and Stephen H Bryant (2011). "CDD: a Conserved Domain Database for

the functional annotation of proteins." *Nucleic acids research* 39.Database issue, pp. D225–9 (cit. on p. 302).

Markley, J L, A Bax, Y Arata, C W Hilbers, R Kaptein, B D Sykes, P E Wright, and K Wüthrich (1998). "Recommendations for the presentation of NMR structures of proteins and nucleic acids–IUPAC-IUBMB-IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy." *European journal of biochemistry / FEBS* 256.1, pp. 1–15 (cit. on p. 119).

Martí-Renom, M A, Ashley C Stuart, A Fiser, R Sánchez, Francisco Melo, and Andrej Sali (2000). "Comparative protein structure modeling of genes and genomes." *Annual review of biophysics and biomolecular structure* 29, pp. 291–325 (cit. on p. 75).

Medema, Marnix H, Kai Blin, Peter Cimermancic, Victor de Jager, Piotr Zakrzewski, Michael A Fischbach, Tilmann Weber, Eriko Takano, Rainer Breitling, and Victor De Jager (2011). "antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences." *Nucleic acids research* 39.Web Server issue, W339–46 (cit. on pp. 66, 67, 308).

Menke, Matthew, Bonnie Berger, and Lenore Cowen (2008). "Matt: local flexibility aids protein multiple structure alignment." *PLoS computational biology* 4.1, e10 (cit. on p. 87).

Menzella, Hugo G, Ralph Reid, John R Carney, Sunil S Chandran, Sarah J Reisinger, Kedar G Patel, David a Hopwood, and Daniel V Santi (2005). "Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes." *Nature biotechnology* 23.9, pp. 1171–1176 (cit. on p. 221).

Mihalek, I, I Res, and O Lichtarge (2004). "A family of evolution-entropy hybrid methods for ranking protein residues by importance." *Journal of molecular biology* 336.5, pp. 1265–82 (cit. on p. 92).

Minowa, Yohsuke, Michihiro Araki, and Minoru Kanehisa (2007). "Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes." *Journal of molecular biology* 368.5, pp. 1500–17 (cit. on p. 308).

Mishra, Bhuwan B and Vinod K Tiwari (2011a). "Natural products: an evolving role in future drug discovery." *European journal of medicinal chemistry* 46.10, pp. 4769–807 (cit. on p. 3).

— (2011b). *Natural products in drug discovery : Clinical evaluations and investigations*. Vol. 661. 2, pp. 1–62 (cit. on p. 3).

Misra, Ila and Henry M Miziorko (1996). "Evidence for the interaction of avian 3-hydroxy-3-methylglutaryl-CoA synthase histidine 264 with acetoacetyl-CoA." *Biochemistry* 35.29, pp. 9610–6 (cit. on p. 124).

Moche, M, G Schneider, P Edwards, K Dehesh, and Y Lindqvist (1999). "Structure of the complex between the antibiotic cerulenin and its target, beta-ketoacyl-acyl carrier protein synthase." *The Journal of biological chemistry* 274.10, pp. 6031–4 (cit. on p. 11).

Moir, A, E Lafferty, and D A Smith (1979). "Genetics analysis of spore germination mutants of Bacillus subtilis 168: the correlation of phenotype with map location." *Journal of general microbiology* 111.1, pp. 165–80 (cit. on p. 97).

Mooers, Blaine H M, Deepshikha Datta, Walter a. Baase, Eric S. Zollars, Stephen L. Mayo, and Brian W. Matthews (2003). "Repacking the core of T4 lysozyme by automated design". *Journal of Molecular Biology* 332.3, pp. 741–756 (cit. on p. 152).

Mootz, Henning D., Robert Finking, and Mohamed a. Marahiel (2001). "4'-phosphopantetheine transfer in primary and secondary metabolism of Bacillus subtilis." *The Journal of biological chemistry* 276.40, pp. 37289–98 (cit. on p. 24).

Morgan, Daniel H, David M Kristensen, David Mittelman, and Olivier Lichtarge (2006). "ET viewer: an application for predicting and visualizing functional sites in protein structures." *Bioinformatics (Oxford, England)* 22.16, pp. 2049–50 (cit. on pp. 92, 140).

Nguyen, TuAnh, Keishi Ishida, Holger Jenke-Kodama, Elke Dittmann, Cristian Gurgui, Thomas Hochmuth, Stefan Taudien, Matthias Platzer, Christian Hertweck, and Jörn Piel (2008). "Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection." *Nature biotechnology* 26.2, pp. 225–33 (cit. on pp. 35, 221).

Olsen, J G, A Kadziola, P von Wettstein-Knowles, M Siggaard-Andersen, Y Lindquist, and S Larsen (1999). "The X-ray crystal structure of beta-ketoacyl [acyl carrier protein] synthase I." *FEBS letters* 460.1, pp. 46–52 (cit. on p. 34).

Olsen, J G, A Kadziola, P von Wettstein-Knowles, M Siggaard-Andersen, and S Larsen (2001). "Structures of beta-ketoacyl-acyl carrier protein synthase I complexed with fatty acids elucidate its catalytic machinery." *Structure (London, England : 1993)* 9.3, pp. 233–43 (cit. on pp. 11, 30).

Oostenbrink, Chris, Alessandra Villa, Alan E. Mark, and Wilfred F. Van Gunsteren (2004). "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6". *Journal of Computational Chemistry* 25, pp. 1656–1676 (cit. on p. 79).

Pan, Hu, Shiou Chuan Tsai, Eric S Meadows, Larry J W Miercke, Adrian T Keatinge-Clay, Joe O'Connell, Chaitan Khosla, and Robert M Stroud (2002). "Crystal structure of the priming beta-ketosynthase from the R1128 polyketide biosynthetic pathway." *Structure (London, England : 1993)* 10.11, pp. 1559–68 (cit. on p. 32).

Paramo, Teresa, Alexandra East, Diana Garzón, Martin B. Ulmschneider, and Peter J. Bond (2014). "Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: trj_cavity". *Journal of Chemical Theory and Computation* 10.5, pp. 2151–2164 (cit. on p. 87).

Parenti, M A, S M Hatfield, and J J Leyden (1987). "Mupirocin: a topical antibiotic with a unique structure and mechanism of action." *Clinical pharmacy* 6.10, pp. 761–70 (cit. on p. 2).

Park, Sung Ryeol, Ah Reum Han, Yeon-Hee Ban, Young Ji Yoo, Eun Ji Kim, and Yeo Joon Yoon (2010). "Genetic engineering of macrolide biosynthesis: past advances, current state, and future prospects." *Applied microbiology and biotechnology* 85.5, pp. 1227–39 (cit. on p. 60).

Parrinello, M., a. Rahman, and Rahman a. Parrinello M (1980). "Crystal Structure and Pair Potentials: A Molecular_Dynamics Study". *Physical Review Letters* 45.14, pp. 1196–1199 (cit. on p. 83).

Patel, Jean B, Rachel J Gorwitz, and John A Jernigan (2009). "Mupirocin resistance." *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 49.6, pp. 935–41 (cit. on p. 63).

Pettersen, Eric F, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin (2004). "UCSF Chimera–a visualization system for exploratory research and analysis." *Journal of computational chemistry* 25.13, pp. 1605–1612 (cit. on p. 76).

Piel, Jörn (2010). "Biosynthesis of polyketides by trans-AT polyketide synthases." *Natural product reports* 27.7, pp. 996–1047 (cit. on p. 60).

Pierce, Brian G, Kevin Wiehe, Howook Hwang, Bong-Hyun Kim, Thom Vreven, and Zhiping Weng (2014). "ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers." *Bioinformatics (Oxford, England)* 30.12, pp. 1771–3 (cit. on p. 94).

Powers, J H (2004). "Antimicrobial drug development–the past, the present, and the future." *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 10 Suppl 4, pp. 23–31 (cit. on pp. 2, 5).

Price, Allen C, Charles O Rock, and Stephen W White (2003). "The 1.3-Angstrom-resolution crystal structure of beta-ketoacyl-acyl carrier protein synthase II from Streptococcus pneumoniae." *Journal of bacteriology* 185.14, pp. 4136–43 (cit. on p. 11).

Prieto, Carlos, Carlos García-Estrada, Diego Lorenzana, and Juan Francisco Martín (2012). "NRPSsp: non-ribosomal peptide synthase substrate predictor." *Bioinformatics (Oxford, England)* 28.3, pp. 426–7 (cit. on p. 67).

Pronk, Sander, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David van der Spoel, Berk Hess, Erik Lindahl, and David Van Der Spoel (2013). "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit." *Bioinformatics (Oxford, England)* 29.7, pp. 845–54 (cit. on p. 79).

Qiu, X, C A Janson, W W Smith, M Head, J Lonsdale, and a K Konstantinidis (2001). "Refined structures of beta-ketoacyl-acyl carrier protein synthase III." *Journal of molecular biology* 307.1, pp. 341–56 (cit. on p. 22).

Qiu, Xiayang, Cheryl A Janson, Alex K Konstantinidis, Silas Nwagwu, Carol Silverman, Ward W Smith, Sanjay Khandekar, John Lonsdale, and S S Abdel-Meguid (1999). "Crystal structure of beta-ketoacyl-acyl carrier protein synthase III. A key condensing enzyme in bacterial fatty acid biosynthesis." *The Journal of biological chemistry* 274.51, pp. 36465–71 (cit. on p. 31).

Rahman, Ayesha S., Joanne Hothersall, John Crosby, Thomas J. Simpson, and Christopher M. Thomas (2005). "Tandemly duplicated acyl carrier proteins, which increase polyketide

antibiotic production, can apparently function either in parallel or in series". *Journal of Biological Chemistry* 280, pp. 6399–6408 (cit. on p. 97).

Rausch, Christian, Tilmann Weber, Oliver Kohlbacher, Wolfgang Wohlleben, and Daniel H Huson (2005). "Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs)." *Nucleic acids research* 33.18, pp. 5799–808 (cit. on pp. 304, 308).

Rausch, Christian, Ilka Hoof, Tilmann Weber, Wolfgang Wohlleben, and Daniel H Huson (2007). "Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution." *BMC evolutionary biology* 7, p. 78 (cit. on p. 308).

Reid, Ralph, Misty Piagentini, Eduardo Rodriguez, Gary Ashley, Nina Viswanathan, John Carney, Daniel V. Santi, C. Richard Hutchinson, and Robert McDaniel (2003). "A model of structure and catalysis for ketoreductase domains in modular polyketide synthases". *Biochemistry* 42, pp. 72–79 (cit. on pp. 39, 40).

Rennell, D., S. E. Bouvier, L. W. Hardy, and a. R. Poteete (1991). "Systematic mutation of bacteriophage T4 lysozyme". *Journal of Molecular Biology* 222.1, pp. 67–87 (cit. on p. 152).

Röttig, Marc, Marnix H Medema, Kai Blin, Tilmann Weber, Christian Rausch, and Oliver Kohlbacher (2011). "NRPSpredictor2–a web server for predicting NRPS adenylation domain specificity." *Nucleic acids research* 39.Web Server issue, W362–7 (cit. on pp. 67, 304, 308).

Saga, Tomoo and Keizo Yamaguchi (2009). "History of Antimicrobial Agents and Resistant". *Jmaj* 137.3, pp. 103–108 (cit. on pp. 1, 2).

Sali, A and T L Blundell (1993). "Comparative protein modelling by satisfaction of spatial restraints." *Journal of molecular biology* 234.3, pp. 779–815 (cit. on p. 75).

Shafqat, Naeem, Andrew Turnbull, Johannes Zschocke, Udo Oppermann, and Wyatt W Yue (2010). "Crystal structures of human HMG-CoA synthase isoforms provide insights into inherited ketogenesis disorders and inhibitor design." *Journal of molecular biology* 398.4, pp. 497–506 (cit. on pp. 111, 112, 124, 129, 130).

Shapovalov, Maxim V. and Roland L. Dunbrack (2011). "A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions." *Structure (London, England : 1993)* 19.6, pp. 844–58 (cit. on p. 119).

Sharma, Krishna K. and Christopher N. Boddy (2007). "The thioesterase domain from the pimaricin and erythromycin biosynthetic pathways can catalyze hydrolysis of simple thioester substrates". *Bioorganic and Medicinal Chemistry Letters* 17, pp. 3034–3037 (cit. on p. 44).

Simon, R., U. Priefer, and A. Pühler (1983). "A Broad Host Range Mobilization System for In Vivo Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria". *Bio/Technology* 1.9, pp. 784–791 (cit. on p. 97).

Simunovic, Vesna, Josef Zapp, Shwan Rachid, Daniel Krug, Peter Meiser, and Rolf Müller (2006). "Myxovirescin A biosynthesis is directed by hybrid polyketide synthases/nonribosomal peptide synthetase, 3-hydroxy-3-methylglutaryl-CoA synthases, and trans-acting

acyltransferases." *Chembiochem : a European journal of chemical biology* 7.8, pp. 1206–20 (cit. on p. 109).

Singh, N, S J Wakil, and J K Stoops (1984). "On the question of half- or full-site reactivity of animal fatty acid synthetase." *The Journal of biological chemistry* 259.6, pp. 3605–11 (cit. on p. 9).

Smith, S and S Abraham (1971). "Fatty acid synthetase from lactating rat mammary gland. 3. Dissociation and reassociation." *The Journal of biological chemistry* 246.21, pp. 6428–35 (cit. on p. 9).

Smith, Stuart and Shiou-Chuan Tsai (2007). "The type I fatty acid and polyketide synthases: a tale of two megasynthases." *Natural product reports* 24.5, pp. 1041–72 (cit. on pp. 10, 29, 30, 33).

Spoel, David Van Der, Paul J Van Maaren, Herman J C Berendsen, and I Introduction (1998). "A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field". *Journal of Chemical Physics* 108.24, pp. 10220–10230 (cit. on p. 79).

Stachelhaus, T, H D Mootz, and M A Marahiel (1999). "The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases." *Chemistry & biology* 6.8, pp. 493–505 (cit. on p. 305).

Starcevic, Antonio, Jurica Zucko, Jurica Simunkovic, Paul F Long, John Cullum, and Daslav Hranueli (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, pp. 6882–92 (cit. on pp. 67, 303, 308).

Starcevic, Antonio, Kerstin Wolf, Janko Diminic, Jurica Zucko, Ida Trninic Ruzic, Paul F Long, Daslav Hranueli, and John Cullum (2012). "Recombinatorial biosynthesis of polyketides." *Journal of industrial microbiology & biotechnology* 39.3, pp. 503–11 (cit. on p. 304).

Staunton, J, P Caffrey, J F Aparicio, G A Roberts, S S Bethell, and P F Leadlay (1996). "Evidence for a double-helical structure for modular polyketide synthases." *Nature structural biology* 3.2, pp. 188–92 (cit. on pp. 12, 13).

Staunton, James and Kira J. Weissman (2001). "Polyketide biosynthesis: a millennium review." *Natural Product Reports* 18.4, pp. 380–416 (cit. on pp. 3, 7).

Steussy, C Nicklaus, Anthony A Vartia, John W Burgner, Autumn Sutherlin, Victor W Rodwell, and Cynthia V Stauffacher (2005). "X-ray crystal structures of HMG-CoA synthase from Enterococcus faecalis and a complex with its second substrate/inhibitor acetoacetyl-CoA." *Biochemistry* 44.43, pp. 14256–67 (cit. on pp. 96, 111, 112, 124, 129).

Stoops, J K, S J Wakil, E C Uberbacher, and G J Bunick (1987). "Small-angle neutron-scattering and electron microscope studies of the chicken liver fatty acid synthase." *The Journal of biological chemistry* 262.21, pp. 10246–51 (cit. on p. 46).

Sugimoto, Yuki, Ling Ding, Keishi Ishida, and Christian Hertweck (2014). "Rational Design of Modular Polyketide Synthases: Morphing the Aureothin Pathway into a Luteoreticulin Assembly Line." *Angewandte Chemie (International ed. in English)*, pp. 1–5 (cit. on p. 2).

Szu, Ping-Hui Hui, Sridhar Govindarajan, Michael J. Meehan, Abhirup Das, Don D. Nguyen, Pieter C. Dorrestein, Jeremy Minshull, and Chaitan Khosla (2011). "Analysis of the ketosynthase chain length factor heterodimer from the fredericamycin polyketide synthase." *Chemistry & biology* 18.8, pp. 1021–31 (cit. on p. 30).

Tae, Hongseok, Eun-Bae Kong, and Kiejung Park (2007). "ASMPKS: an analysis system for modular polyketide synthases." *BMC bioinformatics* 8, p. 327 (cit. on pp. 67, 302).

Tang, Yi, Shiou-Chuan Tsai, and Chaitan Khosla (2003). "Polyketide chain length control by chain length factor." *Journal of the American Chemical Society* 125.42, pp. 12708–9 (cit. on p. 30).

Tang, Yinyan, Ho Young Lee, Yi Tang, Chu-Young Kim, Irimpan Mathews, and Chaitan Khosla (2006a). "Structural and functional studies on SCO1815: a beta-ketoacyl-acyl carrier protein reductase from Streptomyces coelicolor A3(2)." *Biochemistry* 45.47, pp. 14085–93 (cit. on p. 21).

Tang, Yinyan, Chu-Young Kim, Irimpan I Mathews, David E Cane, and Chaitan Khosla (2006b). "The 2.7-Angstrom crystal structure of a 194-kDa homodimeric fragment of the 6-deoxyerythronolide B synthase." *Proceedings of the National Academy of Sciences of the United States of America* 103.30, pp. 11124–9 (cit. on pp. 27–30, 50, 55, 57).

Tang, Yinyan, Alice Y Chen, Chu-Young Kim, David E Cane, and Chaitan Khosla (2007). "Structural and mechanistic analysis of protein interactions in module 3 of the 6-deoxyerythronolide B synthase." *Chemistry & biology* 14.8, pp. 931–43 (cit. on pp. 50, 55, 57, 203, 306).

Teague, Simon J (2003). "Implications of protein flexibility for drug discovery." *Nature reviews. Drug discovery* 2.7, pp. 527–41 (cit. on p. 93).

Theisen, Michael J, Ila Misra, Dana Saadat, Nino Campobasso, Henry M Miziorko, and David H T Harrison (2004). "3-hydroxy-3-methylglutaryl-CoA synthase intermediate complex observed in "real-time"." *Proceedings of the National Academy of Sciences of the United States of America* 101.47, pp. 16442–7 (cit. on pp. 112, 124, 129, 130).

Thomas, Christopher M, Joanne Hothersall, Christine L Willis, and Thomas J Simpson (2010). "Resistance to and synthesis of the antibiotic mupirocin." *Nature reviews. Microbiology* 8.4, pp. 281–9 (cit. on pp. 63, 65).

Tokuriki, Nobuhiko and Dan S. Tawfik (2009). "Stability effects of mutations and protein evolvability". *Current Opinion in Structural Biology* 19.5, pp. 596–604 (cit. on p. 152).

Tsai, S C, L J Miercke, J Krucinski, R Gokhale, J C Chen, P G Foster, D E Cane, C Khosla, and R M Stroud (2001). "Crystal structure of the macrocycle-forming thioesterase domain of the erythromycin polyketide synthase: versatility from a unique substrate channel." *Proceedings of the National Academy of Sciences of the United States of America* 98, pp. 14808–14813 (cit. on pp. 45, 46).

Tsai, Shiou-Chuan Sheryl and Brian Douglas Ames (2009). "Structural enzymology of polyketide synthases." *Methods in enzymology* 459.09, pp. 17–47 (cit. on pp. 27, 306).

Tuan, James S, J.Mark Weber, Michael J Staver, Judith O Leung, Stefano Donadio, and Leonard Katz (1990). "Cloning of genes involved in erythromycin biosynthesis from Saccharopolyspora erythraea using a novel actinomycete-Escherichia coli cosmid". *Gene* 90.1, pp. 21–29 (cit. on p. 6).

Valley, Christopher C, Alessandro Cembran, Jason D Perlmutter, Andrew K Lewis, Nicholas P Labello, Jiali Gao, Jonathan N Sachs, and N Jonathan (2012). "The methionine-aromatic motif plays a unique role in stabilizing protein structure." *The Journal of biological chemistry* 287.42, pp. 34979–91 (cit. on pp. 144, 152).

Vanquelef, Enguerran, Sabrina Simon, Gaelle Marquant, Elodie Garcia, Geoffroy Klimerak, Jean Charles Delepine, Piotr Cieplak, and François Yves Dupradeau (2011). "R.E.D. Server: A web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments". *Nucleic Acids Research* 39.May, pp. 511–517 (cit. on pp. 81, 180).

Vriend, G (1990). "WHAT IF: a molecular modeling and drug design program." *Journal of molecular graphics* 8.1, pp. 52–6, 29 (cit. on p. 136).

Vries, Sjoerd J de, Marc van Dijk, and Alexandre M J J Bonvin (2010). "The HADDOCK web server for data-driven biomolecular docking." *Nature protocols* 5.5, pp. 883–97 (cit. on pp. 94, 135).

Wakil, Sahil J. and Jim K. Stoops (1983). "Structure and Mechanism of Fatty Acid Synthetase". *The Enzymes*. Ed. by P. D. Boyer. XVI. New York: Academic Press. Chap. 3 (cit. on p. 9).

Wallace, A C, R A Laskowski, and J M Thornton (1995). "LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions." *Protein engineering* 8.2, pp. 127–34 (cit. on p. 127).

Wang, J M, R M Wolf, J W Caldwell, P A Kollman, and D A Case (2004). "Development and testing of a general amber force field". *J. Comput. Chem.* 25, pp. 1157–1174 (cit. on p. 81).

Watanabe, Kenji, Clay C C Wang, Christopher N Boddy, David E Cane, and Chaitan Khosla (2003). "Understanding substrate specificity of polyketide synthase modules by generating hybrid multimodular synthases." *The Journal of biological chemistry* 278.43, pp. 42020–6 (cit. on p. 221).

Weber, T, C Rausch, P Lopez, I Hoof, V Gaykova, D H Huson, and W Wohlleben (2009). "CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters." *Journal of biotechnology* 140.1-2, pp. 13–7 (cit. on pp. 67, 307, 308).

Weissman, Kira J (2006). "The structural basis for docking in modular polyketide biosynthesis." *Chembiochem : a European journal of chemical biology* 7.3, pp. 485–94 (cit. on p. 307).

Weissman, Kira J and Peter F Leadlay (2005). "Combinatorial biosynthesis of reduced polyketides." *Nature reviews. Microbiology* 3.12, pp. 925–36 (cit. on pp. 3, 60).

Weissman, Kira J and Rolf Müller (2008). "Protein-protein interactions in multienzyme megasynthetases." *Chembiochem : a European journal of chemical biology* 9.6, pp. 826–48 (cit. on p. 307).

Whicher, Jonathan R, Somnath Dutta, Douglas a Hansen, Wendi A Hale, Joseph a Chemler, Annie M Dosey, Alison R H Narayan, Kristina Hå kansson, David H Sherman, Janet L Smith, and Georgios Skiniotis (2014). "Structural rearrangements of a polyketide synthase module during its catalytic cycle." *Nature* 510.7506, pp. 560–4 (cit. on p. 59).

Winn, P J, G G Ferenczy, and C a Reynolds (1997). "Towards Improved Force fields: {I}. Multipole-derived atomic charges". *J. Phys. Chem. A.* 101.97, pp. 5437–5445 (cit. on p. 84).

Witkowski, A, A Joshi, and S Smith (1996). "Fatty acid synthase: in vitro complementation of inactive mutants." *Biochemistry* 35.32, pp. 10569–75 (cit. on p. 9).

Witkowski, Andrzej, Anil K Joshi, and Stuart Smith (2004a). "Characterization of the beta-carbon processing reactions of the mammalian cytosolic fatty acid synthase: role of the central core." *Biochemistry* 43, pp. 10458–10466 (cit. on p. 44).

Witkowski, Andrzej, Alokesh Ghosal, Anil K Joshi, H Ewa Witkowska, Francisco J Asturias, and Stuart Smith (2004b). "Head-to-head coiled arrangement of the subunits of the animal fatty acid synthase." *Chemistry & biology* 11.12, pp. 1667–76 (cit. on pp. 12, 47).

Wu, Ji'en, Sian M Cooper, Russell J Cox, John Crosby, Matthew P Crump, Joanne Hothersall, Thomas J Simpson, Christopher M Thomas, and Christine L Willis (2007). "Mupirocin H, a novel metabolite resulting from mutation of the HMG-CoA synthase analogue, mupH in Pseudomonas fluorescens." *Chemical communications (Cambridge, England)* 8.20, pp. 2040–2 (cit. on pp. 64, 97, 110, 133).

Wu, Ji'en, Joanne Hothersall, Carlo Mazzetti, Yvonne O'Connell, Jennifer A Shields, Ayesha S Rahman, Russell J Cox, John Crosby, Thomas J Simpson, Christopher M Thomas, and Christine L Willis (2008). "In vivo mutational analysis of the mupirocin gene cluster reveals labile points in the biosynthetic pathway: the "leaky hosepipe" mechanism." *Chembiochem : a European journal of chemical biology* 9.9, pp. 1500–8 (cit. on pp. 202, 220).

Yadav, Gitanjali, Rajesh S Gokhale, and Debasisa Mohanty (2003a). "Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases." *Journal of molecular biology* 328.2, pp. 335–63 (cit. on pp. 29, 66, 303, 305, 308).

— (2003b). "SEARCHPKS: A program for detection and analysis of polyketide synthase domains." *Nucleic acids research* 31.13, pp. 3654–8 (cit. on pp. 30, 302).

— (2009). "Towards prediction of metabolic products of polyketide synthases: an in silico analysis." *PLoS computational biology* 5.4, e1000351 (cit. on pp. 305, 307, 308).

Zhang, Yong-Mei, Bainan Wu, Jie Zheng, and Charles O Rock (2003). "Key residues responsible for acyl carrier protein and beta-ketoacyl-acyl carrier protein reductase (FabG) interaction." *The Journal of biological chemistry* 278.52, pp. 52935–43 (cit. on pp. 25, 48).

Zhang, Yong-mei M, Mohan S Rao, Richard J Heath, Allen C Price, Arthur J Olson, Charles O Rock, and Stephen W White (2001). "Identification and analysis of the acyl carrier protein (ACP) docking site on beta-ketoacyl-ACP synthase III." *The Journal of biological chemistry* 276.11, pp. 8231–8 (cit. on p. 25).

Zheng, Jianting and Adrian T. Keatinge-Clay (2011). “Structural and functional analysis of C2-type ketoreductases from modular polyketide synthases”. *Journal of Molecular Biology* 410.1, pp. 105–117 (cit. on p. 39).

Chapter A

Appendix I

A.1 Steps involved in the HMM analysis

Step 1. Built the HMM models for the branching ACPs (38 seq; wacp.hmm) and non-branching ACPs (178 seq; stdacp.hmm) using the sequences provided by Dr. Anthony Haines.
Step 2. Searched, scored and plotted the HMMs against the sequences used for model building of both the models, to identify the respective clusters.
Step 3. Searched, scored and plot the HMMs against the test sequences (provided by Dr. Anthony Haines) for both the models, to confirm the selectivity of the models and clustering.
Step 4. Fetched the sequences using stdACP hmm from Refseq microbial (6408654 seq) and Uniprot Trembl (20127441 seq) database.
Step 5. Used Perl scripts (Script A.2.1 and Script A.2.2) to read the hmmsearch (a tool in HMMER3 suite) output, extracted the sequence for the each domain matched and arranged the results in decreasing order of sequence length. Only the sequences above the length of 60 residues were considered. The length of the model used was 67.
Step 6. Duplicate sequences were removed from both the sets using Script A.2.3.
Step 7. A multiple sequence alignment was carried out for both the sets and checked for the presence of the active site serine (Script A.2.4). The sequences which lacked the active site serine at the aligned position were removed.
Step 8. Both the sets were merged and again searched for any duplicates resulting in the final set of 10076 sequences.
Step 9. The filtered set resulted in 16490 sequences which were then further extended by 7 residues on both the ends to ensure they would cover the full length of the models (Script A.2.5).
Step 10. The final set of 16490 extended sequences was then scored using both the HMM models and plotted.

A.2 Scripts used in the HMM analysis

A.2.1 Script to extract the individual domain sequence matched by stdACP model against the RefSeq database

```
#!/usr/bin/perl
```

```
use warnings;
use strict;
use Data::Dumper;

# This script is used to extract the domain sequences matched by the stdACP
    hmm model against refseq database
# It reads the hmmsearch output and search for each domain sequences and the
    domain number.

my @fastatag;
my @sequence;
my @winnDomains;
my @len;
my $winnDomain;

while( <> ) {
       my $fa = undef;
       my $seq = undef;
       $winnDomain = $1 if /\=\= domain\s(\d+)/; # Search for the line which
           starts with ==domain num and read the number of the domain
       if( m/^\s+(gi\|[\w\|\_\-\.]+|sp\|[\w\|\_\-\.]+)\s+\d+.([A-Za-z_\-]+)/ )
           {
               $fa = $1; # Store the name or tag
               $seq = $2; # Store the sequence
       }

       # Push everything catched above in an array
       if( $fa && $seq ) {
              $seq =~ s/\-//g;
              push @winnDomains, $winnDomain;
              push @fastatag, $fa;
              push @sequence, $seq;
              push @len, length( $seq );
       }
}

# Sort all the sequences according to lenght in decreasing order
my @sorted = sort { $len[$b] <=> $len[$a] } 0 .. $#len;

#print Dumper @fastatag; # This is to crosscheck

# Output the final list
for my $i ( @sorted ) {
       if(length($sequence[$i])>=60)
       {
       print '>'. $fastatag[$i] . "Domain_" . $winnDomains[$i] . "\n";
       print $sequence[$i] . "\n\n\n";
       }
```

```
        else{ last;}
}

print "Number of sequences in sorted array", scalar @sorted;
```

A.2.2 Script to extract the individual domain sequence matched by stdACP model against the TrEMBL database

```
#!/usr/bin/perl

use warnings;
use strict;
use Data::Dumper;

# This script is used to extract the domain sequences matched by the stdACP
    hmm model against trembl database
# It reads the hmmsearch output and search for each domain sequences and the
    domain number.

my @fastatag;
my @sequence;
my @winnDomains;
my @len;
my $winnDomain;

while( <> ) {
        my $fa = undef;
        my $seq = undef;
        $winnDomain = $1 if /\=\= domain\s(\d+)/; # Search for the line which
            starts with ==domain num and read the number of the domain
        if( m/^\s+(tr\|[\w\|\_\-\.]+|sp\|[\w\|\_\-\.]+)\s+\d+.([A-Za-z_\-]+)/ )
            {
                $fa = $1;#Store the name or tag
                $seq = $2;#Sotre the sequence
        }

        # Push everything catched above in an array
        if( $fa && $seq ) {
                $seq =~ s/\-//g;
                push @winnDomains, $winnDomain;
                push @fastatag, $fa;
                push @sequence, $seq;
                push @len, length( $seq );
        }
}

# Sort all the sequences according to lenght in decreasing order
my @sorted = sort { $len[$b] <=> $len[$a] } 0 .. $#len;
```

```
#print Dumper @fastatag; # This is to crosscheck

# Output the final list
for my $i ( @sorted ) {
       if(length($sequence[$i])>=60)
       {
       print '>'. $fastatag[$i] . "|Domain_" . $winnDomains[$i] . "\n";
       print $sequence[$i] . "\n\n\n";
       }
       else{ last;}
}

print "Number of sequences in sorted array: ",scalar @sorted;
```

A.2.3 Script to eliminate the duplicate sequences

```
#!/usr/bin/perl

use strict;
use Bio::SeqIO;

# This script is used to filter the duplicate sequences using BioPerl module

my %unique;
my $file  = $ARGV[0];
my $seqio = Bio::SeqIO->new(-file => $file, -format => "fasta");
my $outseq = Bio::SeqIO->new(-file => ">$file.unique.fasta", -format =>
    "fasta");

while(my $seqs = $seqio->next_seq) {
  my $id = $seqs->display_id;
  my $seq = $seqs->seq;
  unless(exists($unique{$seq})) {
    $outseq->write_seq($seqs);
    $unique{$seq} +=1;
  }
}
```

A.2.4 Script to check for active site serine in the multiple sequence alignment output file

```
#!/usr/bin/perl

use strict;
use warnings;
use Bio::SeqIO;

# This script is used to check for catalytic S at the aligned position in a
    MSA file
```

```
my $in = Bio::SeqIO->new(-file =>
    "rerun_refseq_trembl_amalgamated_unique_aligned.fasta" , '-format' =>
    'Fasta');
my $count =0;
open(my $out, ">rerun_refseq_trembl_amalgamated_filtered_S.fasta");

while ( my $seq = $in->next_seq() )
{
 my $s = $seq->subseq(361,361);
  if($s eq 'S')
   {
    print $out ">".$seq->id,"\n";
    print $out $seq->seq(),"\n";
    $count++;
   }
}
print $count;
```

A.2.5 Script to extend the sequences on either ends by 7 residues

Note: This script was written with the help of Dr. Anthony Haines.

```
#!/usr/bin/perl -w
use strict;

# Open the database containing 10076 sequences
open(FILE1, "rerun_refseq_trembl_amalgamated_filtered_S_nodash.fasta") || die
    "can't open file";
my $seq;
my $tag;
my @arr;
my %tag_seq;
my %useful;
while(<FILE1>)
 {
  if($_ =~ m/^\>(.+)/ ) # Match for the fasta tag line and store it in @arr
    {
    $tag = $1;
    push @arr,$1;
    $tag_seq{$tag}= "";
    $tag =~ m/^(.+\|)Domain\_(\d+)/; # Exclude the Domain_num from the fasta
        tag line and extract the highest domain number
                                    # To read refseq database
   #$tag =~ m/^(.+)\|Domain\_(\d+)/; # To read the trembl database

    if(defined($useful{$1}))
     {
      if($useful{$1}<$2)
      {
```

```
      $useful{$1}=$2;
     }

    }
    else
    {
        $useful{$1}=$2;
    }
  # print "$tag $useful{$1}\n";
   }
  else
   {
    $seq=$_;
    chomp($seq);
    $tag_seq{$tag}.=$seq; # Store the sequence in the has %tag_seq
    }
}
 #print $#arr;
 #print $tag_seq{$arr[0]};
close(FILE1);

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

# Read the original database files
open(DATA2, "refseq_all.fasta") || die "Can't open file";
#open(DATA2, "uniprot_trembl.fasta") || die "Can't open file";

my $name="FOE";
$seq="";
my $domain;

while (<DATA2>)
{
 if($_ =~ m/^>(\S+)/) # Extract the name till there is no space in the fasta
     tag line
 {
  if(defined($useful{$name}))
  {
   for($domain = $useful{$name}; $domain>0;$domain--) # Iterate the loop till
       the maximum number of domain name found in the previous block
   {
    #check if the name read exist in the hash storing tag and sequence in the
        previous block
    #if yes then check whether the corresponding sequence matches and extract
        the 7 residues on either sides
    if(defined($tag_seq{"${name}Domain_$domain"}))# For refseq database;
   #if(defined($tag_seq{"${name}|Domain_$domain"}))# For trembl database
    {
   # The reg exp to extract 7 residues on either side of the sequence matched
```

```
      $seq=~ m/(\w{0,7})($tag_seq{"${name}Domain_$domain"})(\w{0,7})/; # For
          refseq database
      #$seq=~ m/(\w{0,7})($tag_seq{"${name}\|Domain_$domain"})(\w{0,7})/; #
          For trembl database
      print ">${name}Domain_$domain\n"; # For refseq database
   #  print ">${name}|Domain_$domain\n"; # For trembl database
      print "$1$2$3\n";
    }
   }
  }
  $name=$1; # Store the name of the squences from the original database
  $seq="";
 }
 else
 {
  chomp($_);
  $seq.=$_; # Store sequence from the original database
 }
 #print $1,"\n";
}
close(DATA2);
```

A.3 Script to generate the mutant sequences

```
#!/usr/bin/perl -w

use strict;

#This script was used to generate single point mutation in a sequence with all
    the other 19 amino acids.
#The highest scoring sequence from the previous generation.
my td3a=split(//,"SVICEALSDALKVPKKMIDPTEAFSDYGLDSITGVNV
               AQTISSVLNVDLKTTALFDYVCIDQLARYV");

my @amino=split(//,"ARNDCEQGHILKMFPSTWYV"); #Read the 20 amino acids
my $i;
my $j;
my @temp;
my $count=0;

for($i=0; $i<=$#td3a; $i++) #Iterate every position in the original sequence
{
@temp=@td3a;
 for($j=0;$j<=$#amino;$j++) #Iterate every amino acid in the array
 {
  if($td3a[$i] ne $amino[$j]) #Check if the amino acid is in the position or
      not if not then replace it with the amino acids from the amino acid array
  {
```

```
  $count++;
  $temp[$i]=$amino[$j];
  print ">$count\n";
  print @temp,"\n";
  }
 }
undef @temp;
}
```

A.4 Scripts to convert GAFF parameters into Gromacs format

These scripts were written during the course of the PhD and were used to convert GAFF parameters into Gromacs format in order to integrate the values into the AMBER 99sb-ILDB forcefield parameter databases.

A.4.1 Script to convert atom type parameters from GAFF to Gromacs format

```
#!/usr/bin/perl

# This script was used to convert the format of atom name and mass from GAFF
    to Gromacs format (atomtypes.atp).
# The input required was a file containing the list of atoms with mass from
    GAFF in Amber suite.
use strict;
use warnings;

if ($#ARGV < 0) {
    print STDERR "Usage: $0 <file.txt>\n";
    exit -1;
}

open(FILE, "$ARGV[0]") || die "Can't open file";

while(<FILE>)
{
    chomp($_);
    my $atomname=substr($_,0,2);
    my $atommass=substr($_,3,5);
    my $atommassfloat=sprintf("%9.5f",$atommass);
    my $description=substr($_,36);
    print $atomname,"          ",$atommassfloat,"   ;",$description, "\n";
}

close(FILE);
```

A.4.2 Script to convert bond length parameters from GAFF to Gromacs format

```
#!/usr/bin/perl

# This script was used to convert the format of bond parameters from GAFF to
    Gromacs format (ffbonded.itp).
# The input required was a file containing the list of bond parameters from
    GAFF in Amber suite.
use strict;
use warnings;

if ($#ARGV < 0) {
    print STDERR "Usage: $0 <file.txt>\n";
    exit -1;
}

open(FILE, "$ARGV[0]") || die "Can't open file";

while(<FILE>)
{
    chomp($_);
    my $bondatom1=substr($_,0,2);
    my $bondatom2=substr($_,3,2);
    my $b0=substr($_,16,6);
    my $b0nm=sprintf("%6.5f",$b0/10);
    my $kb=substr($_,6,6);
    my $kbkj=sprintf("%8.1f",($kb*4.184*100*2));
    my $description=substr($_,29);
    print " ",$bondatom1," ",$bondatom2,"    1    ",$b0nm," ",$kbkj, " ;
        $description\n";
}

close(FILE);
```

A.4.3 Script to convert bond angle parameters from GAFF to Gromacs format

```
#!/usr/bin/perl

# This script was used to convert the format of bond angle parameters from
    GAFF to Gromacs format (ffbonded.itp).
# The input required was a file containing the list of bond angle parameters
    from GAFF in Amber suite.
use strict;
use warnings;

if ($#ARGV < 0) {
```

```
    print STDERR "Usage: $0 <file.txt>\n";
    exit -1;
}

open(FILE, "$ARGV[0]") || die "Can't open file";

while(<FILE>)
{
    chomp($_);
    my $angleatom1=substr($_,0,2);
    my $angleatom2=substr($_,3,2);
    my $angleatom3=substr($_,6,2);
    my $th0=substr($_,22,7);
    my $cth=substr($_,10,6);
    my $cthkj=sprintf("%8.3f",($cth*4.184*2));
    my $description=substr($_,30);
    print $angleatom1," ",$angleatom2," ",$angleatom3,"    1   ",$th0,"
        ",$cthkj," ; $description\n";
}

close(FILE);
```

A.4.4 Script to convert dihedral angle parameters from GAFF to Gromacs format

```
#!/usr/bin/perl

# This script was used to convert the format of dihedral angle parameters from
    GAFF to Gromacs format (ffbonded.itp).
# The input required was a file containing the list of dihedral angle
    parameters from GAFF in Amber suite.
use strict;
use warnings;

if ($#ARGV < 0) {
    print STDERR "Usage: $0 <file.txt>\n";
    exit -1;
}

open(FILE, "$ARGV[0]") || die "Can't open file";

while(<FILE>)
{
    chomp($_);
    my $dihedralatom1=substr($_,0,2);
    my $dihedralatom2=substr($_,3,2);
    my $dihedralatom3=substr($_,6,2);
    my $dihedralatom4=substr($_,9,2);
    my $IDIVF=substr($_,14,1);
```

```
    my $PK=substr($_,18,6);
    my $kth=sprintf("%8.5f",($PK/$IDIVF*4.184));
    my $phase=sprintf("%5.1f",substr($_,31,7));
    my $PN=substr($_,49,1);
    my $description=substr($_,60);
    print " ",$dihedralatom1," ",$dihedralatom2," ",$dihedralatom3,"
        ",$dihedralatom4," 9 ",$phase," ",$kth," ",$PN," ; ",$description,"\n";
}

close(FILE);
```

A.4.5 Script to convert improper angle parameters from GAFF to Gromacs format

```
#!/usr/bin/perl

# This script was used to convert the format of improper angle parameters from
    GAFF to Gromacs format (ffbonded.itp).
# The input required was a file containing the list of improper angle
    parameters from GAFF in Amber suite.
use strict;
use warnings;

if ($#ARGV < 0) {
    print STDERR "Usage: $0 <file.txt>\n";
    exit -1;
}

open(FILE, "$ARGV[0]") || die "Can't open file";

while(<FILE>)
{
    chomp($_);
    my $dihedralatom1=substr($_,0,2);
    my $dihedralatom2=substr($_,3,2);
    my $dihedralatom3=substr($_,6,2);
    my $dihedralatom4=substr($_,9,2);
    #my $IDIVF=substr($_,14,1);
    my $PK=substr($_,20,4);
    my $kth=sprintf("%8.5f",($PK*4.184));
    my $phase=sprintf("%5.2f",substr($_,33,4));
    my $PN=substr($_,47,1);
    my $description=substr($_,60);
    print $dihedralatom1," ",$dihedralatom2," ",$dihedralatom3,"
        ",$dihedralatom4," 4 ",$phase,"  ",$kth,"    ",$PN,"   ;
        ",$description,"\n";
}

close(FILE);
```

A.4.6 Script to convert nonbonded parameters from GAFF to Gromacs format

```
#!/usr/bin/perl

# This script was used to convert the format of non bonded parameters from
    GAFF to Gromacs format (ffnonbonded.itp).
# The input required was a file containing the list of non bonded parameters
    from GAFF in Amber suite.
use strict;
use warnings;

if ($#ARGV < 0) {
    print STDERR "Usage: $0 <file.txt>\n";
    exit -1;
}

open(FILE, "$ARGV[0]") || die "Can't open file";

while(<FILE>)
{

    chomp($_);
    my $atomtype = substr($_,2,2);
    my $sigma = sprintf("%.5e",(substr($_,13,6) * ((2**(5/6))/10)));
    my $epsilon = sprintf("%.5e",(substr($_,22,6) * 4.184));
    my $description=substr($_,41);
    if($atomtype=~ m/^c/ and not $atomtype=~ m/cl/)
    {
     print $atomtype,"       6     12.01   0.0000 A  ",$sigma," ",$epsilon," ;
         $description\n";
    }
    elsif($atomtype=~ m/^h/)
    {
     print $atomtype,"       1      1.008  0.0000 A  ",$sigma," ",$epsilon," ;
         $description\n";
    }
    elsif($atomtype=~ m/^o/)
    {
     print $atomtype,"       8     16.00   0.0000 A  ",$sigma," ",$epsilon," ;
         $description\n";
    }
    elsif($atomtype=~ m/^n/)
    {
     print $atomtype,"       7     14.01   0.0000 A  ",$sigma," ",$epsilon," ;
         $description\n";
    }
    elsif($atomtype=~ m/^s/)
    {
```

```
    print $atomtype,"      16     32.06   0.0000 A  ",$sigma," ",$epsilon," ;
        $description\n";
    }
    elsif($atomtype=~ m/^p/)
    {
    print $atomtype,"      15     30.97   0.0000 A  ",$sigma," ",$epsilon," ;
        $description\n";
    }
    elsif($atomtype=~ m/^f/)
    {
    print $atomtype,"       9     19.00   0.0000 A  ",$sigma," ",$epsilon," ;
        $description\n";
    }
    elsif($atomtype=~ m/^cl/)
    {
    print $atomtype,"      17     35.45   0.0000 A  ",$sigma," ",$epsilon," ;
        $description\n";
    }
    elsif($atomtype=~ m/^br/)
    {
    print $atomtype,"      35     79.90   0.0000 A  ",$sigma," ",$epsilon," ;
        $description\n";
    }
    elsif($atomtype=~ m/^i/)
    {
    print $atomtype,"      53    126.9    0.0000 A  ",$sigma," ",$epsilon," ;
        $description\n";
    }
}

close(FILE);
```

A.5 Script to calculate RMSD using Matt program

```
#!/usr/bin/perl -w

$protein1=$ARGV[0];
$multiplepdb=$ARGV[1];
open(PDB, "<$multiplepdb") or die "Can not open file";
open(OUT, ">protein2.pdb") or die "Can not create file";
open(RMSD, ">RMSD.txt") or die "Can not create file";
$i=0;
while (<PDB>)
{
 if($_=~m/^ATOM/)
 {
  #print $_;
  print OUT $_;
  next;
```

```
 }
 elsif ($_=~m/^END/)
 {
  system ("matt $protein1 protein2.pdb");
  open (MattOutput, "<MattAlignment.txt");
  while(<MattOutput>)
  {
   if ($_ =~m/^Core RMSD:\s(\d+.\d+)/)
   {
        print RMSD "$i,$1\n";
   }
  }
 system ("rm protein2.pdb");
 open(OUT, ">protein2.pdb") or die "Can not create file";
 $i++;
 }
}
```

Chapter B

Appendix II

B.1 List of new residues with charges added to the AMBER99SB-ILDN forcefield in Gromacs format

B.1.1 Charges for phosphopantetheine

Residue name: SPT

Atoms	Atomtype	Charge	Number
N	N	-0.41570	1
H	H	0.27190	2
CA	CT	-0.08800	3
HA	H1	0.15590	4
CB	CT	0.03140	5
HB1	H1	0.08540	6
HB2	H1	0.08540	7
OG	OS	-0.43570	8
PD	p5r	1.04280	9
OE	or	-0.73120	10
OH	or	-0.73120	11
OZ	osr	-0.40080	12
CQ	c3r	-0.02100	13
HQ1	h1r	0.10220	14
HQ2	h1r	0.10220	15
CI	c3r	0.12870	16
CIA	c3r	-0.15840	17
HIA1	hcr	0.03550	18
HIA2	hcr	0.03550	19
HIA3	hcr	0.03550	20
CIB	c3r	-0.15840	21
HIB1	hcr	0.03550	22

HIB2	hcr	0.03550	23
HIB3	hcr	0.03550	24
CIG	c3r	-0.00210	25
HIG	h1r	0.16460	26
OID	ohr	-0.58750	27
HID	hor	0.40570	28
CIE	cr	0.38430	29
OIZ	or	-0.49960	30
NH	nr	-0.28870	31
HH	hnr	0.25640	32
CIQ	c3r	0.03840	33
HIQ1	h1r	0.05000	34
HIQ2	h1r	0.05000	35
CK	c3r	-0.17820	36
HK1	hcr	0.07800	37
HK2	hcr	0.07800	38
CKA	cr	0.51920	39
OKB	or	-0.57540	40
NKG	nr	-0.40580	41
HKG	hnr	0.27300	42
CKD	c3r	0.02980	43
HKD1	h1r	0.06100	44
HKD2	h1r	0.06100	45
CKE	c3r	-0.00750	46
HKE1	h1r	0.09690	47
HKE2	h1r	0.09690	48
SKZ	shr	-0.42300	49
HKZ	hsr	0.21670	50
C	C	0.59730	51
O	O	-0.56790	52

B.1.2 Charges for unbranched monic acid attached to phosphopantetheine

Residue name: SPM

Atoms	Atomtype	Charge	Number
N	N	-0.4157	1
H	H	0.2719	2
CA	CT	-0.0880	3
HA	H1	0.1559	4
CB	CT	0.0314	5
HB1	H1	0.0854	6
HB2	H1	0.0854	7
OG	OS	-0.4357	8
PD	p5r	1.0428	9
OE	or	-0.7312	10
OH	or	-0.7312	11

OZ	osr	-0.4008	12
CQ	c3r	-0.0210	13
HQ1	h1r	0.1022	14
HQ2	h1r	0.1022	15
CI	c3r	0.1287	16
CIA	c3r	-0.1584	17
HIA1	hcr	0.0355	18
HIA2	hcr	0.0355	19
HIA3	hcr	0.0355	20
CIB	c3r	-0.1584	21
HIB1	hcr	0.0355	22
HIB2	hcr	0.0355	23
HIB3	hcr	0.0355	24
CIG	c3r	-0.0021	25
HIG	h1r	0.1646	26
OID	ohr	-0.5875	27
HID	hor	0.4057	28
CIE	cr	0.3843	29
OIZ	or	-0.4996	30
NH	nr	-0.2887	31
HH	hnr	0.2564	32
CIQ	c3r	0.0384	33
HIQ1	h1r	0.0500	34
HIQ2	h1r	0.0500	35
CK	c3r	-0.1782	36
HK1	hcr	0.0780	37
HK2	hcr	0.0780	38
CKA	cr	0.6244	39
OKB	or	-0.5521	40
NKG	nr	-0.5468	41
HKG	hnr	0.2730	42
CKD	c3r	-0.1108	43
HKD1	h1r	0.1309	44
HKD2	h1r	0.1309	45
CKE	c3r	0.0796	46
HKE1	h1r	0.0565	47
HKE2	h1r	0.0565	48
SKZ	ssr	-0.278	49
CKH	cr	0.4936	50
OKQ	or	-0.4762	51
CX	c3r	-0.0376	52
HX1	hcr	0.0292	53
HX2	hcr	0.0292	54
CXA	cr	0.5017	55
OXB	or	-0.4804	56
CXG	c3r	-0.0339	57
HXG1	hcr	0.0325	58
HXG2	hcr	0.0325	59
CXD	c3r	0.1369	60

HXD	h1r	0.1344	61
OXE	ohr	-0.6552	62
HXE	hor	0.4331	63
CXZ	c3r	-0.0199	64
HXZ	h1r	0.1196	65
OXH	ohr	-0.6166	66
HXH	hor	0.3962	67
CXQ	c3r	0.1625	68
HXQ	h1r	0.1158	69
OM	ohr	-0.6569	70
HM	hor	0.4266	71
CMA	c2r	0.1042	72
CMB	c3r	-0.1920	73
HMB1	hcr	0.0639	74
HMB2	hcr	0.0639	75
HMB3	hcr	0.0639	76
CMG	cer	-0.2584	77
HMG	har	0.1528	78
CMD	cer	-0.1381	79
HMD	har	0.1190	80
CME	c2r	-0.1661	81
HME	har	0.1142	82
CMZ	c3r	0.0367	83
HMZ	hcr	0.0647	84
CMH	c3r	-0.1807	85
HMH1	hcr	0.0480	86
HMH2	hcr	0.0480	87
HMH3	hcr	0.0480	88
CMQ	c3r	0.2814	89
HMQ	h1r	0.0165	90
ON	ohr	-0.6655	91
HN	hor	0.4035	92
CNA	c3r	-0.2053	93
HNA1	hcr	0.0630	94
HNA2	hcr	0.0630	95
HNA3	hcr	0.0630	96
C	C	0.5973	97
O	O	-0.5679	98

B.1.3 Charges for fully saturated carbon chain attached to phosphopantetheine

Residue name: SPD

Atoms	Atomtype	Charge	Number
N	N	-0.4157	1
H	H	0.2719	2
CA	CT	-0.0880	3

HA	H1	0.1559	4
CB	CT	0.0314	5
HB1	H1	0.0854	6
HB2	H1	0.0854	7
OG	OS	-0.4357	8
PD	p5r	1.0428	9
OE	or	-0.7312	10
OH	or	-0.7312	11
OZ	osr	-0.4008	12
CQ	c3r	-0.0210	13
HQ1	h1r	0.1022	14
HQ2	h1r	0.1022	15
CI	c3r	0.1287	16
CIA	c3r	-0.1584	17
HIA1	hcr	0.0355	18
HIA2	hcr	0.0355	19
HIA3	hcr	0.0355	20
CIB	c3r	-0.1584	21
HIB1	hcr	0.0355	22
HIB2	hcr	0.0355	23
HIB3	hcr	0.0355	24
CIG	c3r	-0.0021	25
HIG	h1r	0.1646	26
OID	ohr	-0.5875	27
HID	hor	0.4057	28
CIE	cr	0.3843	29
OIZ	or	-0.4996	30
NH	nr	-0.2887	31
HH	hnr	0.2564	32
CIQ	c3r	0.0384	33
HIQ1	h1r	0.0500	34
HIQ2	h1r	0.0500	35
CK	c3r	-0.1782	36
HK1	hcr	0.0746	37
HK2	hcr	0.0746	38
CKA	cr	0.5692	39
OKB	or	-0.5754	40
NKG	nr	-0.4058	41
HKG	hnr	0.2730	42
CKD	c3r	0.0298	43
HKD1	h1r	0.0610	44
HKD2	h1r	0.0610	45
CKE	c3r	-0.0075	46
HKE1	h1r	0.0969	47
HKE2	h1r	0.0969	48
SKZ	ssr	-0.4230	49
CKH	c3r	-0.1364	50
HKH1	hcr	0.0980	51
HKH2	hcr	0.0980	52

CX	c3r	0.0624	53
HX1	hcr	0.0311	54
HX2	hcr	0.0311	55
CXA	c3r	-0.0179	56
HXA1	hcr	-0.0035	57
HXA2	hcr	-0.0035	58
CXG	c3r	0.0472	59
HXG1	hcr	-0.0190	60
HXG2	hcr	-0.0190	61
CXD	c3r	0.0036	62
HXD1	hcr	-0.0024	63
HXD2	hcr	-0.0024	64
CXZ	c3r	0.0074	65
HXZ1	hcr	-0.0008	66
HXZ2	hcr	-0.0008	67
CXQ	c3r	0.0074	68
HXQ1	hcr	-0.0008	69
HXQ2	hcr	-0.0008	70
CMA	c3r	0.0036	71
HMA1	hcr	-0.0024	72
HMA2	hcr	-0.0024	73
CMG	c3r	-0.0198	74
HMG1	hcr	0.0053	75
HMG2	hcr	0.0053	76
CMD	c3r	0.0074	77
HMD1	hcr	-0.0008	78
HMD2	hcr	-0.0008	79
CME	c3r	-0.0091	80
HME1	hcr	-0.0008	81
HME2	hcr	-0.0008	82
CMZ	c3r	0.0190	83
HMZ1	hcr	-0.0013	84
HMZ2	hcr	-0.0013	85
CMQ	c3r	0.0329	86
HMQ1	hcr	-0.0053	87
HMQ2	hcr	-0.0053	88
CNA	c3r	-0.0579	89
HNA1	hcr	0.0097	90
HNA2	hcr	0.0097	91
HNA3	hcr	0.0097	92
C	C	0.5973	93
O	O	-0.5679	94

B.1.4 Charges for C14 mupirocin intermediate attached to phosphopantetheine

Residue name: SPB

Atoms	Atomtype	Charge	Number

```
-----------------------------------
    N    N          -0.4157    1
    H    H           0.2719    2
   CA    CT         -0.0880    3
   HA    H1          0.1559    4
   CB    CT          0.0314    5
  HB1    H1          0.0854    6
  HB2    H1          0.0854    7
   OG    OS         -0.4357    8
   PD    p5r         1.0428    9
   OE    or         -0.7312   10
   OH    or         -0.7312   11
   OZ    osr        -0.4008   12
   CQ    c3r        -0.0210   13
  HQ1    h1r         0.1022   14
  HQ2    h1r         0.1022   15
   CI    c3r         0.1287   16
  CIA    c3r        -0.1584   17
 HIA1    hcr         0.0355   18
 HIA2    hcr         0.0355   19
 HIA3    hcr         0.0355   20
  CIB    c3r        -0.1584   21
 HIB1    hcr         0.0355   22
 HIB2    hcr         0.0355   23
 HIB3    hcr         0.0355   24
  CIG    c3r        -0.0021   25
  HIG    h1r         0.1646   26
  OID    ohr        -0.5875   27
  HID    hor         0.4057   28
  CIE    cr          0.3843   29
  OIZ    or         -0.4996   30
   NH    nr         -0.2887   31
   HH    hnr         0.2564   32
  CIQ    c3r         0.0384   33
 HIQ1    h1r         0.0500   34
 HIQ2    h1r         0.0500   35
   CK    c3r        -0.1134   36
  HK1    hcr         0.0353   37
  HK2    hcr         0.0353   38
  CKA    cr          0.6244   39
  OKB    or         -0.5521   40
  NKG    nr         -0.5991   41
  HKG    hnr         0.3539   42
  CKD    c3r         0.0208   43
 HKD1    h1r         0.0485   44
 HKD2    h1r         0.0485   45
  CKE    c3r         0.0518   46
 HKE1    h1r         0.0613   47
 HKE2    h1r         0.0613   48
```

SKZ	ssr	-0.2812	49
CXA	cr	0.5541	50
OXB	or	-0.4720	51
CXG	c3r	-0.1453	52
HXG1	hcr	0.0647	53
HXG2	hcr	0.0647	54
CXD	c3r	0.2284	55
HXD	h1r	0.1019	56
OXE	ohr	-0.5954	57
HXE	hor	0.4071	58
CXZ	c3r	0.0214	59
HXZ	h1r	0.1137	60
OXH	ohr	-0.6254	61
HXH	hor	0.3978	62
CXQ	c3r	0.1110	63
HXQ	h1r	0.1173	64
OM	ohr	-0.6240	65
HM	hor	0.3929	66
CMA	c2r	0.1146	67
CMB	c3r	-0.1914	68
HMB1	hcr	0.0624	69
HMB2	hcr	0.0624	70
HMB3	hcr	0.0624	71
CMG	cer	-0.2579	72
HMG	har	0.1484	73
CMD	cer	-0.1291	74
HMD	har	0.1234	75
CME	c2r	-0.1780	76
HME	har	0.1196	77
CMZ	c3r	0.0148	78
HMZ	hcr	0.0736	79
CMH	c3r	-0.1712	80
HMH1	hcr	0.0441	81
HMH2	hcr	0.0441	82
HMH3	hcr	0.0441	83
CMQ	c3r	0.3047	84
HMQ	h1r	0.0121	85
ON	ohr	-0.6719	86
HN	hor	0.4045	87
CNA	c3r	-0.2237	88
HNA1	hcr	0.0668	89
HNA2	hcr	0.0668	90
HNA3	hcr	0.0668	91
C	C	0.5973	92
O	O	-0.5679	93

B.1.5 Charges for acetyl moiety attached to a cysteine

Residue name: CYA

Atoms	Atomtype	Charge	Number
N	N	-0.4157	1
H	H	0.2719	2
CA	CT	0.0213	3
HA	H1	0.1124	4
CB	CT	-0.1231	5
HB1	H1	0.1112	6
HB2	H1	0.1112	7
SG	ssr	-0.3079	8
CD	cr	0.5660	9
OE	or	-0.4305	10
CZ	c3r	-0.3248	11
HZ1	hcr	0.1262	12
HZ2	hcr	0.1262	13
HZ3	hcr	0.1262	14
C	C	0.5973	15
O	O	-0.5679	16

Chapter C

Appendix III

C.1 Formation and change in cavity volume in PKS ACPs over time

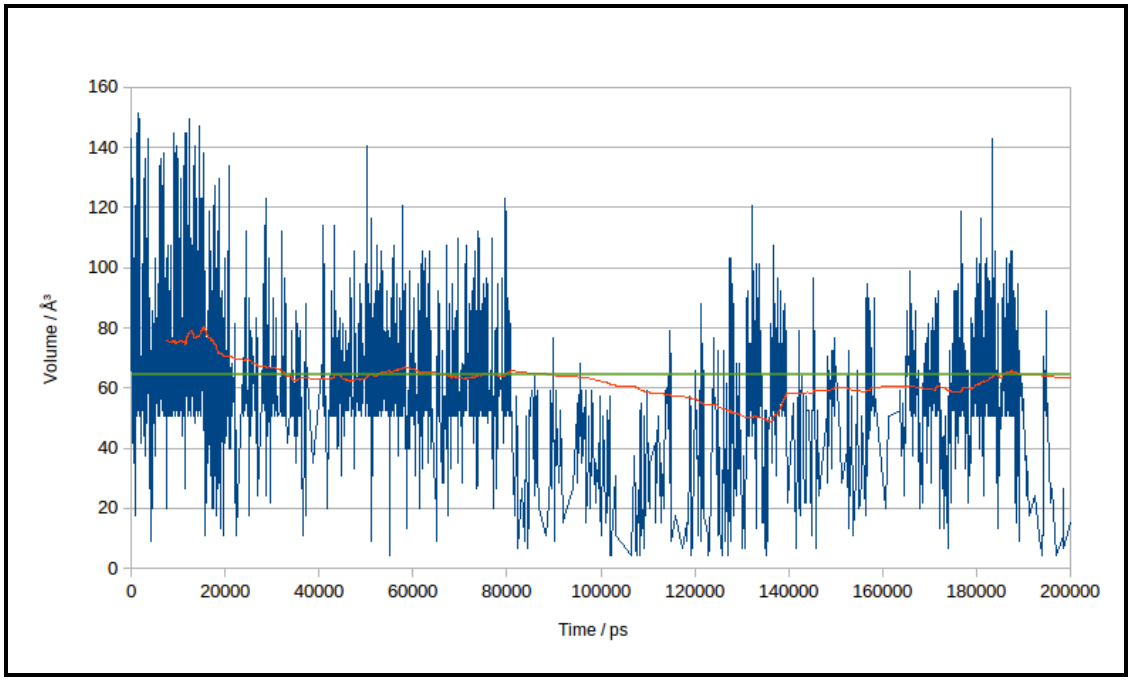

Figure C.1: *Formation and change in cavity volume over time (200 ns) in the apo ACP-mupA3a WT. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

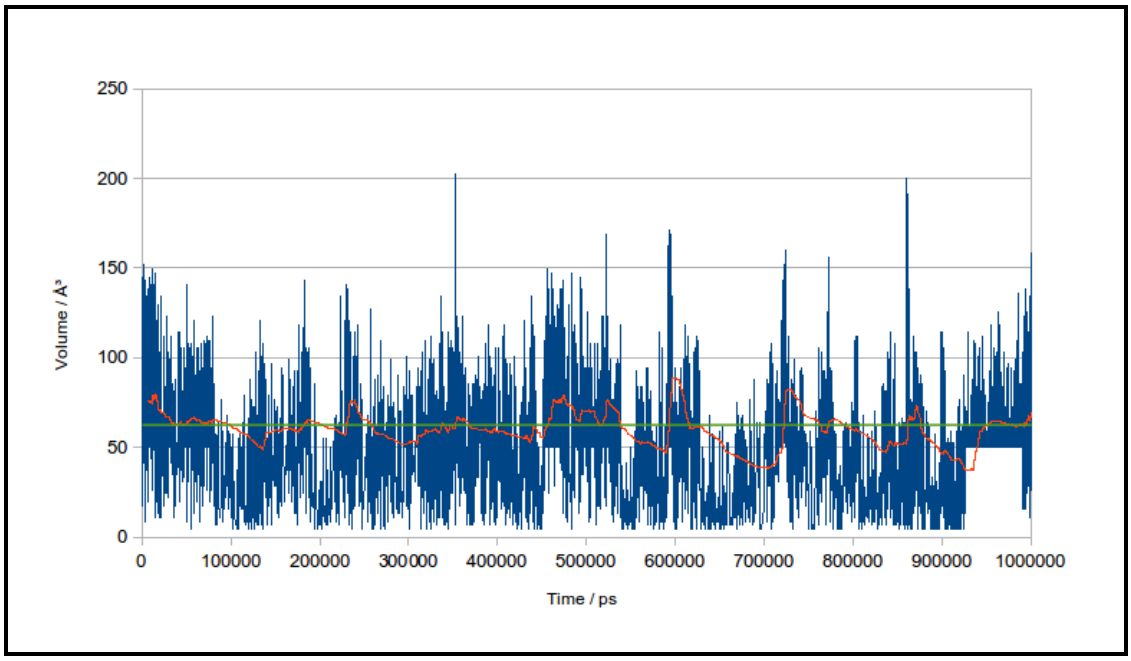

Figure C.2: *Formation and change in cavity volume over time (1 μs) in the apo ACP-mupA3a WT. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

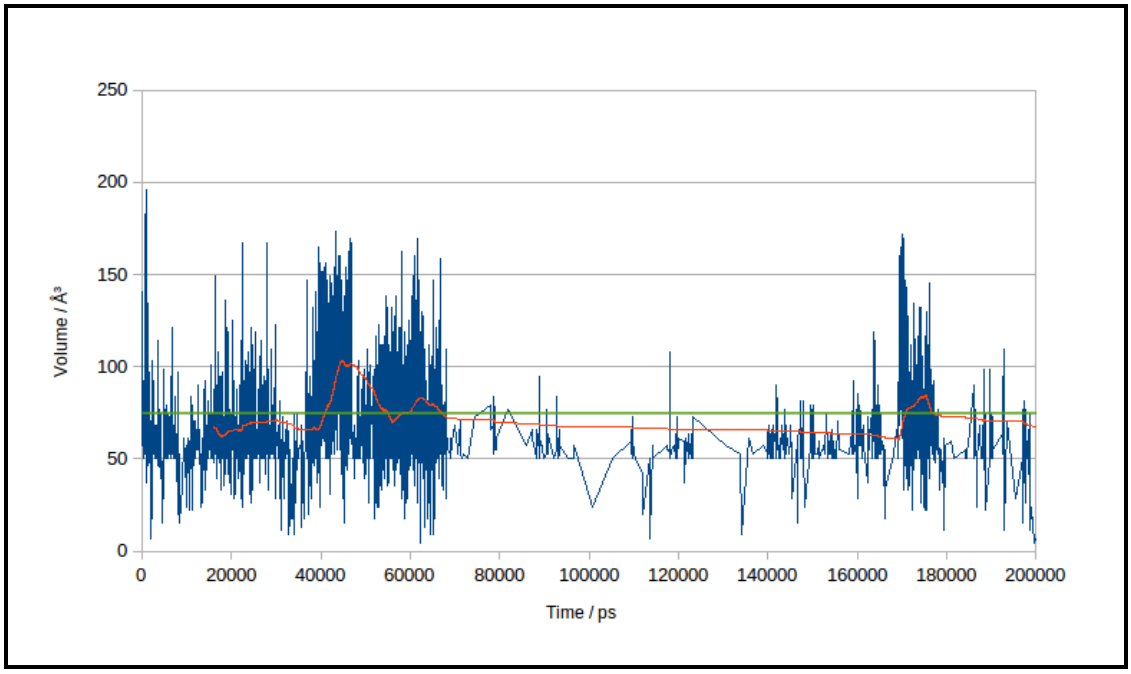

Figure C.3: *Formation and change in cavity volume over time in the apo ACP-mupA3a W44L. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

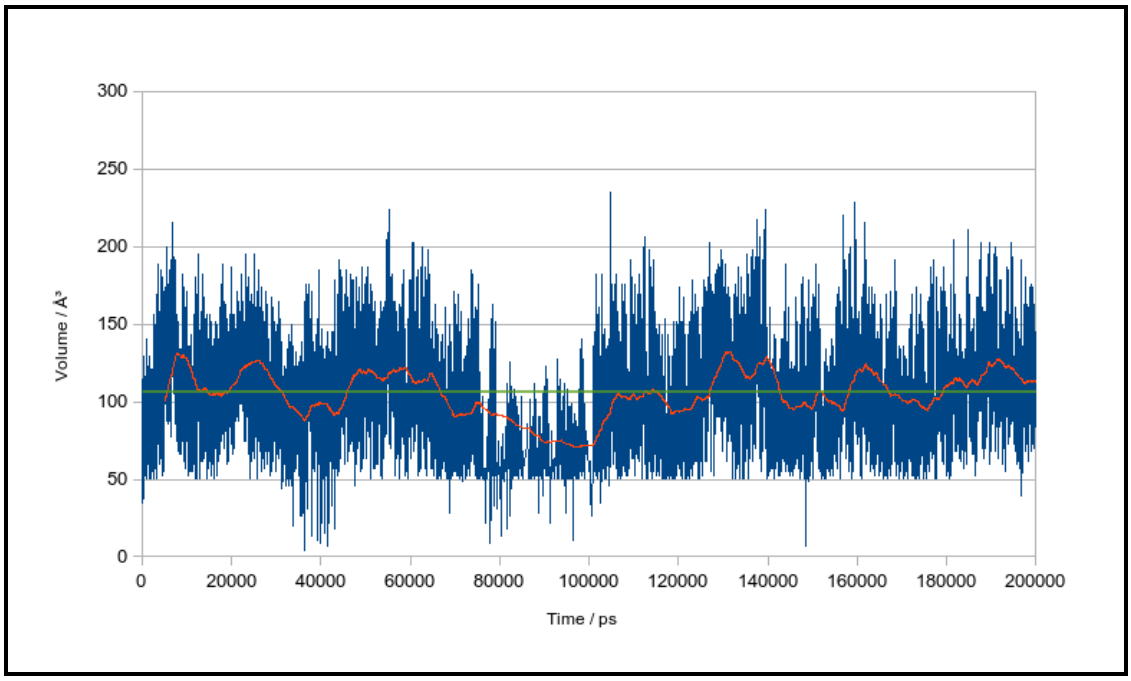

Figure C.4: *Formation and change in cavity volume over time in the holo ACP-mupA3a WT. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

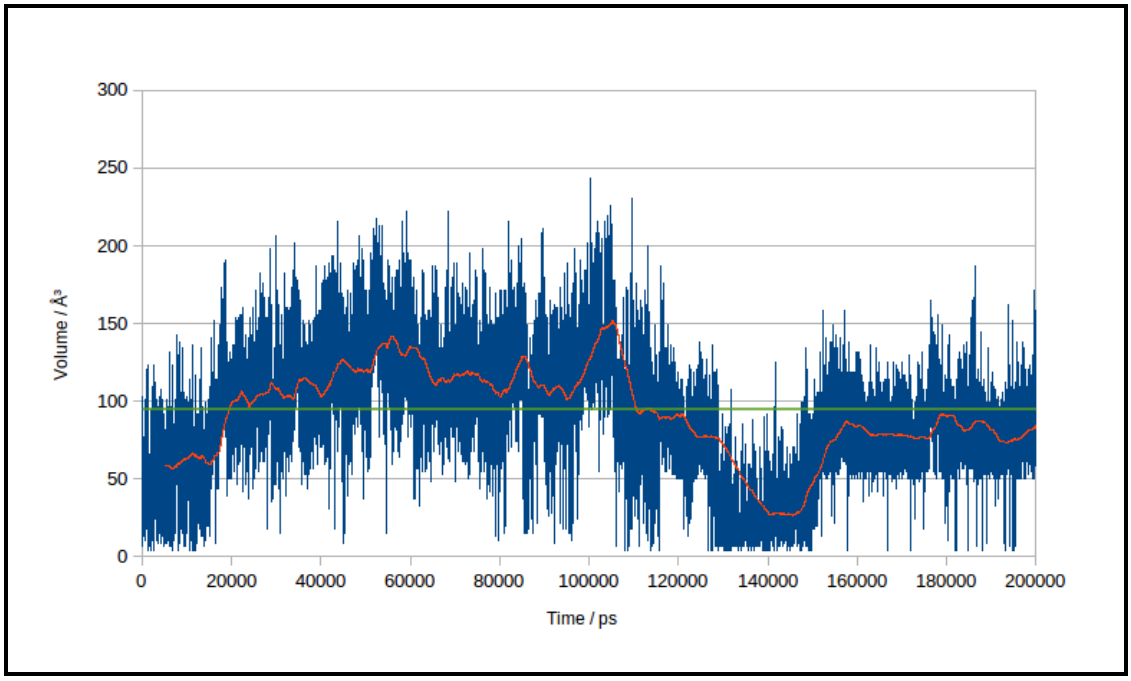

Figure C.5: *Formation and change in cavity volume over time in the holo ACP-mupA3a W44L. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

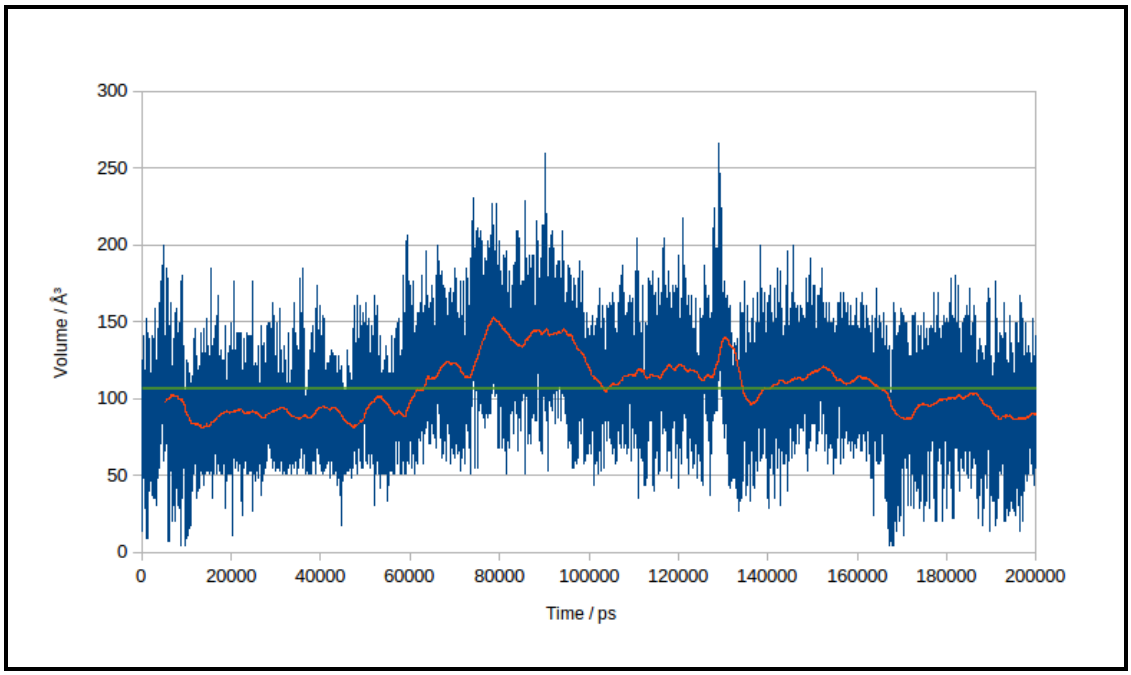

Figure C.6: *Formation and change in cavity volume over time (200ns) in the acyl ACP-mupA3a WT. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

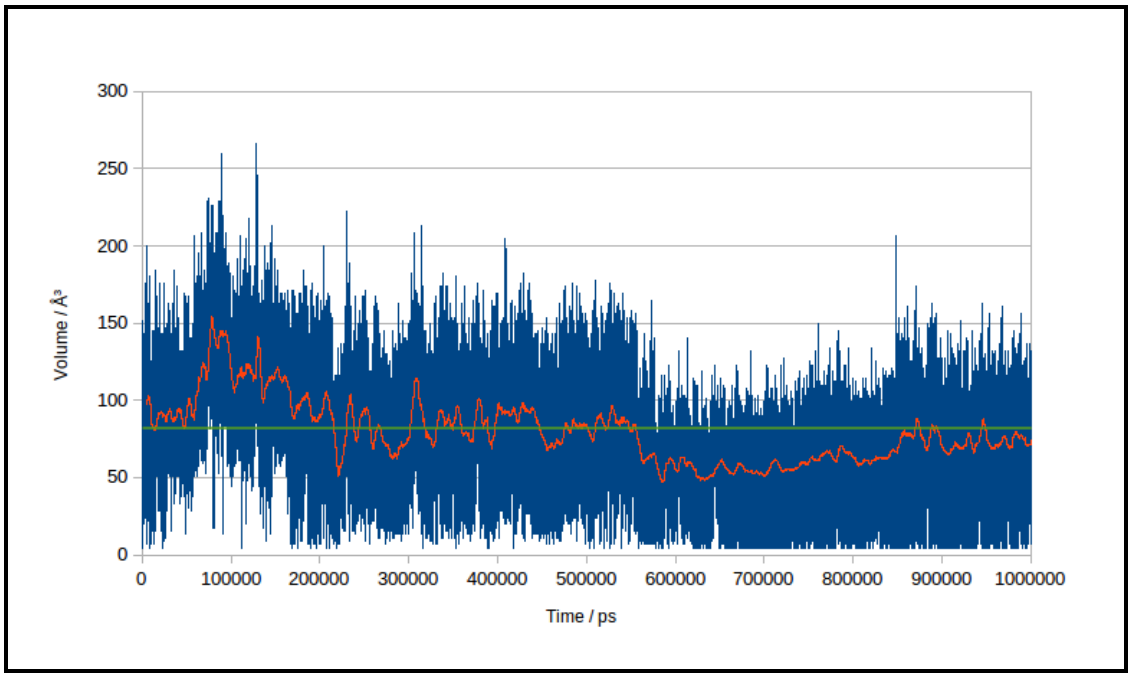

Figure C.7: *Formation and change in cavity volume over time (1 μs) in the acyl ACP-mupA3a WT. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

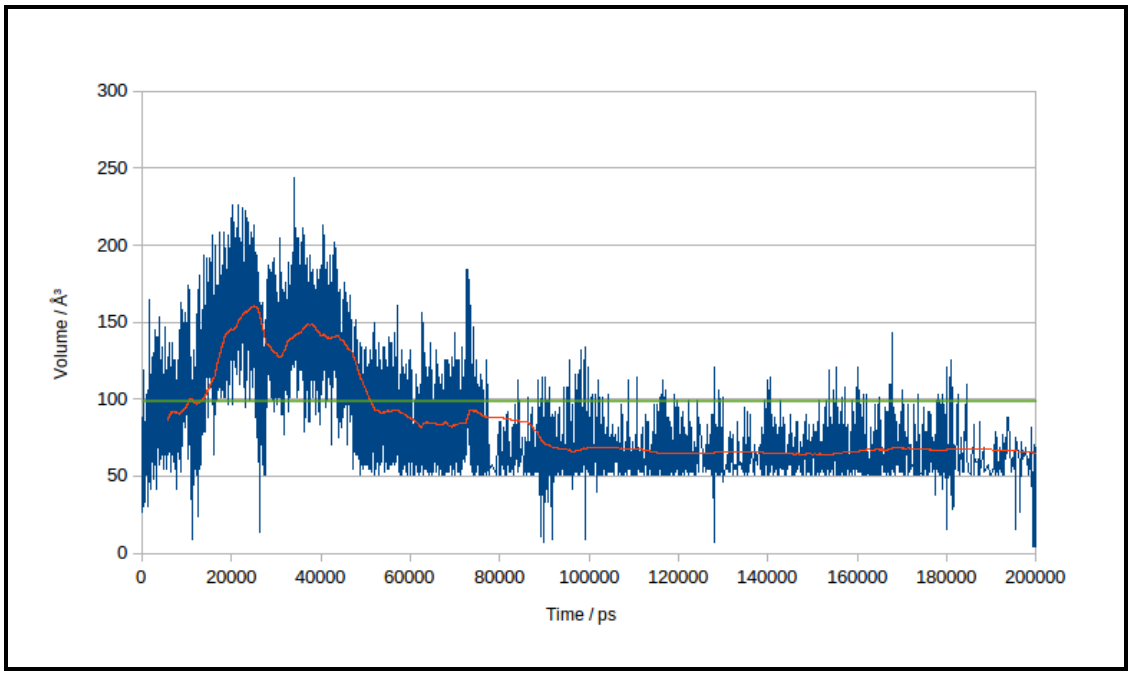

Figure C.8: *Formation and change in cavity volume over time in the acyl ACP-mupA3a W44L. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

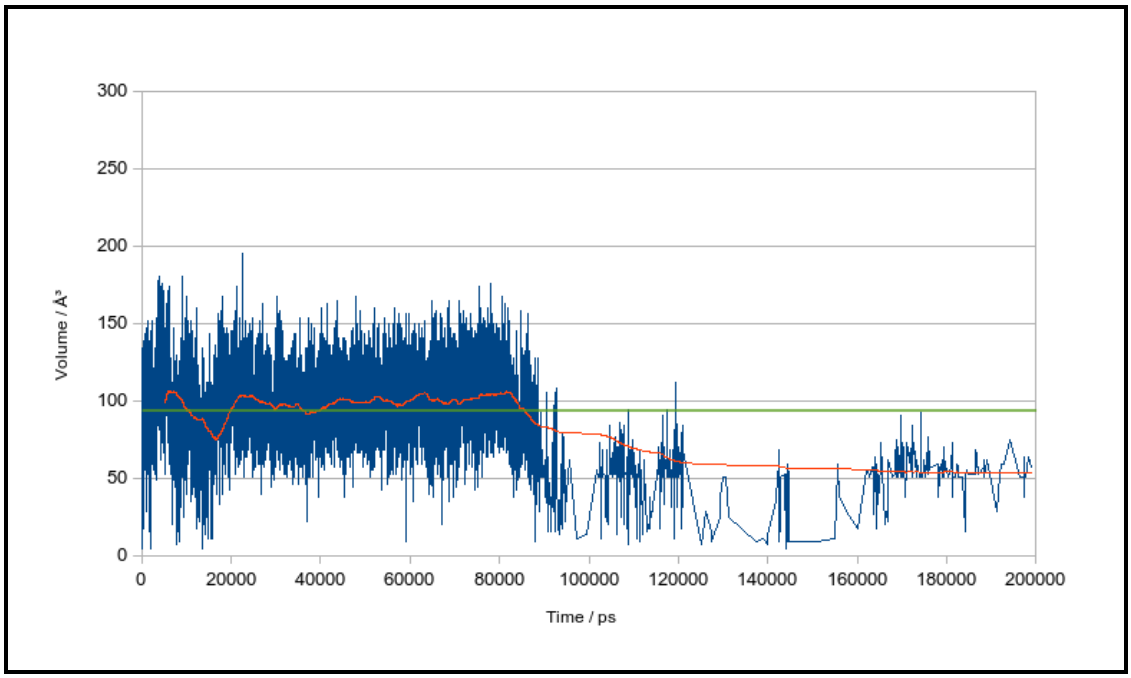

Figure C.9: *Formation and change in cavity volume over time in the acyl 14C ACP-mupA3a. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

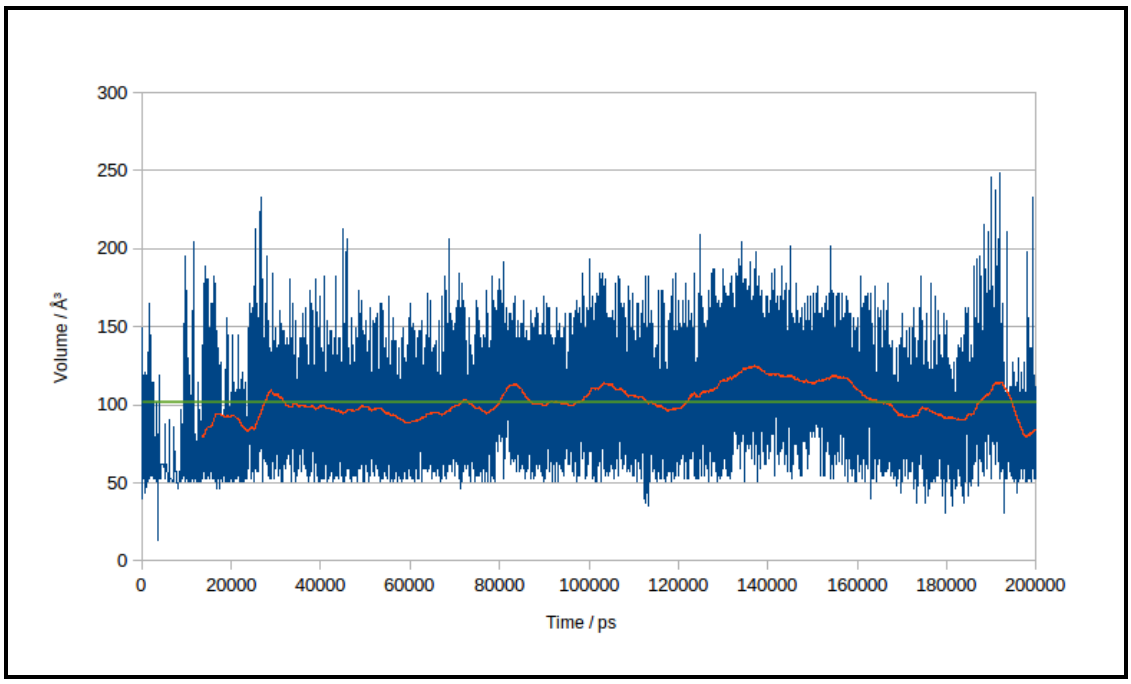

Figure C.10: *Formation and change in cavity volume over time in the acyl ACP-mupA2a. The time frames which had a zero value for the volume were omitted from the plot. Red line represents the running average over 500 frames and green line represents the mean.*

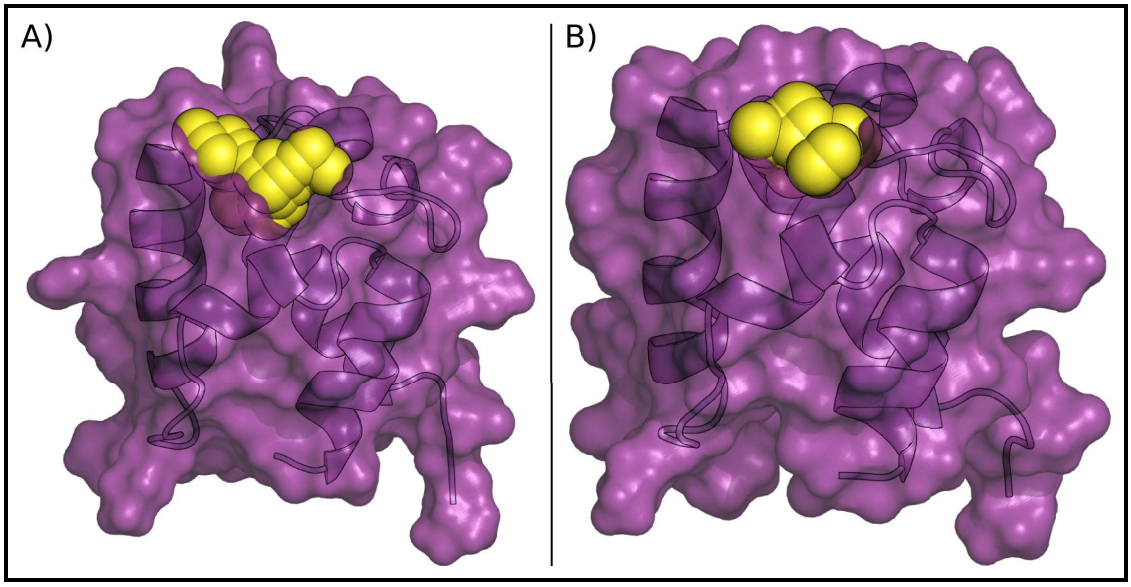

Figure C.11: *Space filled diagram of the largest (A) and the modal (B) cavity volume in the apo ACP-mupA3a WT.*

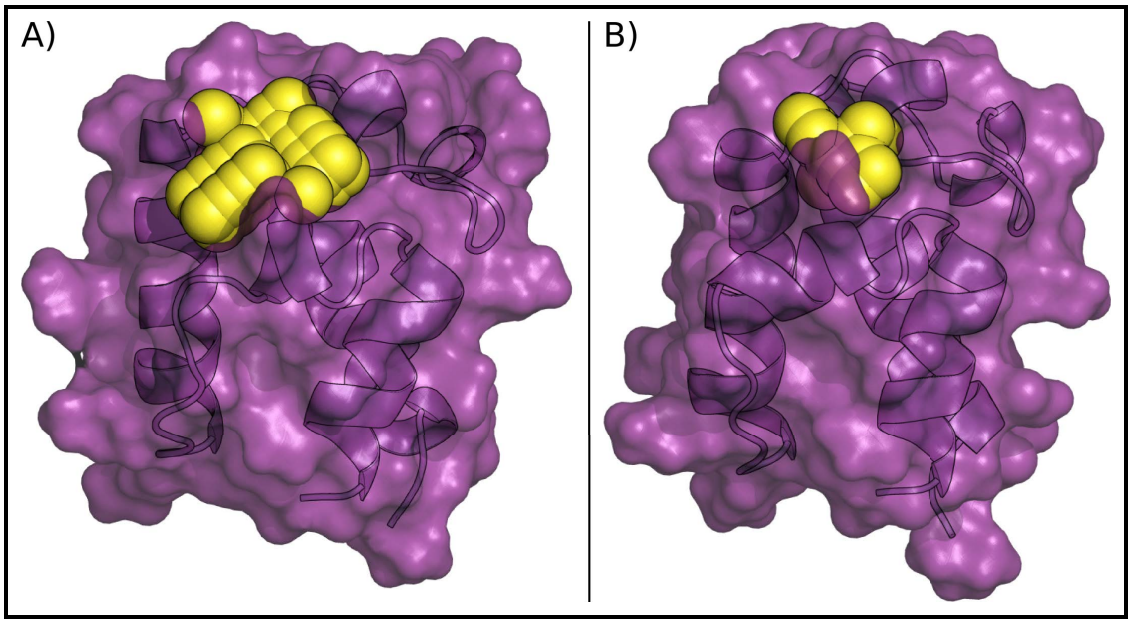

Figure C.12: *Space filled diagram of the largest (A) and the modal (B) cavity volume in the apo ACP-mupA3a W44L.*

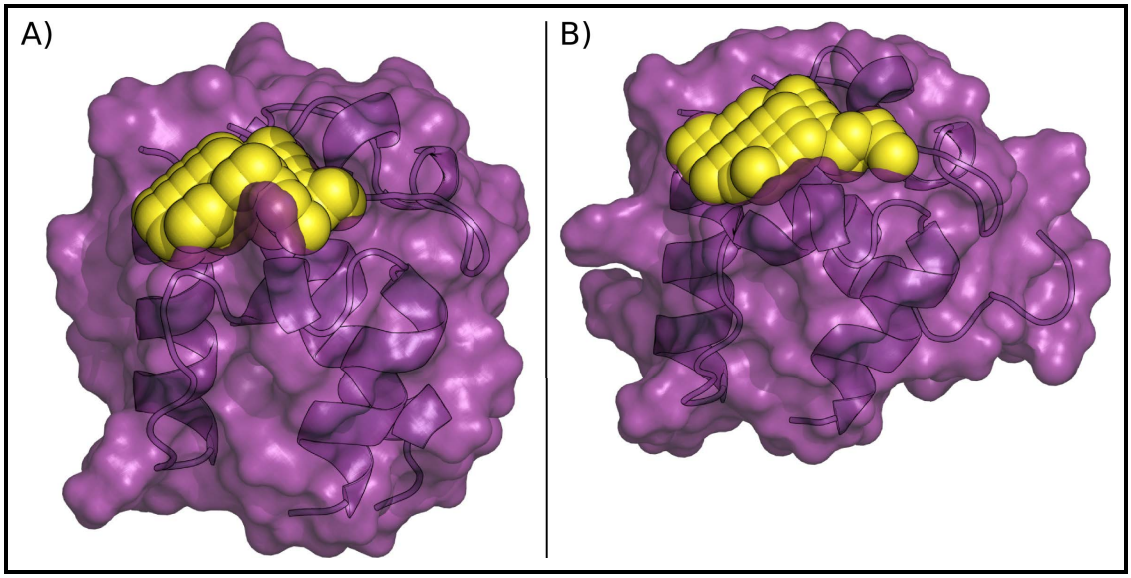

Figure C.13: *Space filled diagram of the largest (A) and the modal (B) cavity volume in the holo ACP-mupA3a WT.*

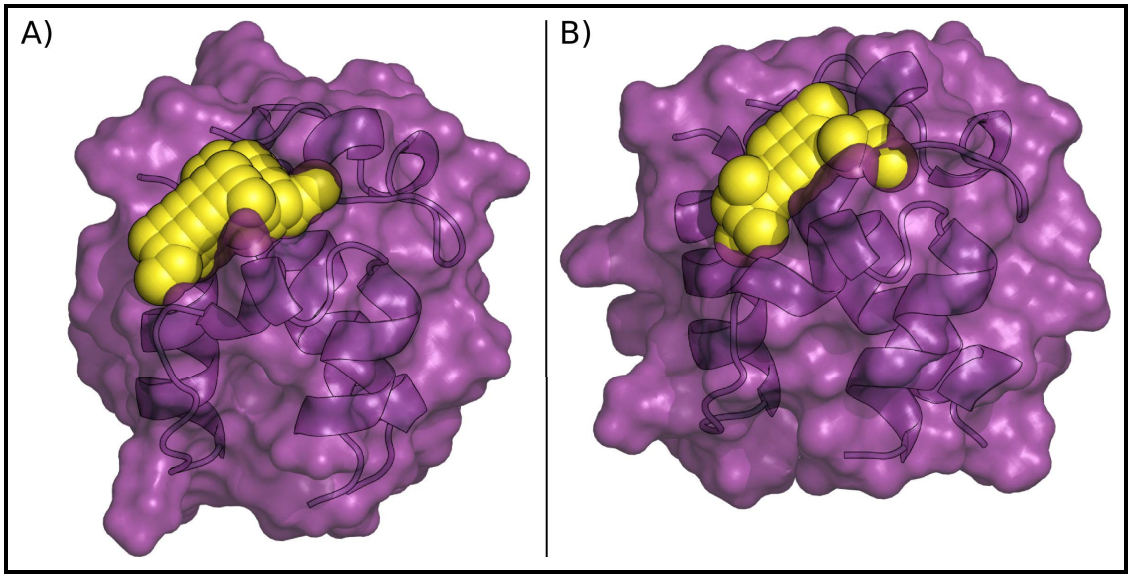

Figure C.14: *Space filled diagram of the largest (A) and the modal (B) cavity volume in the holo ACP-mupA3a W44L.*

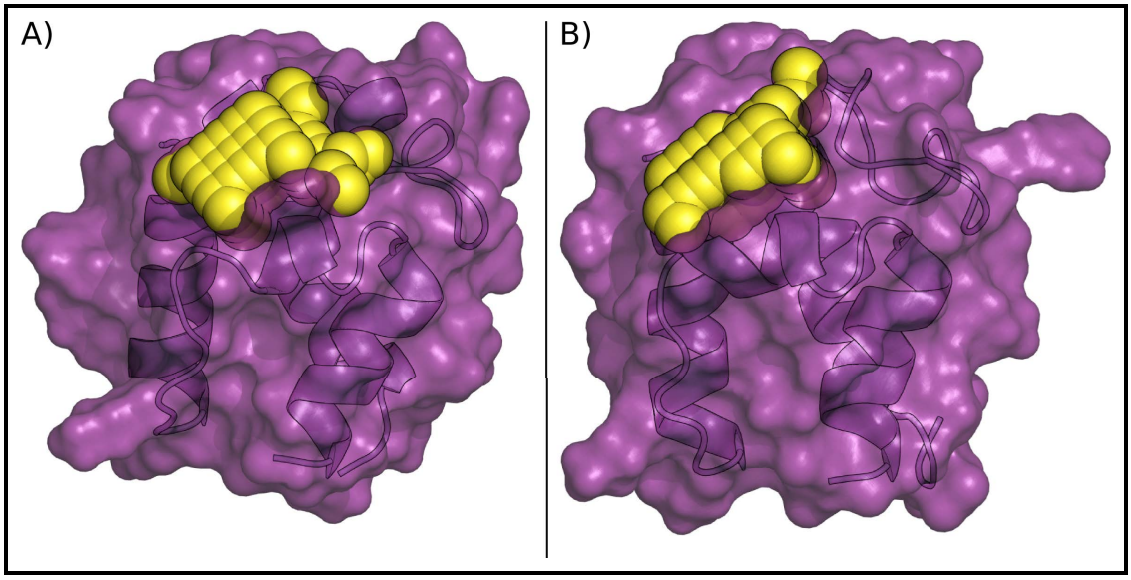

Figure C.15: *Space filled diagram of the largest (A) and the modal (B) cavity volume in the acyl ACP-mupA3a W44L.*

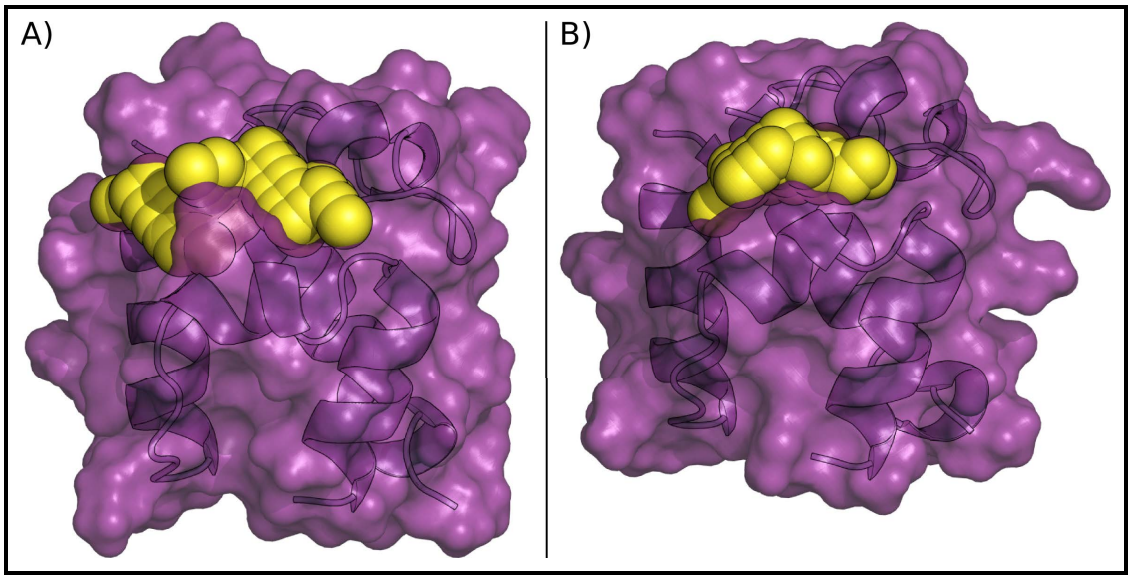

Figure C.16: *Space filled diagram of the largest (A) and the modal (B) cavity volume in the acyl 14C ACP-mupA3a.*

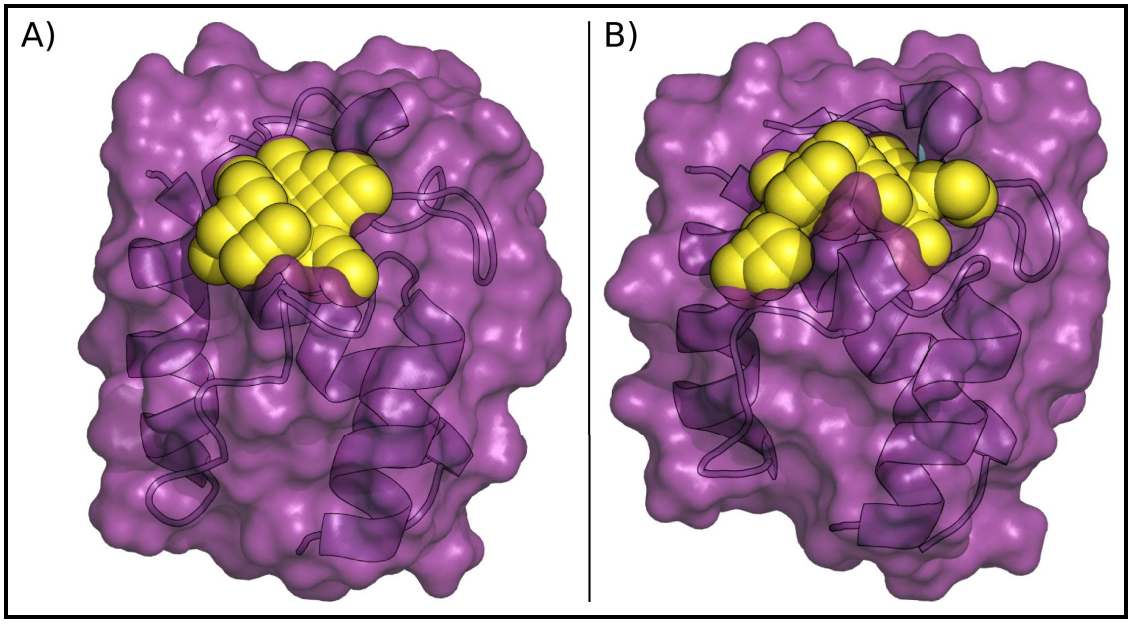

Figure C.17: *Space filled diagram of the largest (A) and the modal (B) cavity volume in the acyl ACP-mupA2a.*

C.1.1 Change in RMSD of PKS ACPs from FAS ACP over time

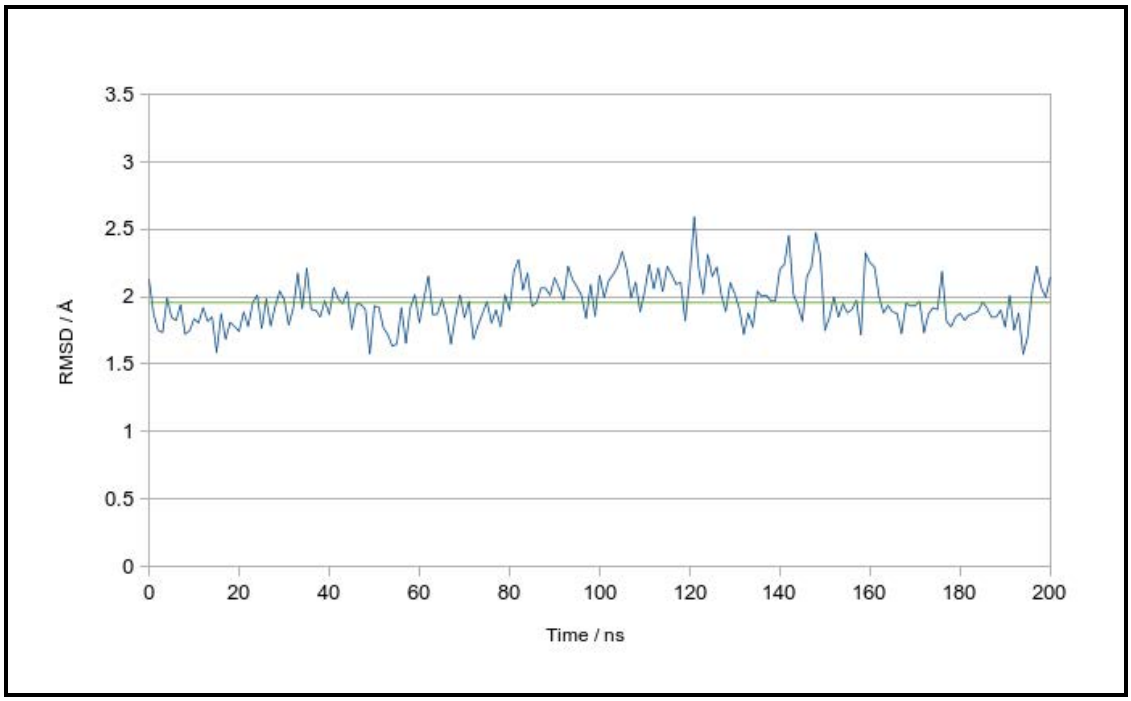

Figure C.18: *RMSD between FAS ACP and apo ACP-mupA3a WT over time (200 ns). Green line represents the mean.*

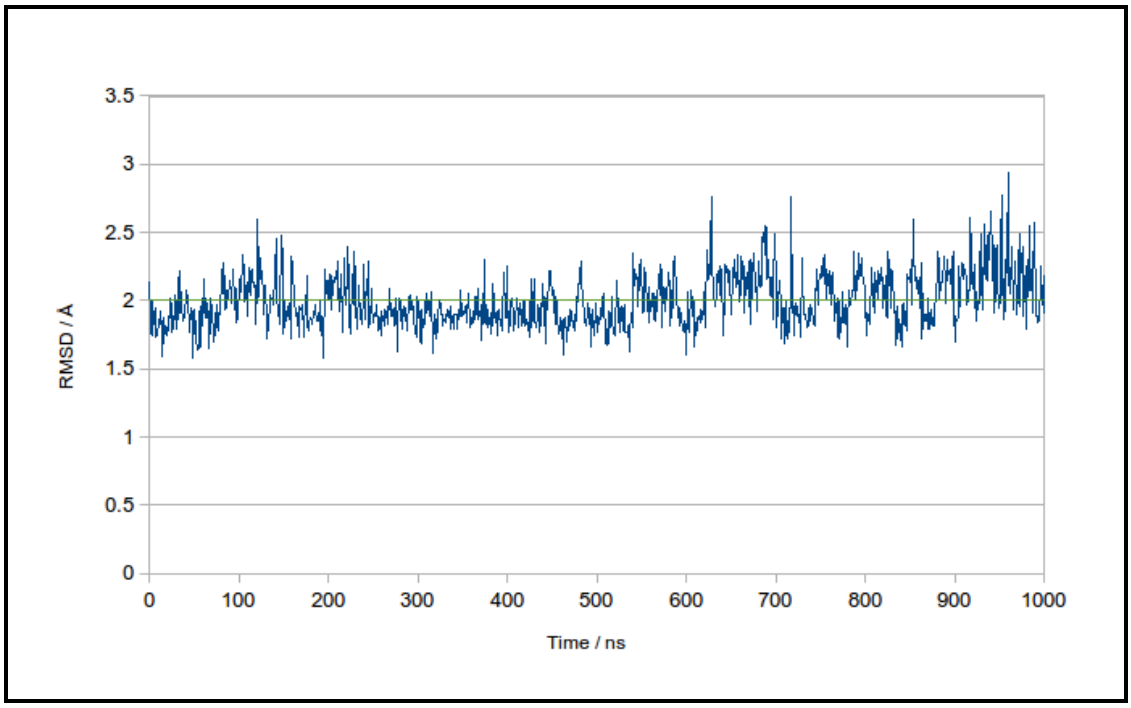

Figure C.19: *RMSD between FAS ACP and apo ACP-mupA3a WT over time (1 µs). Green line represents the mean.*

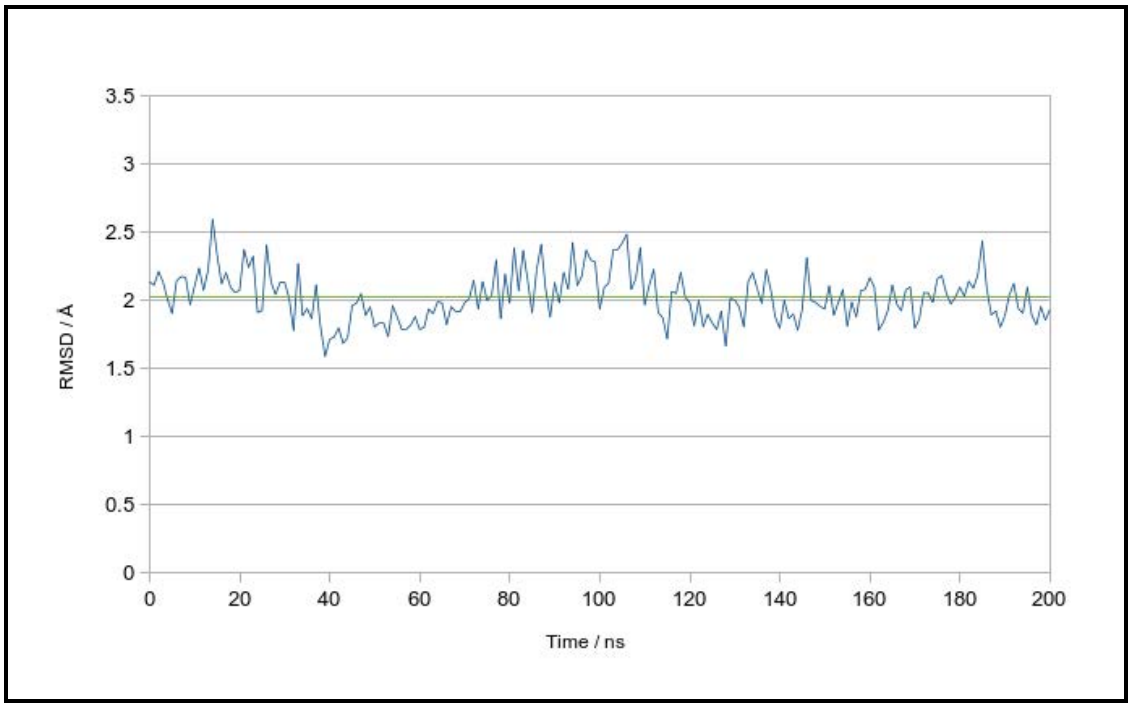

Figure C.20: *RMSD between FAS ACP and apo ACP-mupA3a W44L over time. Green line represents the mean.*

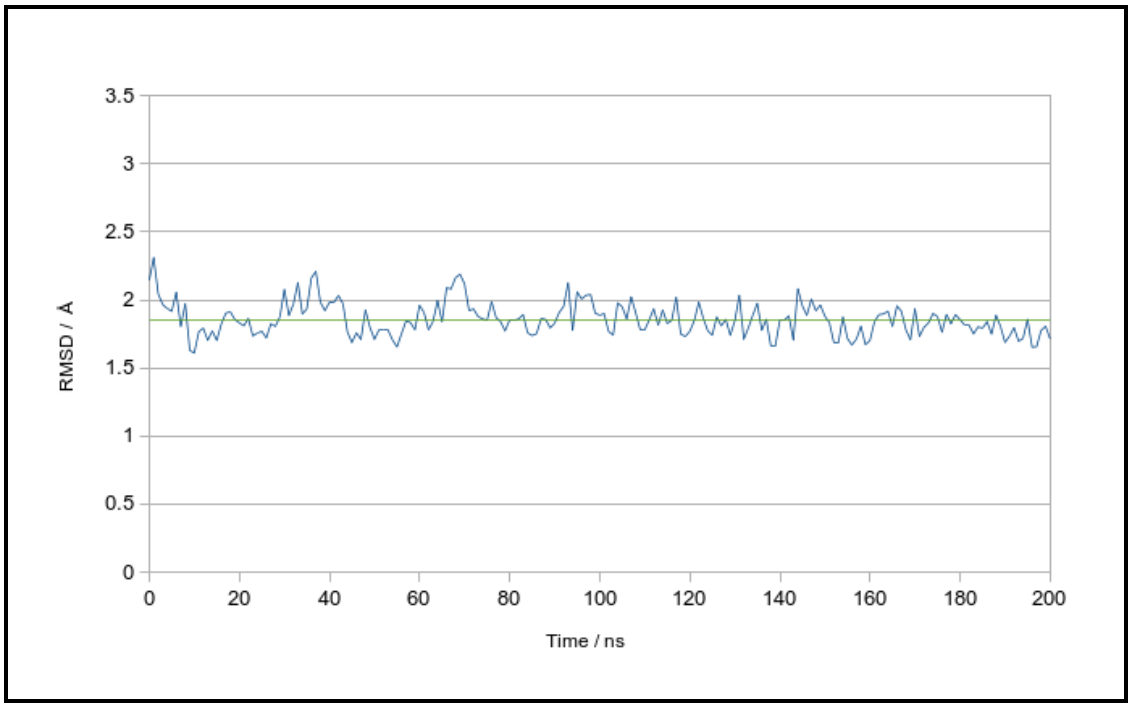

Figure C.21: *RMSD between FAS ACP and the holo ACP-mupA3a WT over time. Green line represents the mean.*

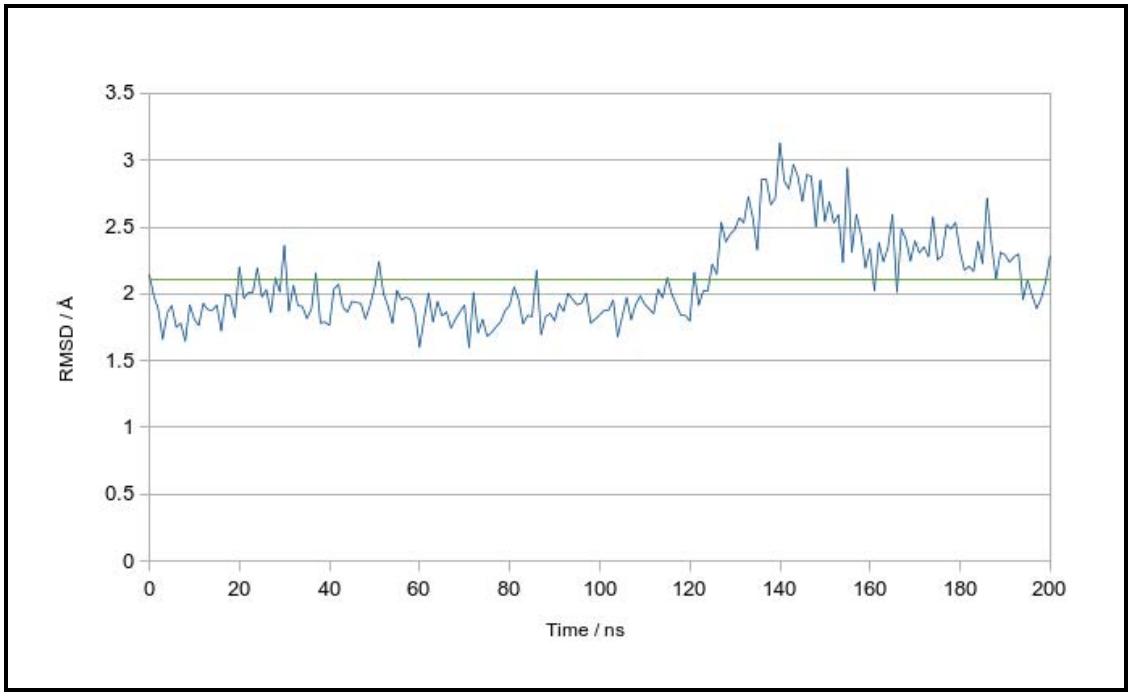

Figure C.22: *RMSD between FAS ACP and the holo ACP-mupA3a W44L over time. Green line represents the mean.*

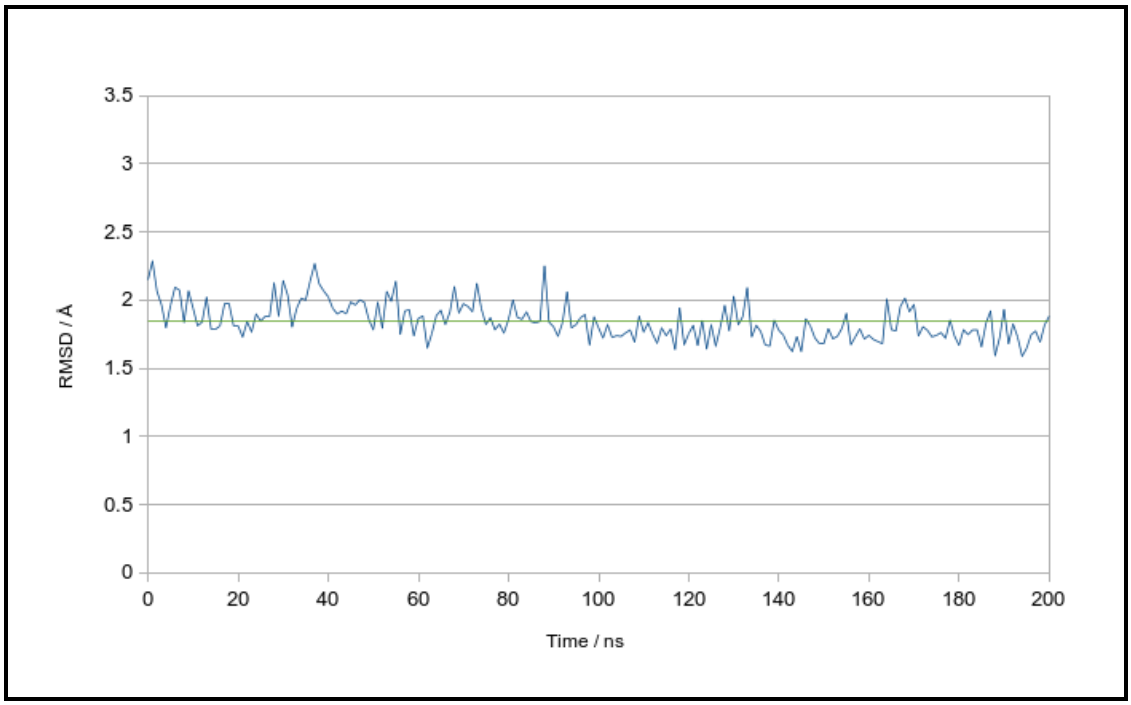

Figure C.23: *RMSD between FAS ACP and the acyl ACP-mupA3a WT over time (200 ns). Green line represents the mean.*

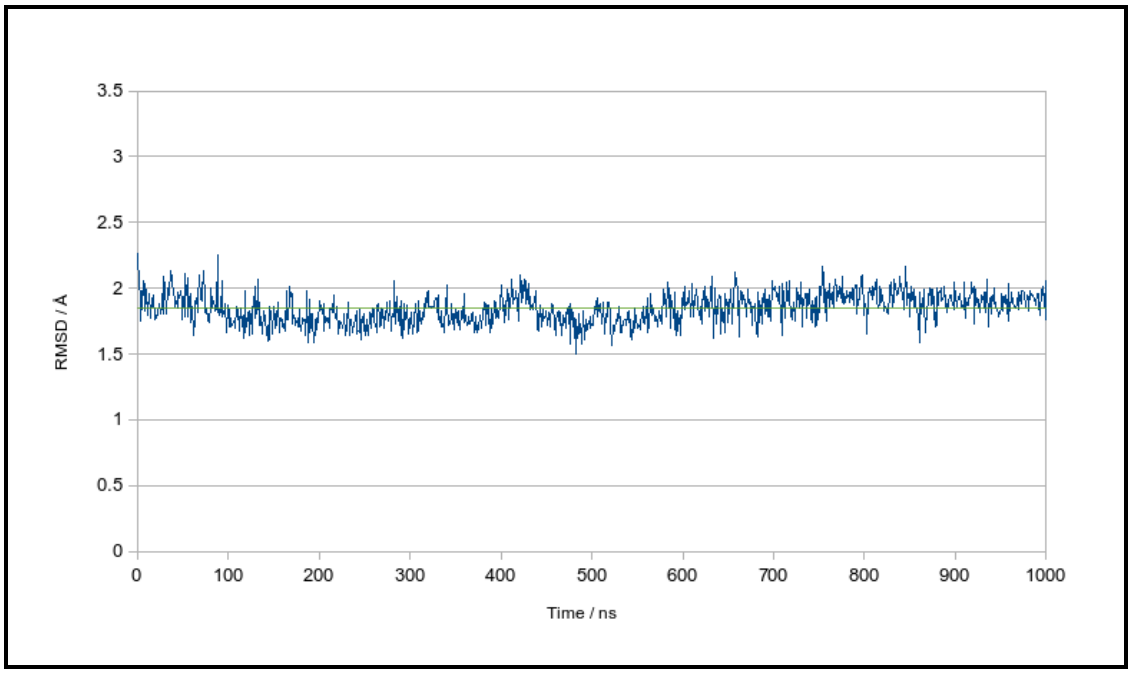

Figure C.24: *RMSD between FAS ACP and the acyl ACP-mupA3a WT over time (1 μs). Green line represents the mean.*

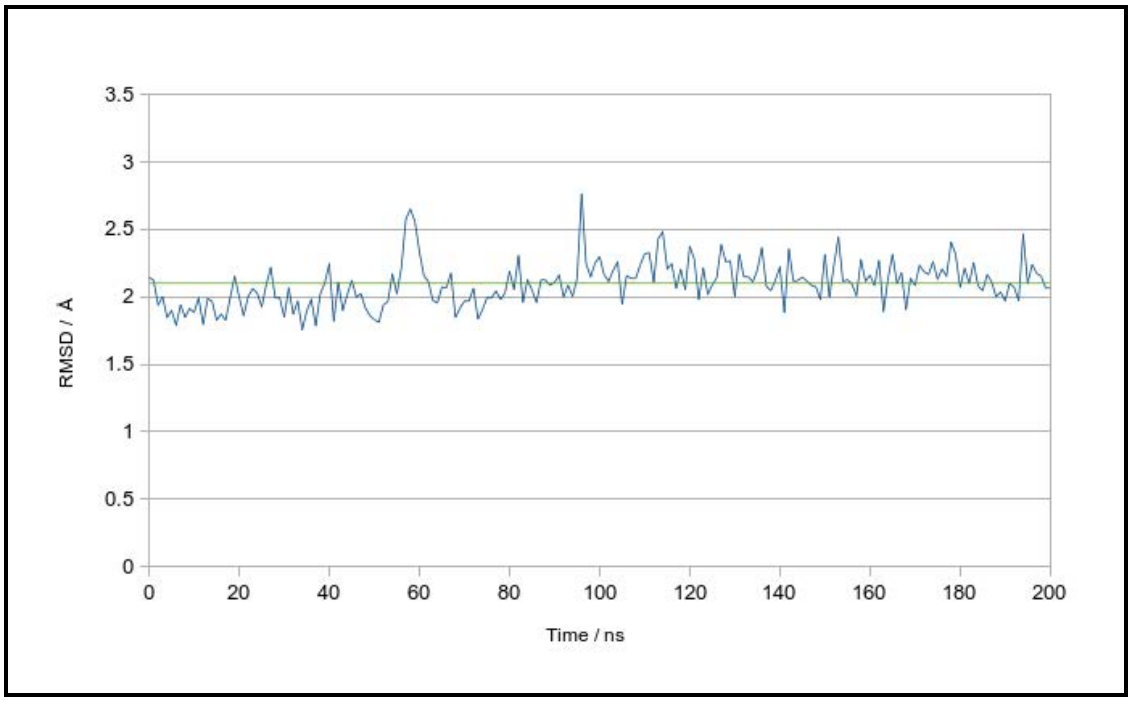

Figure C.25: *RMSD between FAS ACP and the acyl ACP-mupA3a W44L over time. Green line represents the mean.*

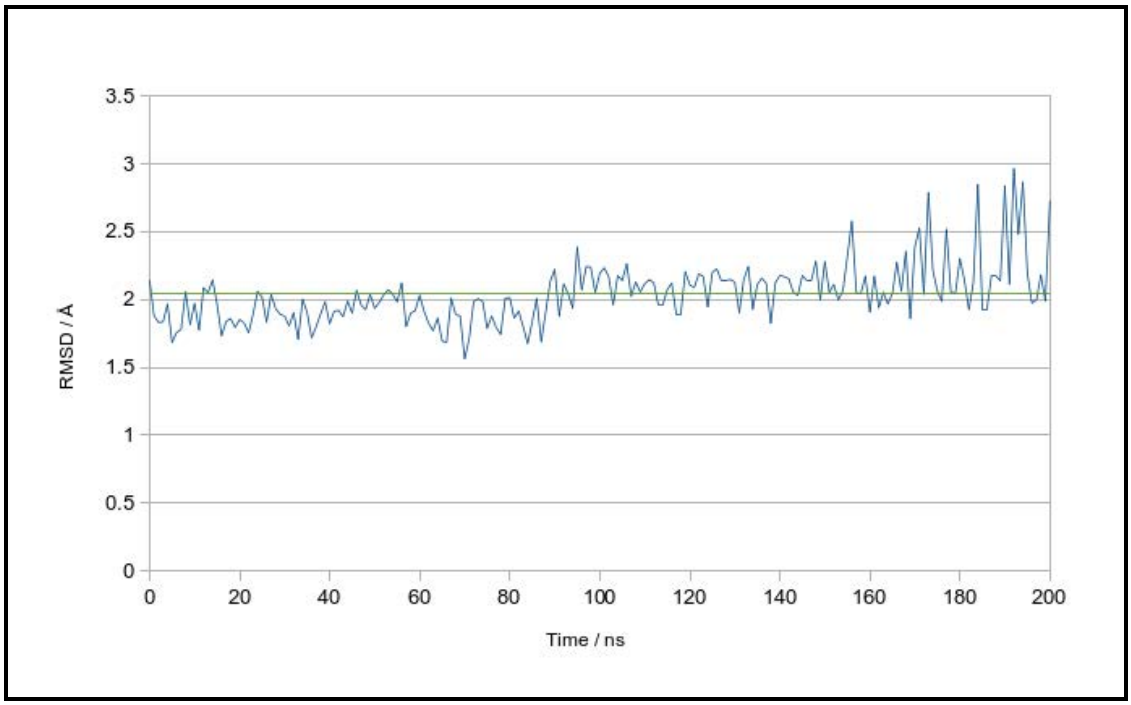

Figure C.26: *RMSD between FAS ACP and the acyl 14C ACP-mupA3a over time. Green line represents the mean.*

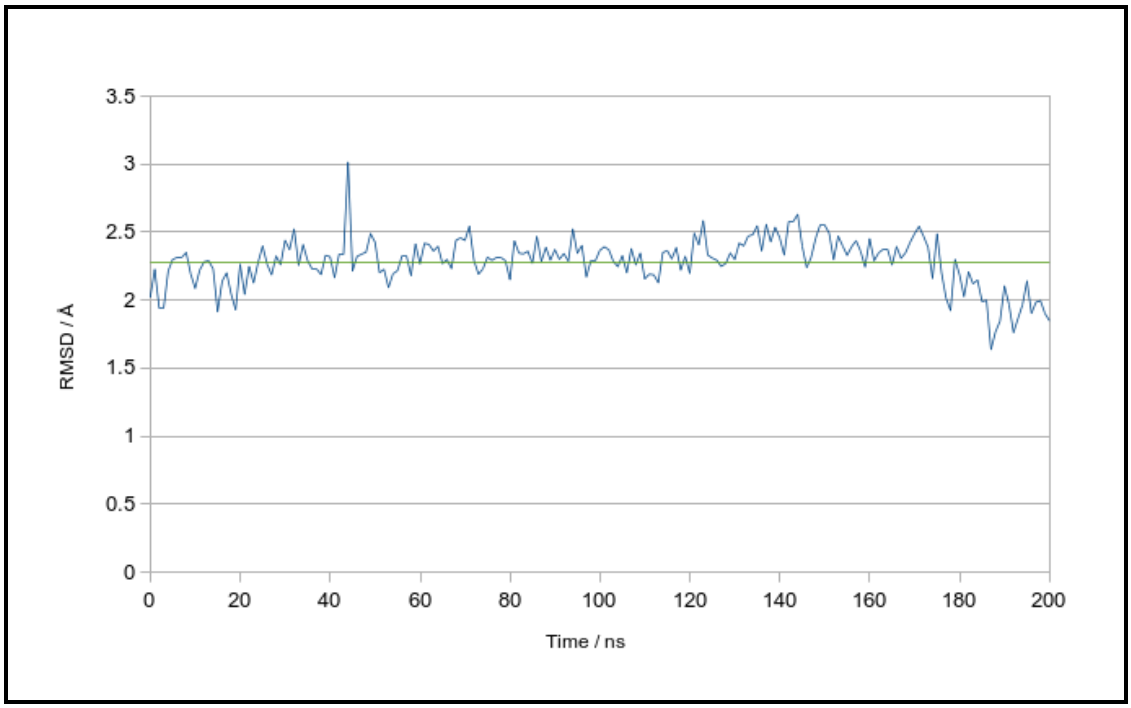

Figure C.27: *RMSD between FAS ACP and the acyl ACP-mupA2a over time. Green line represents the mean.*

C.1.2 Hydrogen bonding between the phosphopantetheine, acyl groups and protein/solvent

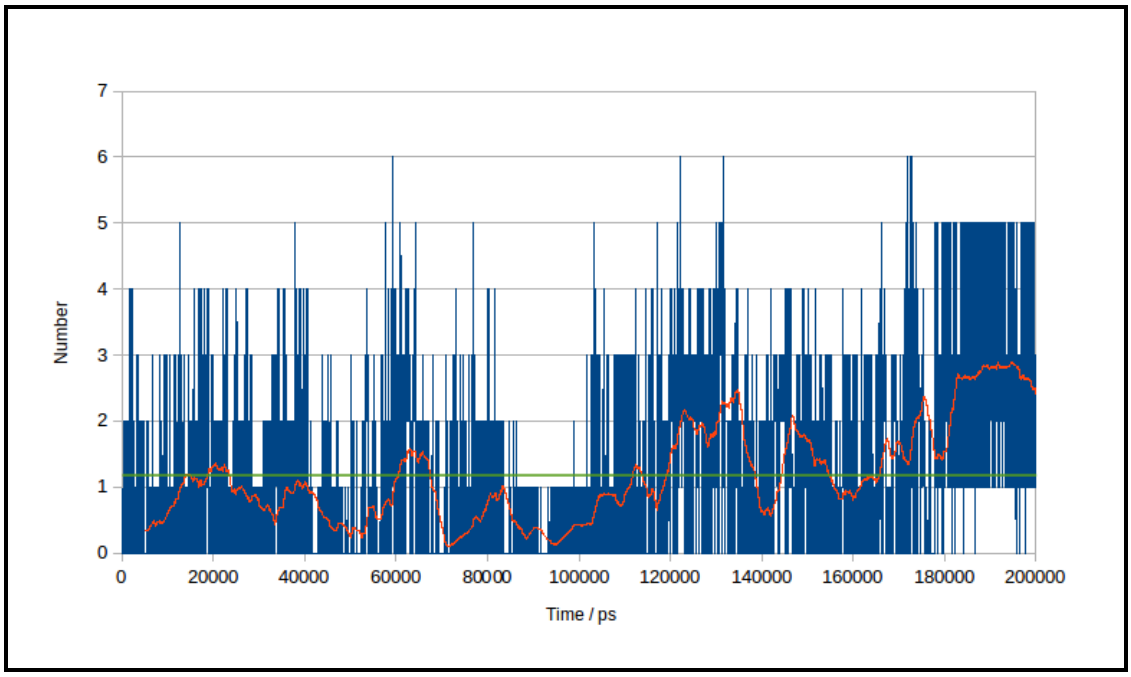

Figure C.28: *Number of hydrogen bonds formed between the phosphopantetheine and the ACP-mupA3a WT surface residues. Red line represents the running average over 500 frames and green line represents the mean.*

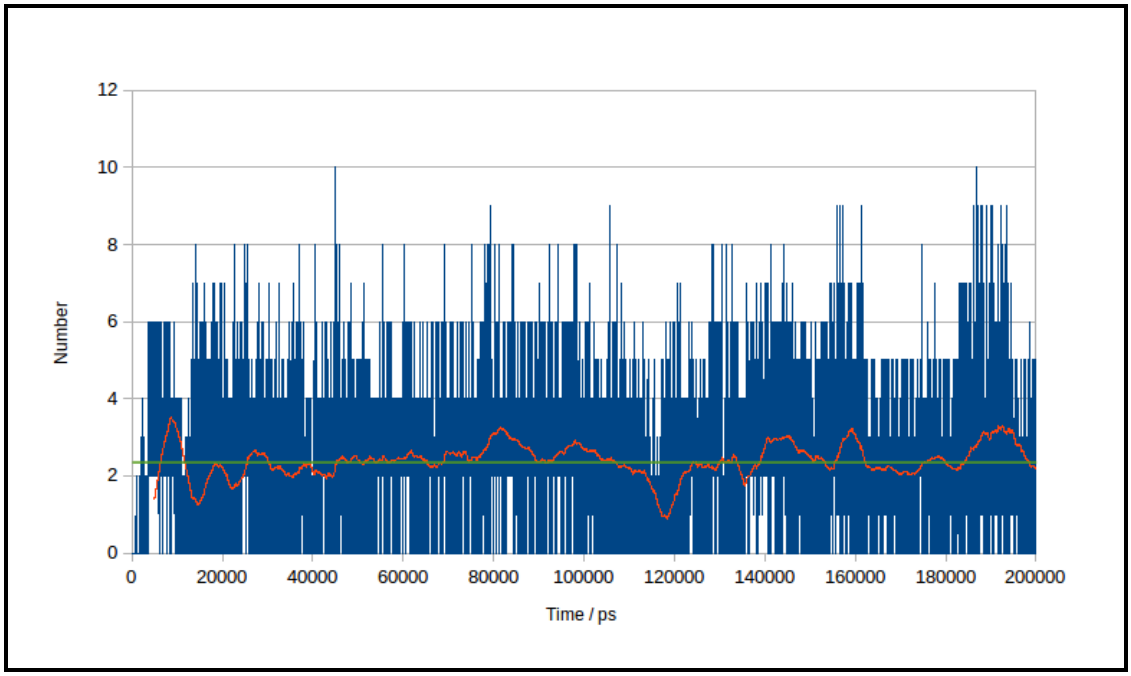

Figure C.29: *Number of hydrogen bonds formed between the phosphopantetheine and the ACP-mupA3a W44L surface residues. Red line represents the running average over 500 frames and green line represents the mean.*

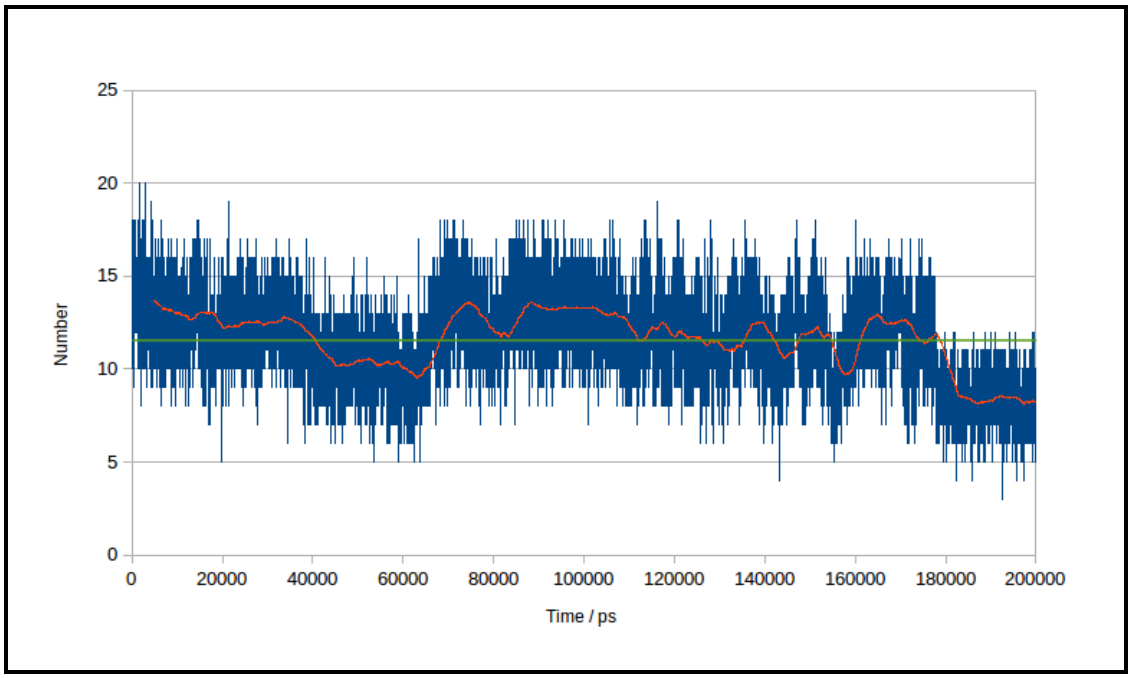

Figure C.30: *Number of hydrogen bonds formed between the phosphopantetheine and the solvent in the ACP-mupA3a WT. Red line represents the running average over 500 frames and green line represents the mean.*

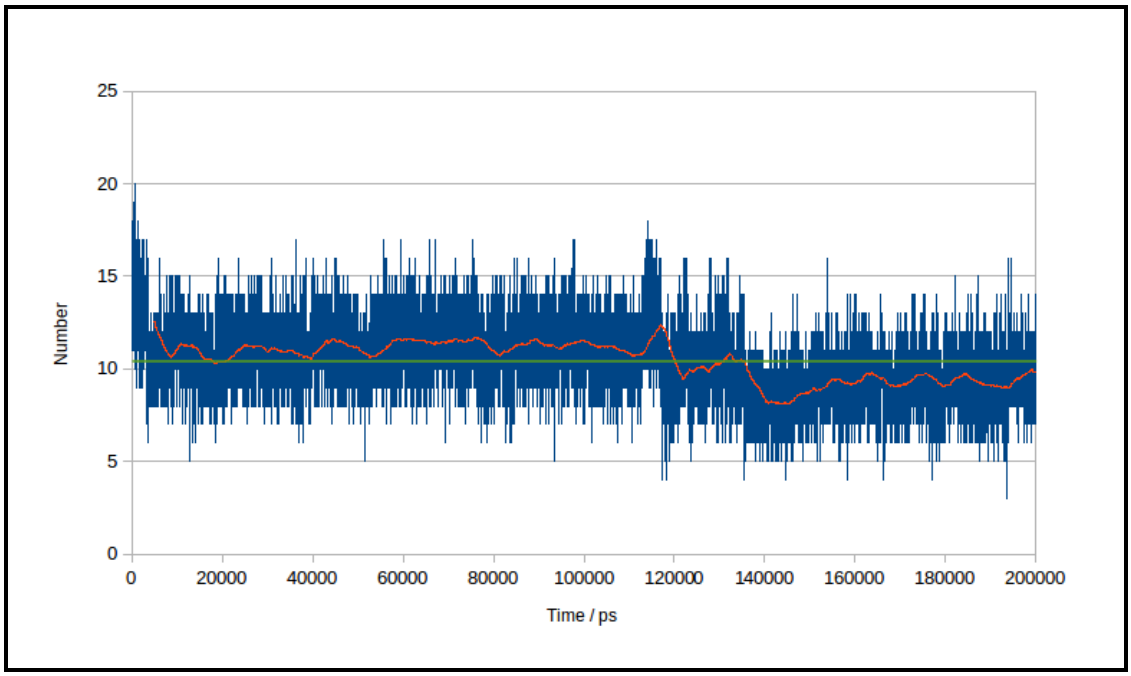

Figure C.31: *Number of hydrogen bonds formed between the phosphopantetheine and the solvent in the ACP-mupA3a W44L. Red line represents the running average over 500 frames and green line represents the mean.*

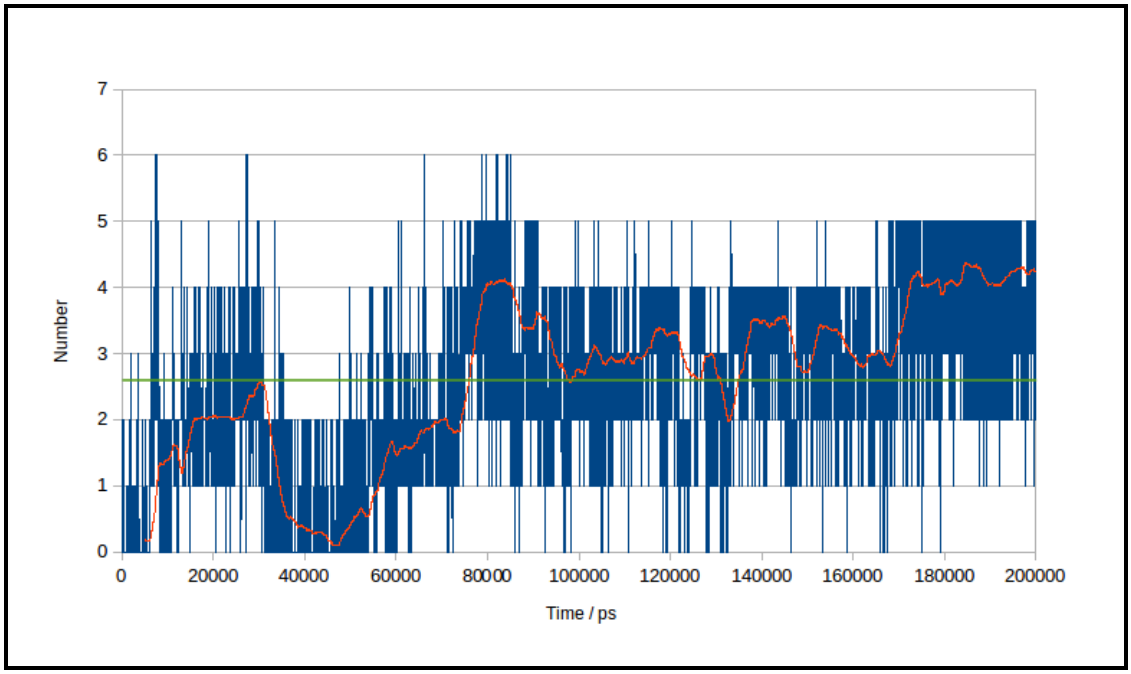

Figure C.32: *Number of hydrogen bonds formed between the ACP-mupA3a cognate substrate and the ACP-mupA3a WT surface residues over time (200 ns). Red line represents the running average over 500 frames and green line represents the mean.*

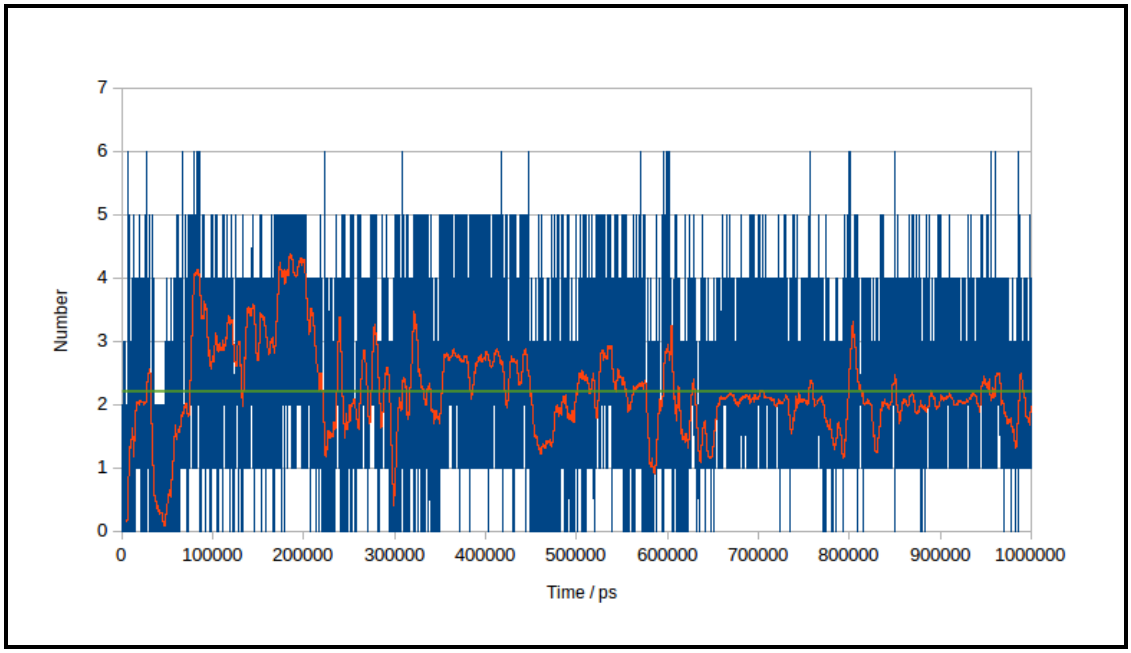

Figure C.33: *Number of hydrogen bonds formed between the ACP-mupA3a cognate substrate and the ACP-mupA3a WT surface residues over time (1 μs). Red line represents the running average over 500 frames and green line represents the mean.*

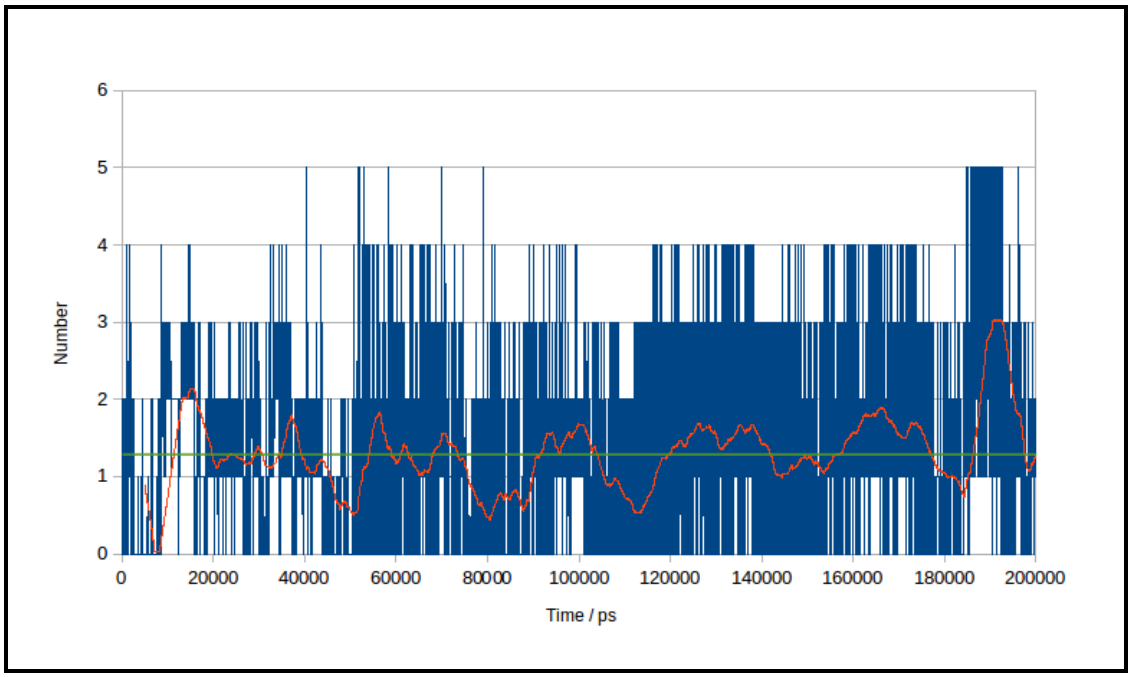

Figure C.34: *Number of hydrogen bonds formed between the phosphopantetheine and the ACP-mupA3a W44L surface residues over time. Red line represents the running average over 500 frames and green line represents the mean.*

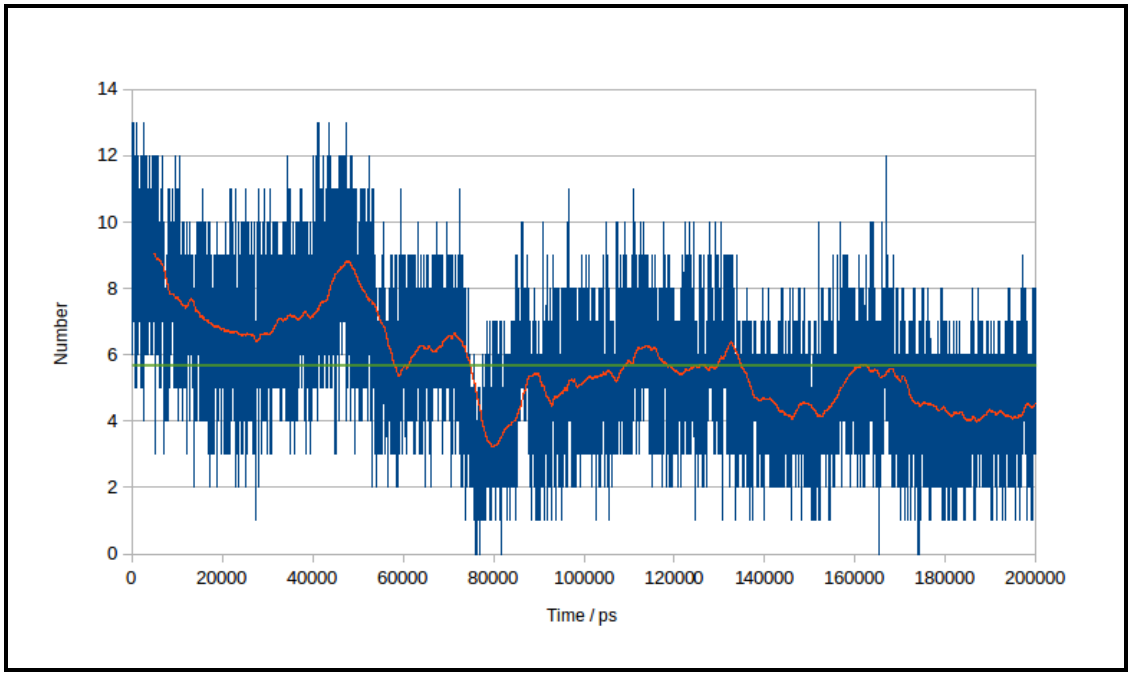

Figure C.35: *Number of hydrogen bonds formed between the ACP-mupA3a cognate substrate and the solvent in the acyl ACP-mupA3a WT simulation (200 ns). Red line represents the running average over 500 frames and green line represents the mean.*

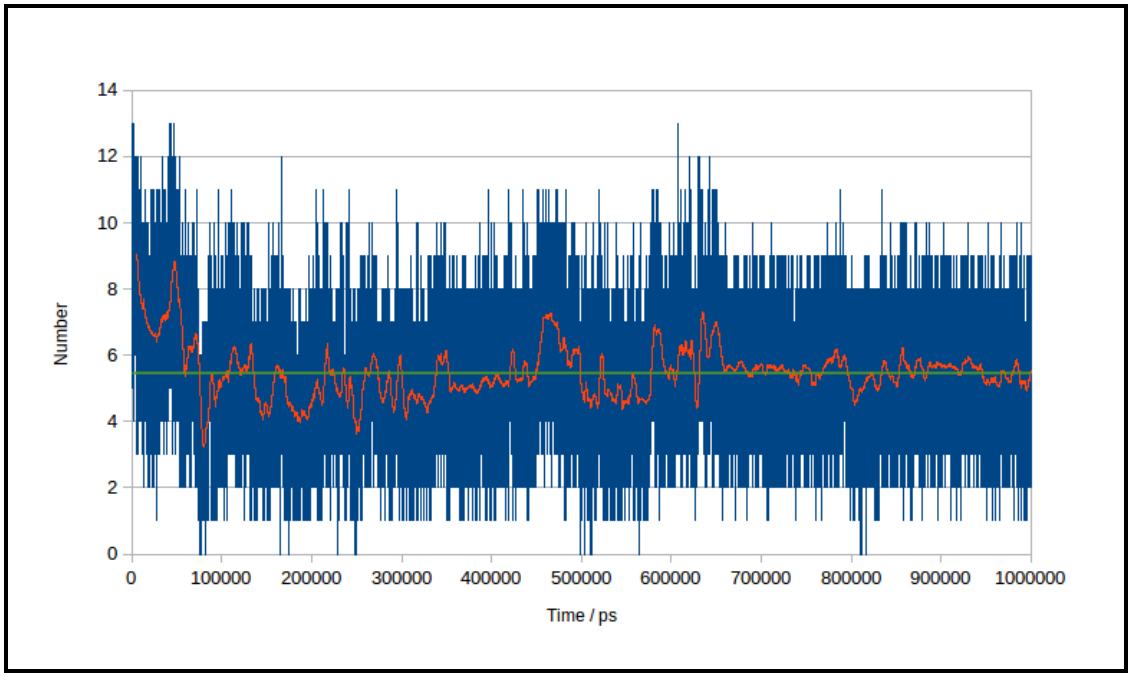

Figure C.36: *Number of hydrogen bonds formed between the ACP-mupA3a cognate substrate and the solvent in the acyl ACP-mupA3a WT simulation (1 μs). Red line represents the running average over 500 frames and green line represents the mean.*

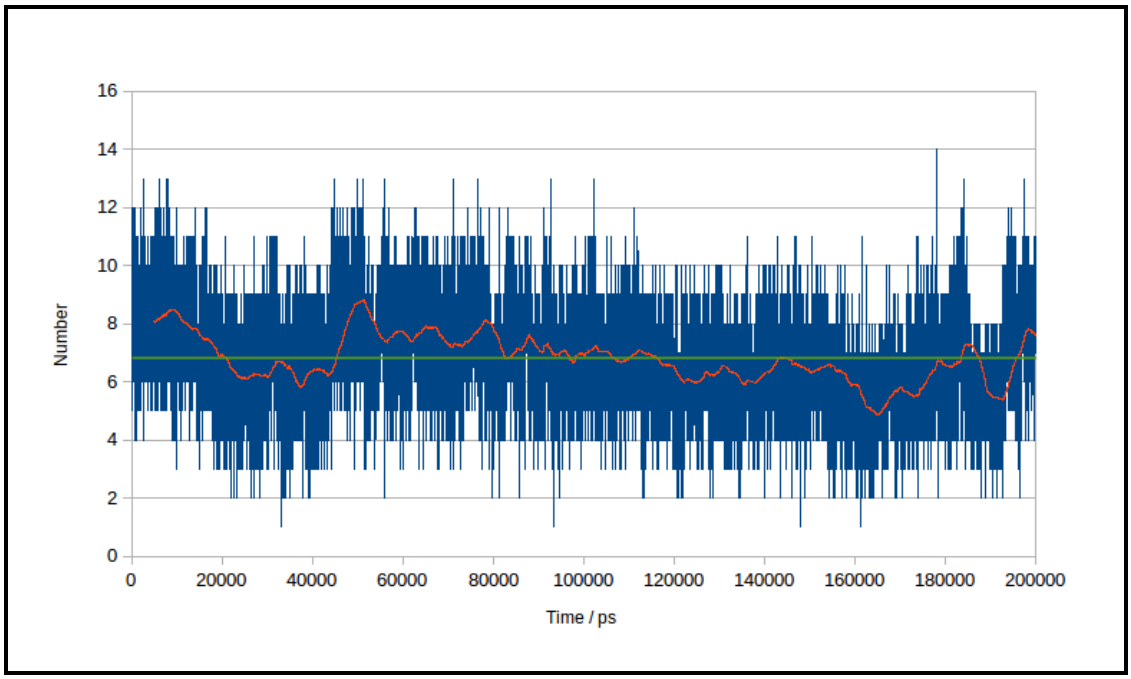

Figure C.37: *Number of hydrogen bonds formed between the ACP-mupA3a cognate substrate and the solvent in the acyl ACP-mupA3a W44L simulation. Red line represents the running average over 500 frames and green line represents the mean.*

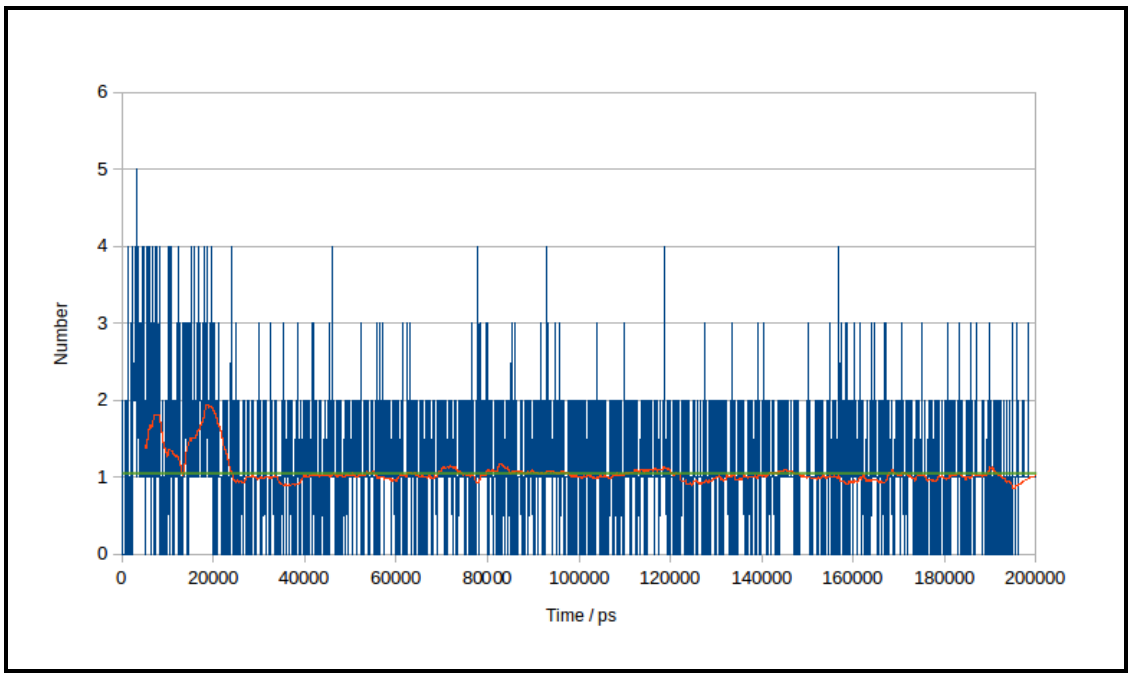

Figure C.38: *Number of hydrogen bonds formed between the ACP-mupA2 cognate substrate and the ACP-mupA2a surface residues. Red line represents the running average over 500 frames and green line represents the mean.*

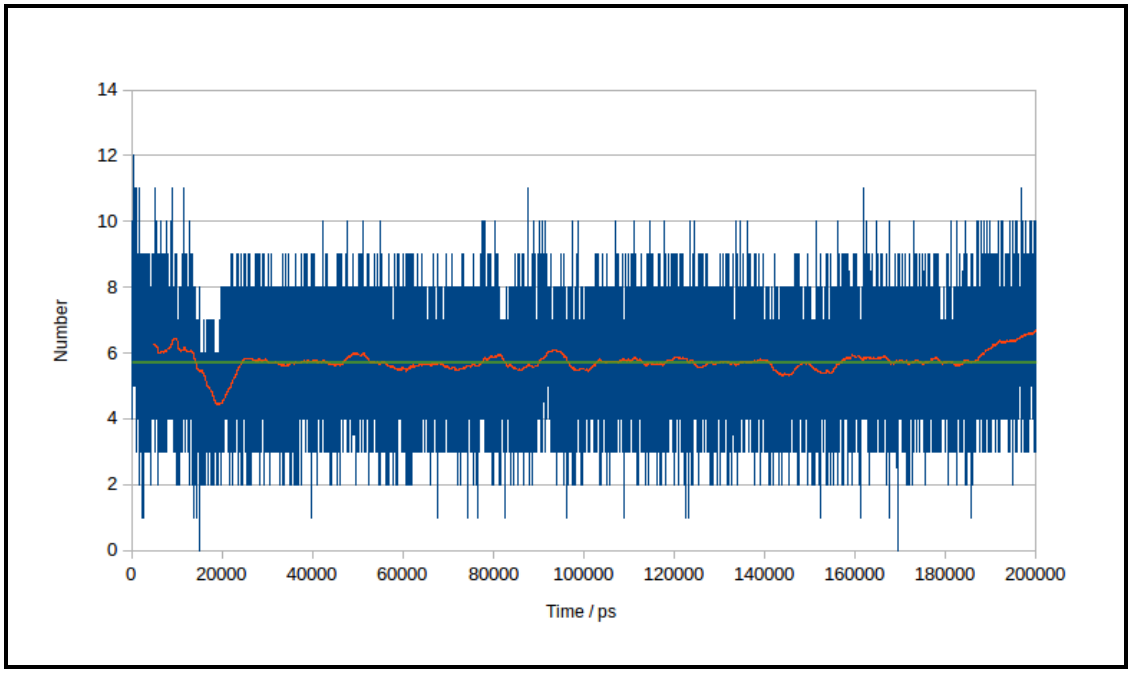

Figure C.39: *Number of hydrogen bonds formed between the ACP-mupA2 cognate substrate and the solvent in the acyl ACP-mupA2a simulation. Red line represents the running average over 500 frames and green line represents the mean.*

C.1.3 Change in solvent accessible surface area of the ligand over time

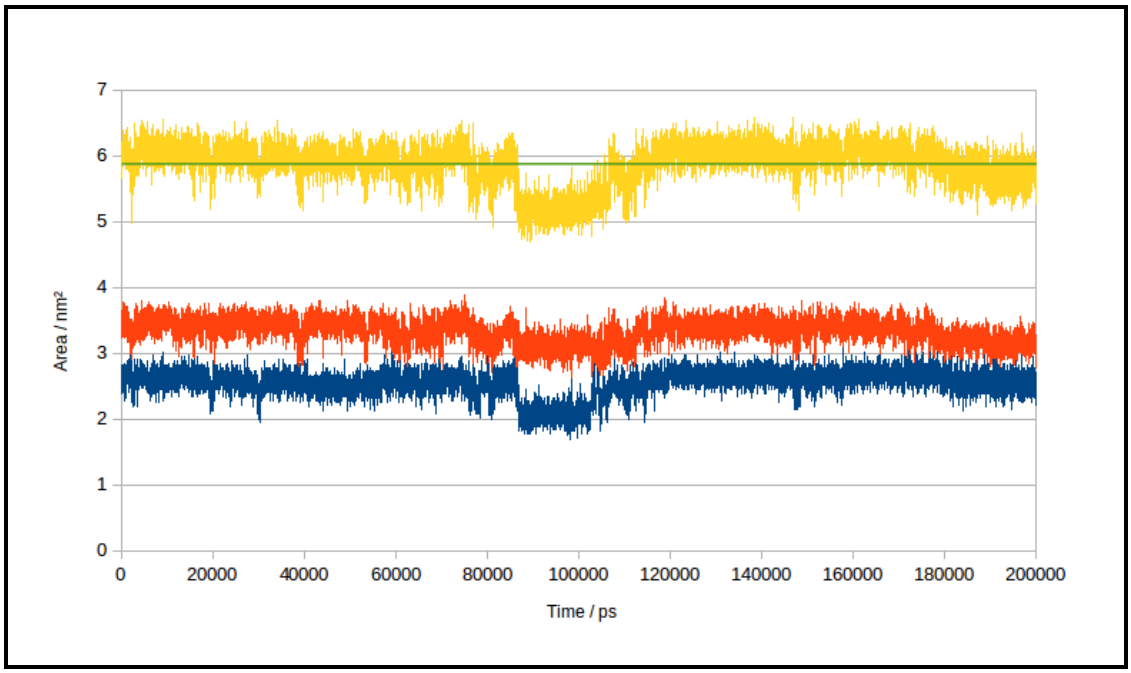

Figure C.40: *Change in the solvent accessible area of phosphopantetheine over time in the holo ACP-mupA3a WT simulation. Blue line represents hydrophobic SAS, red line represents hydrophilic SAS, yellow line represents total SAS and green line represents the mean of the total SAS.*

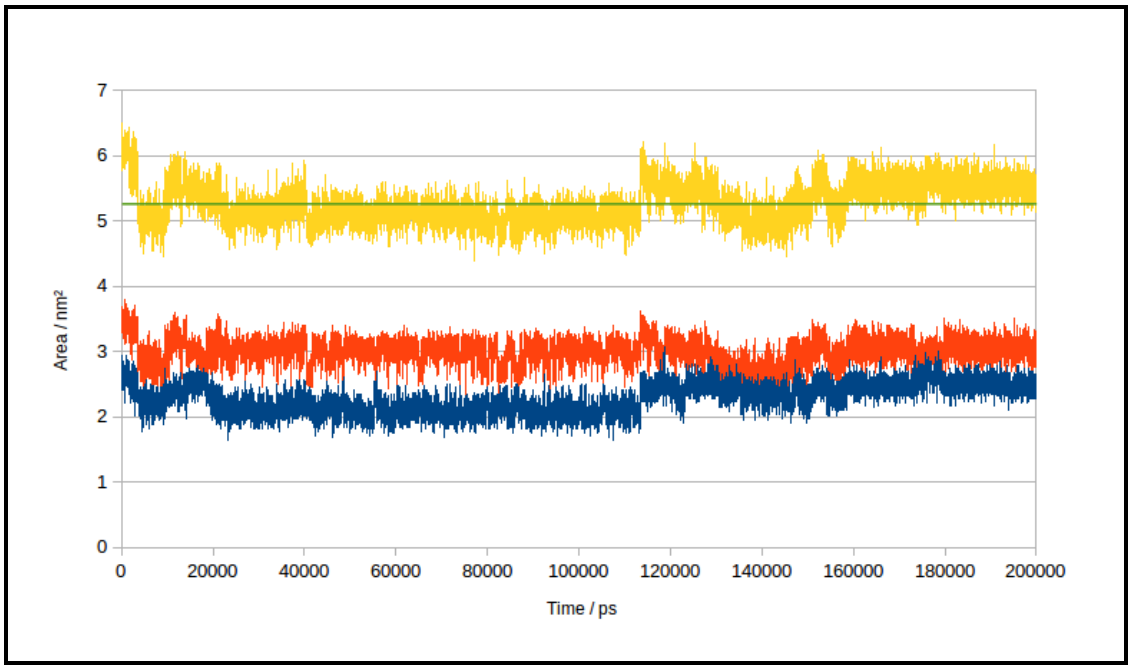

Figure C.41: *Change in the solvent accessible surface (SAS) of phosphopantetheine over time in the holo ACP-mupA3a W44L simulation. Blue line represents hydrophobic SAS, red line represents hydrophilic SAS, yellow line represents total SAS and green line represents the mean of the total SAS.*

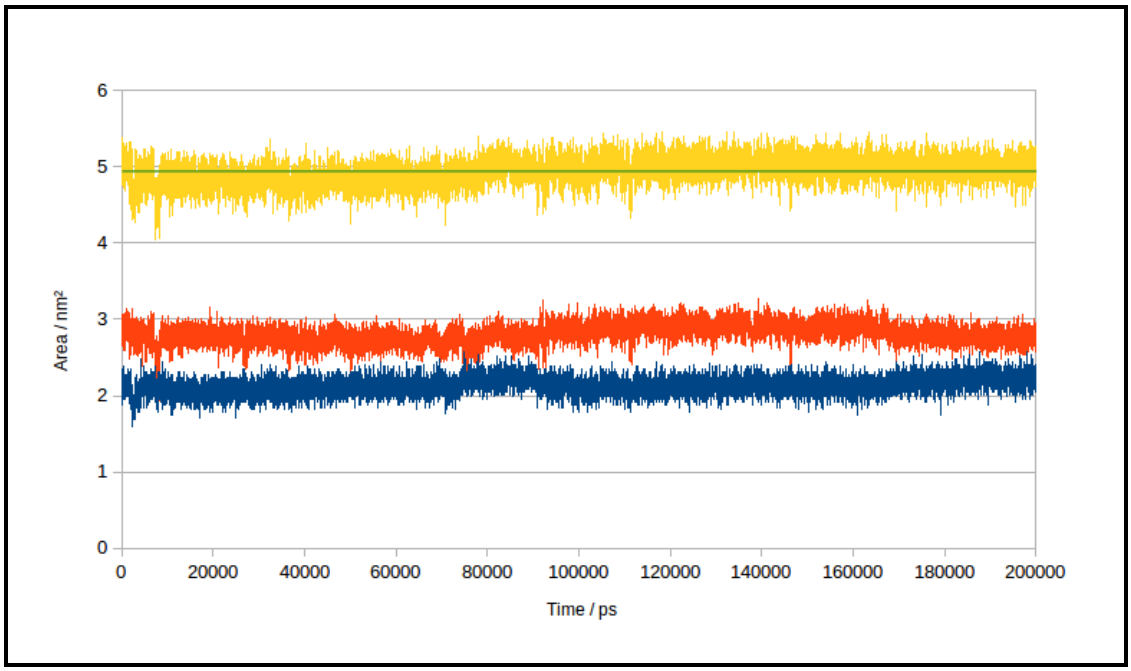

Figure C.42: *Change in the solvent accessible area of the ACP-mupA3a cognate substrate over time (200 ns) in the acyl ACP-mupA3a WT simulation. Blue line represents hydrophobic SAS, red line represents hydrophilic SAS, yellow line represents total SAS and green line represents the mean of the total SAS.*

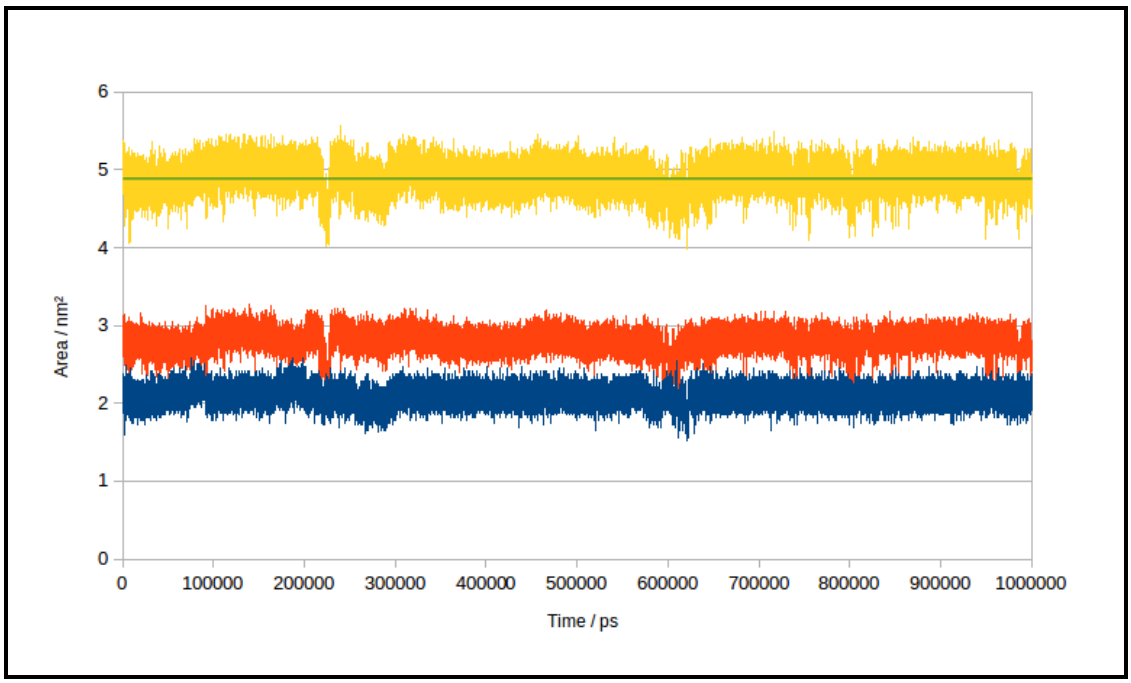

Figure C.43: *Change in the solvent accessible surface (SAS) of the ACP-mupA3a cognate substrate over time (1 μs) in the acyl ACP-mupA3a WT simulation.. Blue line represents hydrophobic SAS, red line represents hydrophilic SAS, yellow line represents total SAS and green line represents the mean of the total SAS.*

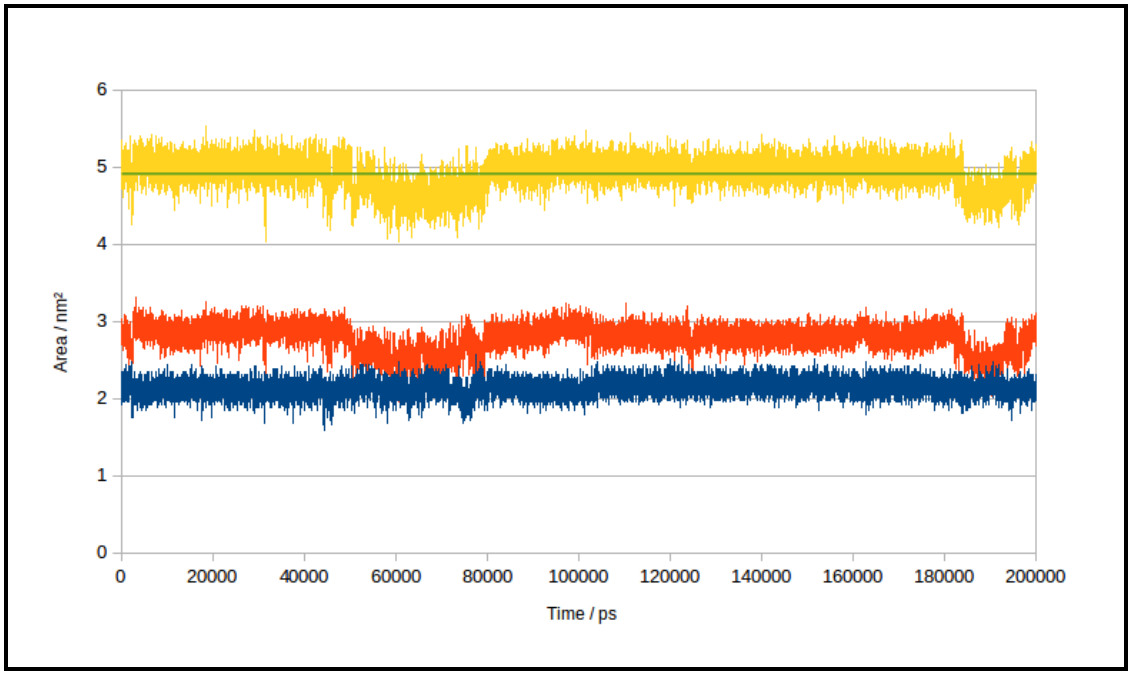

Figure C.44: *Change in the solvent accessible surface (SAS) of the ACP-mupA3a cognate substrate over time in the acyl ACP-mupA3a W44L simulation. Blue line represents hydrophobic SAS, red line represents hydrophilic SAS, yellow line represents total SAS and green line represents the mean of the total SAS.*

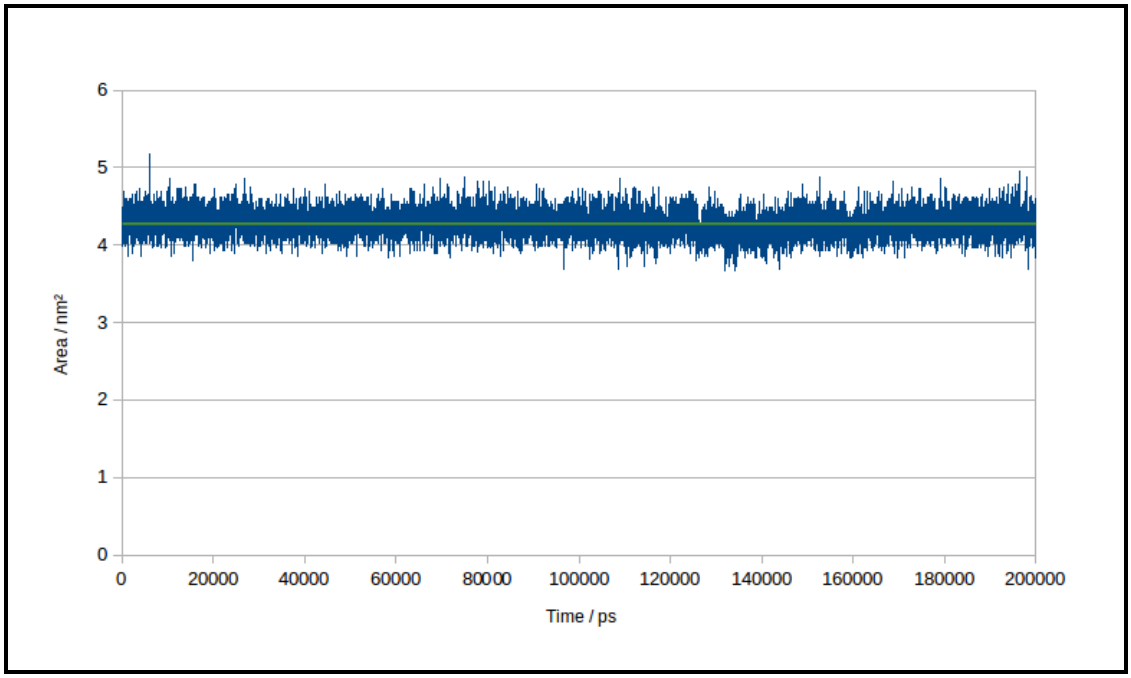

Figure C.45: *Change in the solvent accessible surface (SAS) of the 14C saturated chain over time in the acyl 14C ACP-mupA3a simulation. Blue line represents hydrophobic SAS and green line represents the mean of the total SAS.*

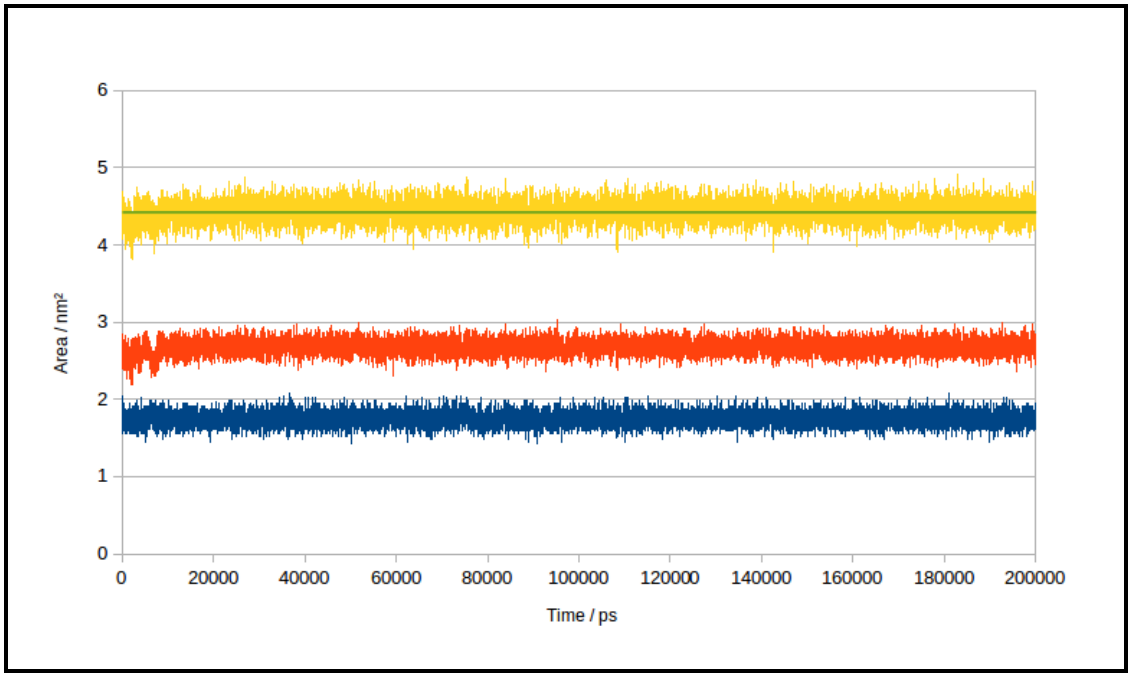

Figure C.46: *Change in the solvent accessible surface (SAS) of the ACP-mupA2 cognate substrate over time in the acyl ACP-mupA2a simulation. Blue line represents hydrophobic SAS, red line represents hydrophilic SAS, yellow line represents total SAS and green line represents the mean of the total SAS.*

C.1.4 Sequence logos

On the next page.

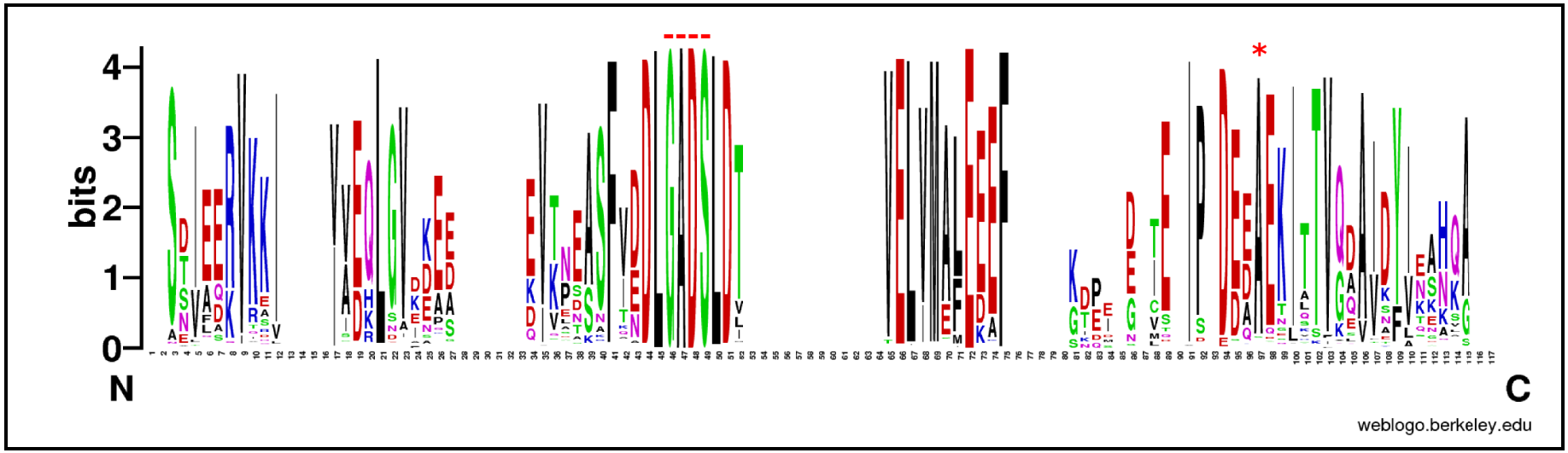

Figure C.47: *Sequence logo built on 2078 unique FAS ACP sequences with GADS motif. These sequences were searched with normal BLASTp without any pattern and post search filtered for sequences carrying GADS motif.* - - - - - *indicates the GADS position and* * *indicates the position equivalent to A59 in the FAS ACPs.*

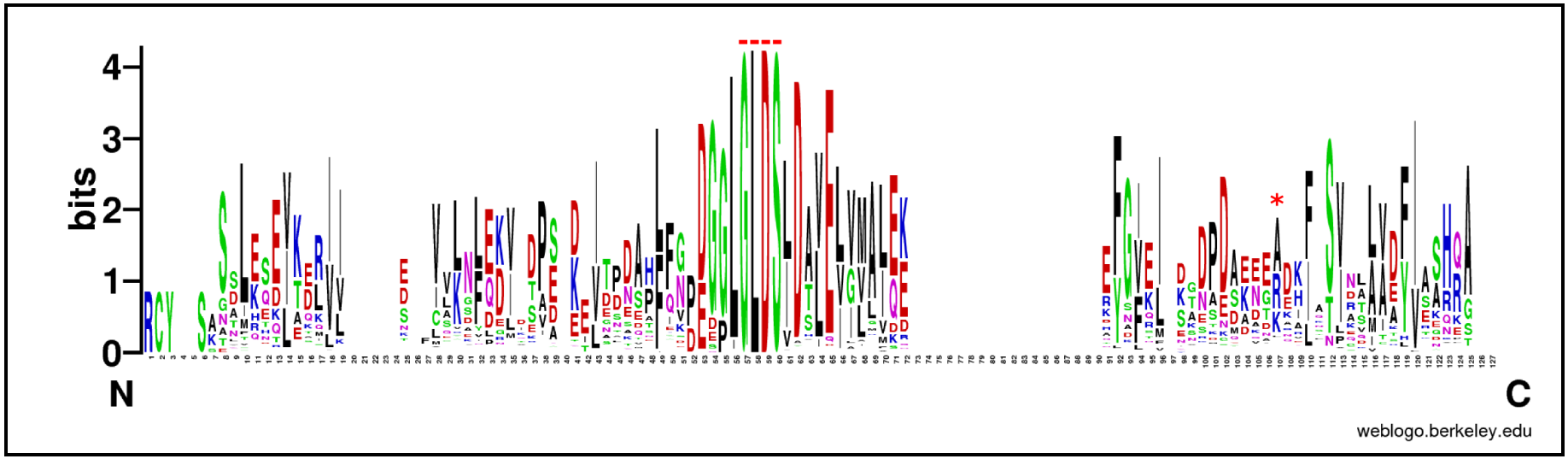

Figure C.48: *Sequence logo built on 257 unique FAS ACP sequences with GLDS motif. These sequences were searched with normal BLASTp without any pattern and post search filtered for sequences carrying GLDS motif.* - - - - - *indicates the GLDS position and* * *indicates the position equivalent to A59 in the FAS ACPs.*

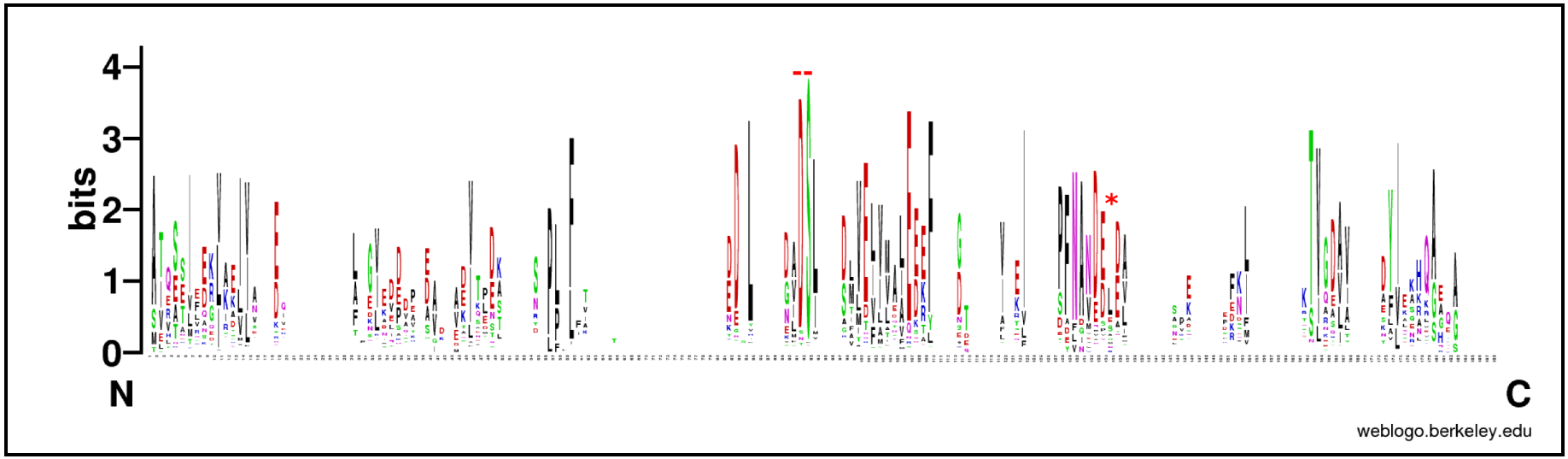

Figure C.49: *Sequence logo built on 541 unique FAS ACP sequences with neither GADS or GLDS motif. These sequences were searched with normal BLASTp without any pattern and post search filtered for sequences devoid of GADS or GLDS motif. - - - - - indicates the GXDS position and * indicates the position equivalent to A59 in the FAS ACPs.*

Chapter D

Appendix IV

D.1 Computational Tools available for the PKS researcher.

Comparing all the resources mentioned in Section 1.5 the various tasks that can be performed can be divided into four major categories: 1) obtaining a well curated PKS/NRPS cluster database for further analysis; 2) domain detection for various modular PKSs and NRPs; 3) predicting substrate specificity for various starter and extender units; and 4) correlating identified clusters to their corresponding metabolites.

The NRPS-PKS database serves as an amalgamated source for PKS clusters (both modular and iterative), NRPS clusters and chalcone like products. At the time of first publication data in this database was derived from the PKSDB (19 modular PKS cluster), NRPSDB (17 NRPS clusters and 5 hybrid PKS+NRPS clusters), ITERDB (21 iterative PKSs) and CHSDB (11 plant chalcone PKSs and 3 bacterial chalcone PKSs) databases (Yadav *et al.* 2003b). Another database ASMPKS at the time of publication contained 41 characterized PKS pathways including everything in PKSDB, with more entries being added (Tae *et al.* 2007). Users may also add or delete their own entries. NORINE is a database for non-ribosomal peptides containing 1122 peptide products and over 500 monomers as of 04/2010, however it does not provide information on biosynthesis (Caboche *et al.* 2008).

Why is there a need for specialized databases for PKS/NRPS research when databases like CDD (Marchler-bauer *et al.* 2011) and interPro (Hunter *et al.* 2012), or domain finding software

like SMART (Letunic *et al.* 2012) exist? It was observed by Yadav *et al.* (2003a) during the construction of PKSDB/SEARCHPKS that inspite of CDD and interPro being a vast source of protein domains they suffer from being general and not tailored for a specific purpose. Comparing the CDD results from their domain identification program they found that at that time CDD failed to detect any of the DH domains in the modular PKS clusters and was also not able to distinguish between the KR and ER domains. However, over the time CDD has improved and in my analysis with the MmpD subunit from the mupirocin pathway it is able to predict all the domains except the catalytically inactive DH domain in module 1.

Tailored NRPS/PKS domain prediction has been achieved by either a BLAST (Altschul *et al.* 1990) search of the query sequence against a backend database or through querying a database of profile hidden markov models (HMMs) trained on domains from PKS/NRPS clusters. The domain detection algorithms in NRPS-PKS, ASMPKS and SBSPKS use BLAST whereas ClustScan, NRPSsp, CLUSEAN and antiSMASH uses profile HMMs. HMMs have also been used by Ansari and coworkers (Ansari *et al.* 2008) to identify methyl transferase (MT) domains present in type I PKS and NRPS megasynthases and to sub-group them into N-MT, C-MT and O-MT groups based on the prediction of their site of methylation. In another study, Foerstner *et al.* (2008) used HMMs to screen eight metagenomics shot gun data sets in order to estimate the frequency of type I PKSs, using HMMs for eight domains. They also incorporated analysis of maximum-likelihood phylogenetic trees to increase the reliability and resolution of the dataset and to discriminate true PKS I domains from evolutionarily related but functionally different ones.

ClustScan (Starcevic *et al.* 2008) also utilizes profile HMMs, both extracted from Pfam and specifically constructed profiles. ClustScan works as a client server application with the main program running on a linux server and a Java client running on the user's computer, making it compatible with Windows, Mac OS and Linux operating systems. It uses Glimmer or GeneMark for the gene prediction followed by HMMs for the domain identification in PKS/NRPS/PKS-NRPS hybrid clusters. At the time of publication ClustScan database contained data for 57 PKS clusters, 51 NRPS clusters and 62 PKS-NRPS hybrid clusters (Starcevic

et al. 2008). Apart from domain identification and substrate specificity detection ClustScan can also exports the chemical structures of predicted products in a SMILES/SMARTS format for further analysis by standard chemistry programs. ClustScan also acts as a graphical interface to CompGen, a tool that undertakes *in silico* homologous recombination of PKS gene clusters, predicting whether a particular recombination is likely to be functional, and the polyketide product to be expected (Starcevic *et al.* 2012). ClustScan can be accessed using a 30-day evaluation license; the database behind ClustScan, ClustScanDB is freely available via a web interface (`http://csdb.bioserv.pbf.hr/csdb/ClustScanWeb.html`).

To facilitate the systematic mapping of secondary metabolites in fungal genomes Khaldi *et al.* (2010) developed the SMURF (Secondary Metabolite Unknown Region Finder). SMURF is a web based tool which relies on hidden Markov model searches against Pfam and TIGRFAMs (Haft *et al.* 2001) domains to detect backbone genes in sequenced fungal genomes. SMURF is not only capable of predicting backbone genes but also tailoring enzymes and Khaldi *et al.* (2010) used SMURF to catalogue putative clusters in 27 publically available fungal genomes. They also compared the predicted results with genetically characterized clusters from 6 fungal species and demonstrated that SMURF is capable of predicting accurately all of these clusters.

Recently NRPSpredictor2 (Röttig *et al.* 2011) used an innovative method employing a machine learning algorithm called a transductive support vector machine (SVM) for predicting the specificity of adenylation domains in NRPs from amino acid sequence data. The method is based on previous work (Rausch *et al.* 2005) from the same group however the new version outperforms the previous version by predicting the specificity of adenylation domains at four hierarchical levels, ranging from gross physicochemical properties of an A-domain's substrate to single amino acid substrates as well as predicting A-domain specificity in fungal systems, which was not achieved in the previous version. The NRPSpredictor2 utilizes the active site lining residues within 8 Å of the bound phenylalanine ligand in the crystal structure (PDB ID 1AMU) of the peptide synthetase gramicidin S synthetase 1 (GrsA). These 34 extracted positions were then located in the A-domain sequences of the training data set, and this data was input into the SVM to train predictors of substrate specificity. The NRPSpredictor2 also has a

larger database of training data including the sequences from fungal counterparts, as compared to its previous version, thus enabling a wider and more accurate prediction rate. This work is an extension of ideas built up by a number of researchers over a number of years, as discussed further below.

Many research groups have used structural data to determine conserved residues lining the active site, which they then use for the prediction of substrate specificity. In the case of PKS this is typically specificity for extender or starter units, in the case of NRPs specificity for amino acids. Prior to NRPSpredictor2, Stachelhaus and coworkers (Stachelhaus *et al.* 1999) had utilized the 10 active site lining residues from the same crystal structure of the peptide synthetase gramicidin S synthetase 1 (GrsA, PDB ID 1AMU) and succeeded in predicting the specificity of the A-domain for 20 substrates. Shortly after their work Challis et al. (Challis *et al.* 2000), adopting a similar strategy, extended the number of predictable NRPS substrates to 33. Challis *et al.* (2000) used 8 amino acids within the binding pocket combined the phylogenetic clustering. They also performed modelling for the structures of a variety of binding pockets.

Similarly, during the compilation of the first polyketide synthase database (PKSDB) and associated domain prediction program SEARCHPKS, Yadav *et al.* (2003a) identified 13 active site positions in AT domains required to discriminate between malonate and methylmalonate as starter and extender units in type I modular polyketide synthases. The 13 active site residues were identified on the basis of the crystal structure of acyltransferase from *Escherichia coli* FAS (PDB ID 1MLA). By modelling malonate and methylmalonate in the active site of the AT they identified the residue which is responsible for controlling substrate specificity. The online server at SEARCHPKS assigns a substrate for the AT domain if all the 13 positions in the query amino acid sequence match identically to the corresponding positions in any of the AT domains found in the PKSDB database.

Further utilizing structural modelling of the active site Yadav *et al.* (2009) identified that certain residues in iterative KS domains can potentially control the size of final product by governing the total number of iterations. In the same work Yadav et al. also utilized profile HMMs to distinguish KS domains between modular PKSs and iterative PKSs, they also ob-

served that HMMs are not only capable of broadly classifying KSs as modular or iterative but also of grouping them into subtypes. They proposed that such a method can help the sequencing projects as just by analyzing the KS domains of a novel PKS cluster one can identify its type and subtype and decide whether sequencing the entire cluster would be of interest. From these recent works HMMs prove to be a promising tool for various types of analysis ranging from domain identification to substrate specificity.

Recently available coordinates of crystal structures of various type I PKS catalytic domains (Keatinge-Clay and Stroud 2006; Khosla *et al.* 2007; Tang *et al.* 2007; Keatinge-clay 2008; Khosla 2009; Tsai and Ames 2009) and an almost complete module of the homo dimeric mammalian FAS protein (Maier *et al.* 2008) allow the modelling of PKS domains in a homo dimeric modular context, assuming the PKSs have a similar structure to the mammalian FAS (Gokhale *et al.* 2007; Tsai and Ames 2009). Based on this, Mohanty's group developed SBSPKS (Anand *et al.* 2010). SBSPKS is a web based tool and probably the first which can model the 3D structures of a PKS module in a biologically active dimeric conformation. SBSPKS consists of three main components, MODEL_3D_PKS, DOCK_DOM_ANAL and an updated version of NRPS-PKS which was previously developed by the same group. MODEL_3D_PKS component models the 3D structures of a complete module of a type I modular PKS protein in dimeric form. The homology modelling protocol used involves aligning the query module sequence to the sequence of the templates by standalone BLAST and side chain modelling using SCRWL (Canutescu *et al.* 2003). The templates used for the various domains of PKSs were PDB ID 2HG4, 3LE6, 1IZ0 and 2FR0 along with nine threading model based on 2FR0 for modelling structural subdomains of KR in the cases where DH-KR and DH-ER linkers lacked homology to 2FR0. The modelled domains are then superimposed on to the corresponding mammalian FAS module to provide the relative orientation of the PKS domains in a dimeric state. MODEL_3D_PKS can model 3D structures for any of the four typical combinations of modular PKS i.e. KS-AT-ACP, KS-AT-KR-ACP, KS-AT-DH-KR-ACP, KS-AT-DH-ER-KR-ACP, excluding the ACP domain as there is no experimental information available to provide the relative orientation of ACP domains to the rest of the domains in the module. Although MODEL_3D_PKS can model the

typical combinations of domains found in modular PKSs it is not designed to model the module from trans AT systems, although it may be possible to trick the system into modelling this. The MODEL_3D_PKS web interface also has an embedded JMOL applet for quick visualization of the modelled structure.

The DOCK_DOM_ANAL module of the SBSPKS analyses the docking domains between the related subunits in a modular PKS. The docking domains are the inter-subunit linker region characterized by a four helical bundle, one helix is from the C-terminus of the preceding protein and three helices are from the N-terminus of the succeeding protein. In previous studies (Broadhurst *et al.* 2003; Weissman 2006; Weissman and Müller 2008) a "docking code" was proposed, in which the electrostatic interaction between two residues in the docking domains were responsible for the inter-subunit contacts. DOCK_DOM_ANAL estimates the crucial inter-subunit contacts and predict the preferred order of substrate channelling utilizing the method developed by Yadav *et al.* (2009). As a point of terminology these inter subunit docking domains should not be confused with intra module segments seen in trans AT PKS I systems, which show sequence similarity to the N and C termini of *cis*-AT domain, and may thus be remnants of such domains although their role is still poorly understood (Gurney and Thomas 2011).

The most important updated feature in NRPS-PKS component of SBSPKS is a wider range of substrate specificity detection for the AT domain. The initial version of the AT domain specificity protocol was only able to discriminate between the malonate and methyl malonate selectivity while the new version can now detect specificity for a total of 13 substrates. Another enhanced feature is its integration into the SBSPKS interface, thus providing links for automated input of its results to various other programs in the suite.

Apart from the easy-to-use web servers and client server based applications like ClustScan, recent initiatives by Weber and co-workers resulted in CLUSEAN which is "a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters" (Weber *et al.* 2009). CLUSEAN (CLUster SEquence ANalyzer) is an open source resource for semi-automatic analysis of secondary metabolite gene clusters. It is a modular framework

of Bioperl (Rausch *et al.* 2007) programs that are compatible with LINUX, UNIX, or MS Windows systems. It currently includes BLAST and HMMER as the tools for sequence annotation and domain identification respectively. CLUSEAN scripts search the NCBI non redundant protein database with BLASTp, and use HMMER to search the following Hidden Markov Models: Pfam domains, PKS/NRPS domains and motifs, and the C-domain types and NRPS adenylation domain models developed by Rausch and co-worker (Rausch *et al.* 2005; Rausch *et al.* 2007).

Another software pipeline probably the first and most recent of its type called antiSMASH (antibiotics and Secondary Metabolite Analysis Shell) (Medema *et al.* 2011) has been developed for identifying secondary metabolite biosynthesis gene clusters with the advanced features providing the analysis and annotation of the identified clusters. It serves as a meta server which amalgamates the data and methods available from various sources (Rausch *et al.* 2007; Ansari *et al.* 2008; Yadav *et al.* 2009; Weber *et al.* 2009; Letunic *et al.* 2009; Jong *et al.* 2010; Finn *et al.* 2010). It is capable of analyzing not only PKS/NRPS clusters as most of the software mentioned above do, but also for the identification of the biosynthetic loci for various other secondary metabolites as listed in Table 1.1. AntiSMASH can be accessed via a web server or it can be run as a standalone Java graphical user interface. It utilizes Glimmer3/GlimmerHMM for the gene prediction in the input sequence data and HMMER3 for the prediction of biosynthetic gene clusters using both existing profile HMMs as well as new profile HMMs from seed alignments. The substrate specificity of AT and adenylation domains were performed as proposed by (Yadav *et al.* 2003a) and (Röttig *et al.* 2011) respectively along with the method proposed by (Minowa *et al.* 2007) for both. The stereochemistry predictions for PKSs based on the Ketoreductase (KR) domain were carried out using the method used in the program ClustScan (Starcevic *et al.* 2008). To predict the biosynthetic order of PKS/NRPS modules antiSMASH uses the same method as SBSPKS to match the docking domain residues in the ORFs of type I modular PKSs, and otherwise assumes colinearity in the biosynthetic gene cluster. It also generates the SMILES string for the final predicted core chemical structure along with its picture. antiSMASH can also annotate the accessory genes by utilizing the HMMs constructed on smCOG (secondary metabolite clusters of orthologous groups). It also provides features like

ClusterBlast which can be used for comparative gene cluster analysis between the queried cluster and the clusters in the database. Utilizing CLUSEAN framework modules antiSMASH can also perform various other genome-wide analysis.