UNIVERSITY OF LEEDS

Cloning, expression and characterisation of the starter module from indanomycin biosynthesis

Sasha Rebecca Derrington

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Astbury Centre for Structural Molecular Biology

September 2014

The candidate confirms that the submitted work is her own and that the appropriate credit has been given within the thesis where reference has been made to the work of others. This copy has been supplied on the understanding that it is copyright material and that no quotation from this thesis may be published without proper acknowledgement.

Acknowledgements

Firstly, I would like to thank Professor Alan Berry for giving me the opportunity to work in this lab, allowing me to start something completely new, and for the support that he has provided me throughout the project, especially during the thesis dark days. I would also like to thank my second supervisor, Professor Adam Nelson for his supervision, guidance and encouragement.

I don't want to spoil the ending, but a big part of the results includes all of the work I have completed in the NMR spectroscopy facility. None of this would have been possible without the help and support of Dr Gary Thompson and also Dr Arnout Kalverda as well. I would like to thank my friend Dr Theo Karamanos, who has kept me calm and also helped me in NMR spectroscopy.

Next I would like to thank my friend Dr James Ault, who not only carried out the awesome mass spectrometry work but for our chats and being supportive, and I would like to thank Dr Chi Trinh for all his help with the crystallography. I would also like to thank Prof. Arwen Pearson and Prof. Sheena Radford for all their support and advice.

I have thoroughly enjoyed my four years working in the Astbury centre and have an amazing group of friends who have supported me, made me laugh and above everything made me happy. To Oli, Claire, Rachel, Lydia, Sophia, Jan, Hazel, Rob, Chris, Laura, Tom, Patrick, Paul and everyone else on level 10 (and mass spec), thanks for an amazing four years and putting up with me! And let's not forget the past Berry members and friends, whom I miss, Lucy Woods, Nicole Timms, Adam Daniels and Sophie Forrest.

Finally, I would like to thank my family, my Mum, Dad, Sister and Brother. To my Sister and Brother, thanks for putting up with me and the late phone calls because "I'm scared" and for supporting me. To my Mum and Dad, thank you for everything you have done and continue to do for me, you really are amazing parents and this would have been impossible without you. I really hope I've made you proud!

I would like to dedicate this thesis work to my Bubbe, Zayde, Grandma, Grandpa and the rest of my wonderful family for all of the love, support and encouragement they have always given me. I am so thankful for everything you have done for me and I hope I have made you proud.

Abstract

Nonribosomal peptides and polyketides form important classes of pharmaceutical agents. Several architectures of biosynthetic machinery exist for the construction of these structurally diverse molecules. Many attempts to engineer these proteins have concentrated on the multimodular type I polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS). Here, work was carried out to express the NRPS module responsible for starter unit formation from indanomycin biosynthesis in *E. coli*. Three synthetic genes, for IdmI, IdmJ and IdmK, were cloned and subsequently used in expression trials. Despite multiple attempts, IdmI was always expressed as an insoluble protein. IdmJ was expressed as a soluble protein, and an unexpected post-translational modification was found and investigated. Mutagenesis studies suggested that the unknown post-translational modification was occurring at a cysteine 127. IdmK, the carrier protein, was expressed in a soluble form with good yield. Analysis by mass spectrometry showed that, surprisingly, *E. coli* was able to phosphopantetheinylate IdmK, which is required for a functional module. Preliminary structure determination was carried out by X-ray crystallography. A complete 3D structure was obtained using NMR spectroscopy of the [^{15}N] and [^{13}C, ^{15}N] labelled protein. Structure determination was performed using CS-ROSETTA, which only uses chemical shift data, and ARIA, which assigns ambiguous NOE distance restraints. Both structure calculations produced comparable structures for IdmK. The structure showed a 3 α-helix bundle, with the same topology, lengths and locations of helices as other carrier proteins. Initial investigations into *holo*-IdmK suggest that the phosphopantetheine co-factor is directed into the hydrophobic core of the protein. This research has set up the system for future studies to engineer the pathway for novel product biosynthesis.

Contents

List of figures

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
ACP	Acyl carrier protein
APS	Ammonium persulphate
CS/δ	Chemical shift
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonuclease
dNTP	Deoxyribonucleotide triphosphate
DTT	Dithiothreitol
E. coli	*Escherichia coli*
ESI-MS	Electrospray ionisation mass spectrometry
FAS	Fatty acid synthase
LB	Luria broth
TB	Terrific broth
HSQC	Heteronuclear single quantum coherence
*idm*I	Prolyl dehydrogenase gene
IdmI	Prolyl dehydrogenase protein
*idm*J	Adenylyltransferase gene
IdmJ	Adenylyltransferase protein
*idm*K	Prolyl carrier protein gene
IdmK	Prolyl carrier protein protein
IPTG	Isopropyl β-D-1-galactosidase
LC-MS	Liquid chromatography mass spectrometry
MBP	Maltose binding protein
min	Minute
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect spectroscopy
NRPS	Nonribosomal peptide synthase
OD	Optical density
PCP	Peptidyl carrier protein

PCR	Polymerase chain reaction
PDB	Protein data bank
PKS	Polyketide synthase
PPant	Phosphopantheine
ppm	Parts per million
PPTase	Phosphopantetheinyl transferase
RDC	Residual dipolar coupling
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
BSA	Bovine serum albumin
S. antibioticus	*Streptomyces antibioticus*
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SNAC	*N*-acetylcysteamine thioesters
TCEP	Tris-(Carboxyethyl)phosphine, Hydrochloride
TEMED	N,N,N’,N’,-tetramethylethylenediamine
TOCSY	Total correlation spectroscopy
Tris	Tris(hydroxymethyl)aminomethane
tRNA	Transfer RNA
MS	Mass spectrometry/mass spectrum

1. Introduction

1.1 The Golden Age of Antibiotics

In 1928 Alexander Fleming made the serendipitous discovery of penicillin (Fleming, 1929). Following this, in the 1940's, Howard Florey and Ernst Chain started exploring the pharmaceutical benefits of penicillin, discovering its antibacterial properties (Chain *et al.*, 1940). Penicillin targets peptidoglycan cell wall biosynthesis of bacteria and is therefore most active against Gram positive bacteria (Yocum *et al.*, 1980). Between the discovery and development of penicillin, in 1935, scientists at the German company IG Farben discovered, developed and commercialised a synthetic sulfonamide drug named prontosil (4-[2,4-diaminophenyl)azo] benzenesulfonamide) for the treatment of Gram-positive bacteria by acting in a bacteriostatic manner and targeting folate synthesis (Domagk, 1935). These drugs sparked a cascade of discovery of a variety of pharmaceutically beneficial compounds; this is referred to as the "golden age" of antibiotics. The "golden age" of antibiotics was really in full stride between the 1950s and 1960s with a large percentage of antibiotics still in use today being discovered in this period (Davies, 2006). Figure 1 shows a timeline of antibiotics in use today, the majority of which were discovered during this era.

Since the discovery of antibiotics such as penicillin, overuse and misuse in the field of medicine and in intensive animal farming has caused a large increase in resistant strains of bacteria, rendering some of these well-established treatments generally ineffective. A combination of horizontal gene transfer and mutation has armed bacteria with an arsenal of mechanisms for antibiotic resistance (Davies, 2006). Some examples of mechanisms of resistance include: (1) efflux pumps which will remove the drug from the cell, this occurs in fluoroquinolone resistance (Hooper, 1999; Davies and Davies, 2010); (2) modification of the drug itself rendering it inactive, which plays a key role in penicillin resistance (Fisher *et al.*, 2005; Wright, 2007); and (3) alteration of cellular metabolic pathways, bypassing the process the drug has inhibited, this occurs in sulfonamide resistance (Skold, 2000).

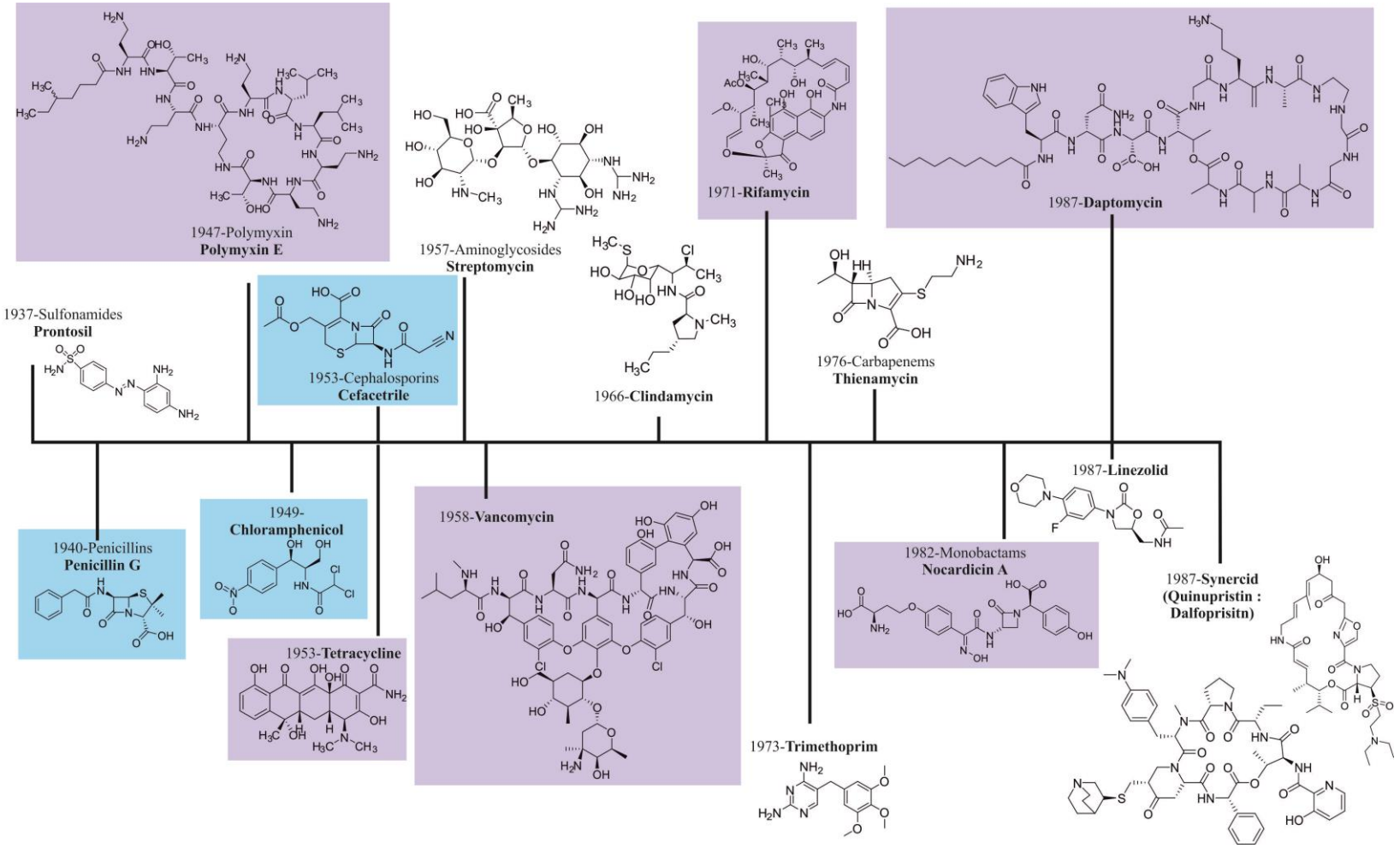


Figure 1.1- A timeline of the discovery of antibiotics that are still in use today. Drug classes (normal) and examples (bold) with chemical structures are shown above. Drugs highlighted in purple boxes are synthesised by non-ribosomal peptide synthetases (NRPS) and Type I modular polyketide synthases (PKS) (Section 1.4 and 1.5), those highlighted in blue boxes are partially synthesised by NRPS or PKS and those not highlighted are synthesised by other methods. Figure adapted from Davies (2006).

Some bacterial strains, such as *Staphylococcus aureus*, are known to be resistant to a number of antibacterial drugs and are therefore called multidrug resistant organisms (MDRO). Multidrug resistance (MDR) leaves only a very limited range of treatments available, such as vancomycin (Figure 1.1) for the treatment of the methicillin-resistant *Staphylococcus aureus* (MRSA) superbug (Alekshun and Levy, 2007). Now, with the "golden age" of antibiotics well and truly over, we have seen a decrease in the number of new antibiotics discovered and a combination of the financial drain and strict Food and Drug Administration requirements has caused a plummet in the investment the pharmaceutical industry (Davies, 2006). In 1990 there were 18 pharmaceutical companies researching antibiotic treatments; today, there are just four; AstraZeneca and GlaxoSmithKline (United Kingdom), Novartis (Switzerland) and finally Sanofi-Aventis (France) (Cooper and Shlaes, 2011).

Antibiotics are a vitally important element in the fight against infectious diseases and maintenance of a good quality of life. Natural products provide a wealthy source of antimicrobials and research into the production and engineering of novel antimicrobials may aid in the fight against, not only drug resistant bacteria, but also a whole host of life threatening conditions. Throughout this introduction I will be presenting examples of these antimicrobials and the reasoning and strategies behind engineering their producers for novel pharmaceutical compound biosynthesis.

1.2 Natural products as antibiotics

Analysis of the literature and public pharmaceutical documentation has highlighted 16 natural product derivatives across stages I, II and III of clinical trials (Butler *et al.*, 2013). Since 2000, over 50% of the drugs released onto the market were either natural products, or derived from natural products. From these 12 natural and natural based products, 9 are produced by actinomycetes and 3 are produced by fungi, both are prolific sources of natural products with pharmaceutical benefits (Butler *et al.*, 2013; Newman and Cragg, 2012).

Statistical analysis suggested that typically, pharmaceutical companies are able to screen tens of thousands of actinomycetes in one year; however undiscovered

antibiotics are likely to occur ≤1 in 10^7 from randomly screened soil actinomycetes. Additionally, only a microscopic percentage of the soil on earth has been screened for actinomycetes, which incidentally is not the only environment in which these pharmaceutically beneficial compounds are found (Baltz, 2008; Clardy *et al.*, 2006).

Efforts are being made to discover new antibiotics produced by Nature, identifying natural products that have otherwise been inaccessible or missed beforehand. Although previously the main source of antibiotics has been from *Streptomyces*, continuation of the traditional culturing methods are still aiding in the discovery of new actinomycete strains not only from soil, but also marine sediment (Clardy *et al.*, 2006; Rocha-Martin *et al.*, 2014). Ribosomal RNA (rRNA) sequencing has been exploited for investigating uncultivable bacteria. 16S rRNA is ubiquitous in bacteria and contains conserved regions that can be probed and variable regions, which enable identification of new bacterial strains. rRNA sequencing has highlighted the fact that only a small fraction of bacteria can be cultured using traditional methods (Clarridge, 2004; Relman, 1999; Clardy *et al.*, 2006). This indicates that other approaches are required to access this untapped resource (Clardy *et al.*, 2006). Strategies for discovering new natural products for antibiotics include heterologous expression of genes in a culturable host, metagenomics in combination with heterologous expression, typically using 16s rRNA, and manipulation of currently known producers of natural products, also known as combinatorial biosynthesis (Clardy *et al.*, 2006).

There are three classes of natural products with pharmaceutical benefits of particular interest: polyketides (PK), non-ribosomal peptides (NRP), and hybrid PK-NRP compounds which will be discussed further.

1.2.1 Polyketides

Polyketides are secondary metabolites known to possess a wealth of pharmaceutically beneficial activities such as antibacterial, antifungal, antiparasitic, antihelminthic and insecticidal properties (Kevin Ii *et al.*, 2009; Dutton *et al.*, 1995). The main producers of these biologically active molecules are the actinomycetes especially *Streptomyces* and *Saccharopolyspora* (Weissman and Leadlay, 2005).

Figure 1.2 shows example structures of polyketides (and rifamycin in Figure 1.1) illustrating the structural diversity seen across these compounds (Staunton and Weissman, 2001).

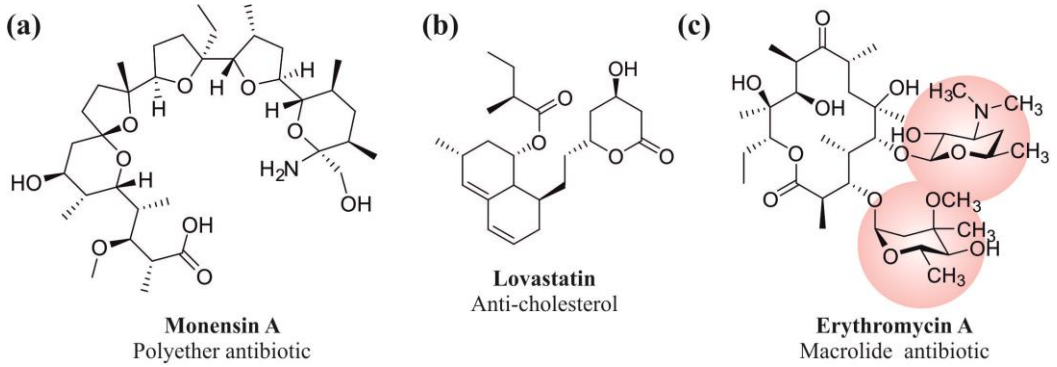


Figure 1.2- Examples of the secondary metabolites, polyketides. (a) Monensin produced by *Streptomyces cinnamonensis*, is a polyether antibiotic (Oliynyk *et al.*, 2003). (b) Lovastatin, produced by *Aspergillus terreus*, is used to treat high cholesterol (Alberts *et al.*, 1980). (c) Erythromycin A, a macrolide antibiotic is produced by *Saccharopolyspora erythraea*. The sugar moieties highlighted are added by auxiliary enzymes after the release from the PKS (Weber *et al.*, 1990; Labeda, 1987; Dhillon *et al.*, 1989).

Polyketides are composed mainly of three basic units; acetate, propionate and butyrate, derived from their CoA equivalents. These units are also utilised by fatty acid synthases for the synthesis of fatty acids and so are readily available in the cell (Figure 1.5) (Staunton and Weissman, 2001; Meier and Burkart, 2009). The structural diversity, wealth of stereochemistry and functional groups observed in polyketides can partly be attributed to the various starter and extender units that can be incorporated (Chan *et al.*, 2009; Smith and Tsai, 2007; Cane *et al.*, 1983).

Polyketides can be classed by their structure and how they were made, and can be divided into unreduced, reduced and unclassified polyketides. Reduced polyketides can be further divided into two different classes; the macrocycles which have large lactone or lactam rings (such as erythromycin A) or polyether antibiotics, these contain between 2 and 5 ether oxygen atoms (for example monensin A) (Figure 1.2) (Staunton and Wilkinson, 1997; Oliynyk *et al.*, 2003; Weissman and Leadlay, 2005). The biological actions carried out by polyketides vary and are dependent on their

structure. Polyether ionophores, such as monensin, have the ability to chelate metal ions and disrupt ion balance in order to kill an organism. This involves transporting metal ions across a membrane either by: (1) binding to an ion in a polar cavity and concealing the charge, ferrying it across the plasma membrane, (2) forming a channel allowing ions to flow more freely into the cell, and finally (3) neutral ionophores aid in the diffusion of ions into a cell (Pressman, 1976; Pressman and Fahim, 1982). Macrolides such as erythromycin, which are constrained by macrocyclisation, work by inhibiting protein synthesis by premature chain termination (Menninger and Otto, 1982; Kohli and Walsh, 2003).

Total synthesis, a well-studied area, makes use of well-known and studied reactions such as the carbon-carbon bond forming aldol reactions to produce fragments of polyketide (Ireland *et al.*, 1985; Faul and Huff, 2000; Schetter and Mahrwald, 2006). Challenges arise in linking such fragments with the correct stereochemistry thus making production of more diverse analogues difficult. Polyketides and specifically ladder polyethers become cyclic by epoxide-opening cascades, and, for total synthesis, work needs to be done on selective activation for these cascades to occur. Additionally efforts need to be made towards facile modifications of epoxide groups including being amenable to methyl substitutions (Ireland *et al.*, 1985; Vilotijevic and Jamison, 2009; Faul and Huff, 2000). In Nature polyketides are produced by giant multimodular proteins termed polyketide synthases (PKS) (Gallimore, 2009; Liu *et al.*, 2009). These macromolecular structures have the capacity to synthesise polyketides and modify them accordingly to efficiently produce complex structures with the correct stereochemistry (Staunton and Weissman, 2001; Weissman and Leadlay, 2005) (Section 1.4).

1.2.2 Nonribosomal peptides

Nonribosomal peptides (NRPs) are another class of natural products produced by bacteria, filamentous fungi and more generally actinomycetes (Perlman and Bodanszky, 1971). NRPs, as implied by the name, are synthesised independently of the ribosome and a nucleic-acid template (Finking and Marahiel, 2004; Grunewald

and Marahiel, 2006). As with polyketides, nonribosomal peptides are a group of structurally diverse compounds with a large repertoire of activities such as antibacterial, immunosuppressive and cytostatic (Schwarzer *et al.*, 2003).

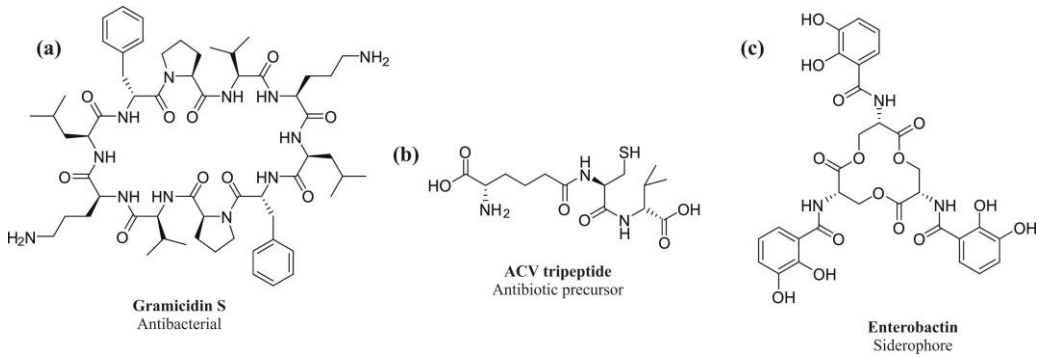


Figure 1.3- Examples of nonribosomal peptides. (a) Gramicidin S, produced by *Bacillus brevis*, has been shown to have antibacterial properties (Gause and Brazhnikova, 1944). (b) ACV tripeptide, is the precursor to penicillin and cephalosporin antibiotics (Martin *et al.*, 1994; Keeler, 2002). (c) Enterobactin is a siderophore which chelates iron, it is found in some Gram-negative bacteria such as *E. coli* (Raymond *et al.*, 2003; Fischbach and Walsh, 2006).

NRPs are known to have several different modes of action; polymyxin antibiotics (Figure 1.1) target and bind to lipopolysaccharide (LPS) in the outer membrane of Gram-negative bacteria, using their hydrophobic tails to disrupt inner and outer membranes (Velkov *et al.*, 2010). Gramicidin S (Figure 1.3) is a cationic compound and the suggested mode of action is that it is taken up into the cell and disrupts the cytoplasmic membrane; polymyxins are also suggested to do this (Katsu *et al.*, 1986; Velkov *et al.*, 2010). Vancomycin (Figure 1.1) targets cell wall biosynthesis (Watanakunakorn, 1981). Siderophores, such as enterobactin (Figure 1.3), are essential for host survival and are released by bacteria in environments where iron is scarce as it has a picomolar affinity for the metal ion (Raymond *et al.*, 2003).

NRPs are formed by oligomerisation through peptide bond formation of between 3 and 22 proteogenic, non-proteogenic amino acids and aryl acids (Sieber and Marahiel, 2003; Schwarzer *et al.*, 2003). NRPs can undergo macrocyclisation such as

gramicidin S, or be formed from multiple copies of identical NRP chains like enterobactin (Figure 1.3) (Rusnak *et al.*, 1991; Walsh *et al.*, 1990). The macrocyclisation and oxidative cross-linking seen in NRPs aids in their biological functions. NRPs can be further modified by introducing small heterocyclic rings, *N*-methylation, *N*-formylation, glycosylation or by the addition of fatty acids (Walsh *et al.*, 2001; Sieber and Marahiel, 2005). These structurally elaborate compounds are synthesised by the multimodular enzymes, nonribosomal peptide synthetases (NRPS) (Section 1.5).

1.2.3 Polyketide-nonribosomal peptide hybrid products

Nonribosomal peptides and polyketides are synthesised in a very similar way, by multimodular enzymes in an assembly line-like manner utilising an enzyme template-directed mechanism (Lipmann, 1980; Hopwood, 1997; Marahiel *et al.*, 1997). Hybrid polyketide-polypeptide products (Figure 1.4) are composed of a combination of monomeric units utilised in both polyketides and nonribosomal peptides, again creating structurally diverse biologically active compounds (Du *et al.*, 2001).

Nonribosomal peptide-polyketide hybrid products are synthesised by a combination of nonribosomal peptide synthetase modules and polyketide synthase modules. These products can be synthesised in one of two ways; by individual NRPS and PKS modules, coupling the NRP and PK moieties using a ligase, for example coronatine biosynthesis (Bender *et al.*, 1999; Rangaswamy *et al.*, 1998) or by a hybrid NRPS-PKS enzyme, as with epothilone biosynthesis (Molnar *et al.*, 2000). The latter will be discussed below.

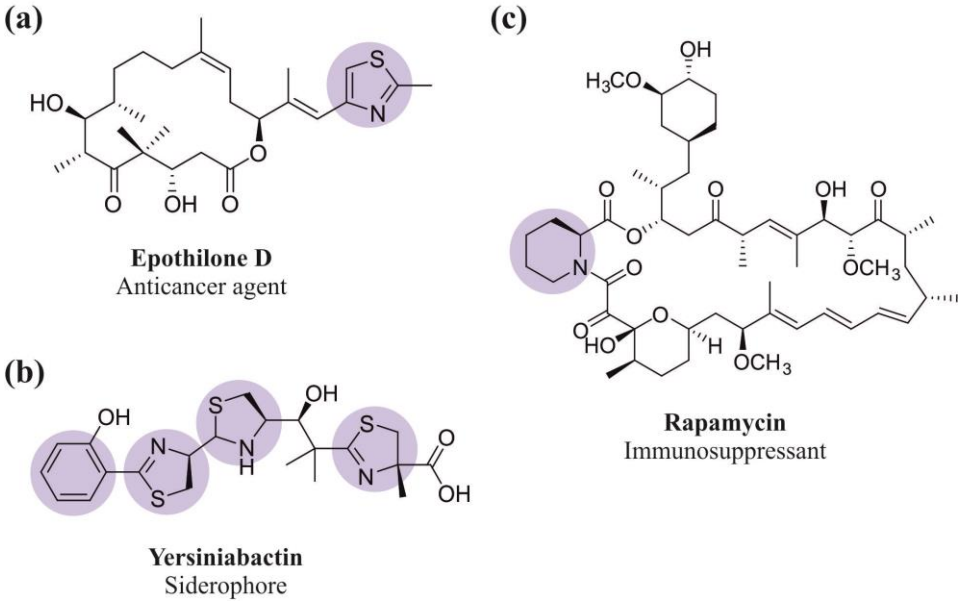


Figure 1.4- Examples of hybrid polyketide-nonribosomal peptides. (a) Epothilone is an NRP-PK synthesised by an NRPS/PKS hybrid enzyme produced by *Sorangium cellulosum* (Molnar *et al.*, 2000). (b) Yersiniabactin is also synthesised by an NRPS/PKS hybrid enzyme found in *Yersinia pestis* (Pelludat *et al.*, 1998). (c) Rapamycin produced by *Streptomyces hygroscopicus* where a PK chain is modified by nonribosomal synthetase to initiate macrocyclisation creating an NRP-PK hybrid product (Molnar *et al.*, 1996; Park *et al.*, 2010; Schwecke *et al.*, 1995). Groups highlighted in blue are incorporated by NRPS.

1.3 Polyketide synthases

There are three classes of polyketide synthases; I, II and III. Type I PKS can be subdivided into two types, iterative and modular. Type I modular polyketide synthases (PKS) are large multimodular enzymes composed of discrete domains covalently linked into modules often on a MDa scale. These enzymes, encoded either by a single operon or gene clusters in bacteria, fungi, plants and more infrequently animals, are responsible for the production of polyketides as secondary metabolites as mentioned previously (1.2.1). In type I modular PKSs a module describes a group of covalently linked discrete domains, with specific roles, responsible for incorporation, and in some cases modification, of a single extender unit to a growing polyketide chain (Staunton and Weissman, 2001). Iterative type I PKS have a similar architecture to a single module in type I modular PKS that is used in continuous rounds of elongation. Examples of type I iterative PKS include the PKS responsible for lovastatin synthesis (Figure 1.2) (Staunton and Weissman, 2001; Campbell and

Vederas, 2010). Type II PKS are composed mainly of monofunctional proteins i.e. two ketosynthase domains (KS) and an acyl carrier protein (ACP), chain elongation occurs in iterative cycles producing aromatic polyketides with additional tailoring domain such as oxygenases and methyl transferases increasing structural diversity in compounds produced. Type II PKS are the source of important pharmaceutical compounds such as tetracyclines (Hertweck *et al.*, 2007; McDaniel *et al.*, 2005). Type III PKS are composed of a KS homodimer that works in an iterative fashion to condense together monomeric units. Type III PKS are usually found in plants and produce compounds such as chalcones (Abe and Morita, 2010; Katsuyama and Ohnishi, 2012).

This research is intended to focus solely on the modular nature of proteins and how they construct such diverse products, with the hopes of engineering the modular enzymes to manufacture novel compounds. Therefore the rest of the introduction will focus on type I modular PKS and NRPS (section 1.4).

1.3.1 PKS biosynthetic gene clusters

In 1990 Peter Leadlay's group cloned a 10 kb gene fragment from *Saccharopolyspora erythaea* encoding the modular polyketide synthase responsible for erythromycin biosynthesis. This was achieved by identifying the gene which encoded erythromycin resistance and sequencing from that point using chromosome walking, making comparisons with sequences of homologous proteins. The method used was based on the fact that the genes which encode the proteins that make up bacterial and fungal PKSs are clustered together within the genome, along with transcriptional regulators and self-resistance genes (Cortes *et al.*, 1990). At approximately the same time Leonard Katz and his group also cloned the DEBS biosynthetic gene cluster using the same principles as Leadlay but working from a mutation within a gene which disabled erythromycin production (Tuan *et al.*, 1990). Figure 1.8 illustrates the biosynthetic gene cluster of 6-deoxyerythronolide synthase (DEBS). The catalytic domains of DEBS were determined by comparison with homologous domains from FAS with identical activities (Bevitt *et al.*, 1992; Cortes *et al.*, 1990; Tuan *et al.*, 1990; Donadio *et al.*, 1991). Characterisation of the

biosynthetic gene cluster of DEBS has provided some insight into gene organisation in PKS.

With decreasing costs of genome sequencing and advances in *in silico* analysis, more than 50 PKS gene clusters have been elucidated (Weissman and Leadlay, 2005). Currently, common techniques used to identify PKS gene clusters include entire genome sequencing and screening cosmid or bacterial artificial chromosome libraries with probes based on homologous sequences, such as the KS domain, using hybridisation or PCR based methods (Weissman and Leadlay, 2005; Wang *et al.*, 2011).

Entire genome sequencing has highlighted an unexpected prevalence of potential genes for antibiotic synthesis in bacteria, entire genome sequencing of the *Bacillus subtilis* genome revealed more than 4% is suspected to encode for gene clusters responsible for polyketide and bacteriocin synthesis (Kunst *et al.*, 1997). Sequencing of certain *Streptomyces* strains have shown that the abundance of PKS and NRPS gene clusters is not unique to *B. subtilis* (Bentley *et al.*, 2002; Ikeda *et al.*, 2003). The elucidation of increasing numbers of PKS gene clusters is providing ever more options for engineering to produce novel compounds.

1.3.2 PKS module organisation

PKSs are composed minimally of three essential domains, an acyltransferase domain (AT), responsible for the substrate selected as the extender unit; a ketosynthase domain (KS) which catalyses the condensation between the selected extender unit and the growing polyketide chain, and an acyl carrier protein (ACP) which houses the growing chain (Staunton and Weissman, 2001; Weissman and Leadlay, 2005; Menzella *et al.*, 2005; Fischbach and Walsh, 2006). Polyketides are synthesised in an assembly line-like fashion with each individual module responsible for incorporating a single unit, and the number of modules dictating the number and specificity of units incorporated (Chan *et al.*, 2009; Staunton and Weissman, 2001; Fischbach and Walsh, 2006). Figure 1.6 illustrates the composition of the minimal module and mechanism by which chain elongation occurs.

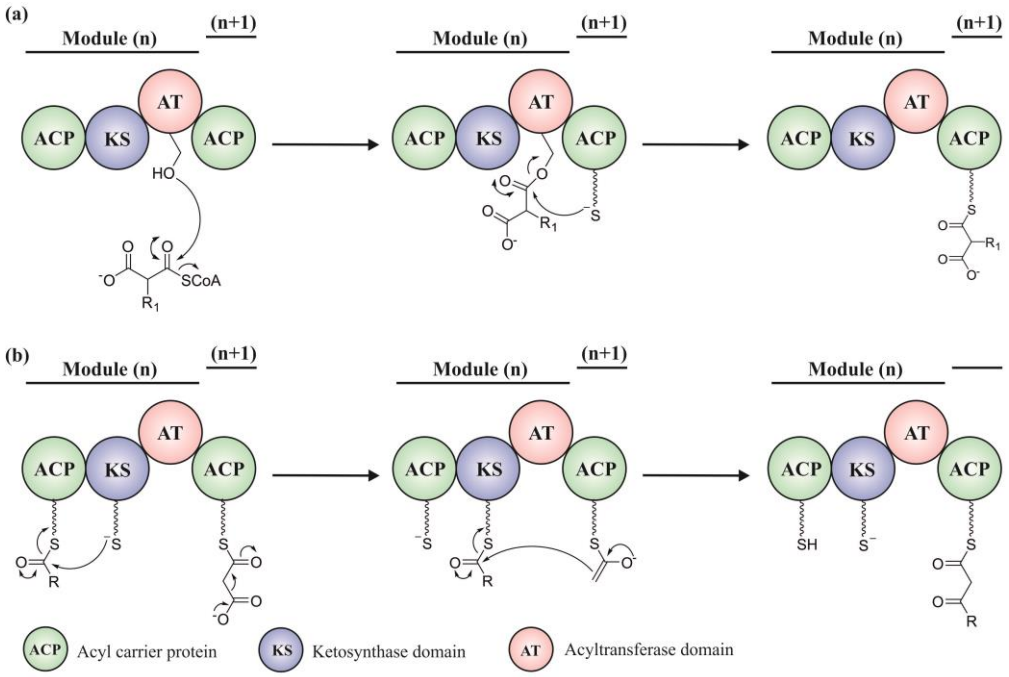


Figure 1.5- Minimal PKS module illustrating the mechanism by which chain elongation occurs. (a) The acyltransferase (AT) domain selects an acyl CoA extender unit and catalyses its transfer onto the acyl carrier protein (ACP). (b) Carbon-carbon bond formation by decarboxylative condensation then occurs between the extender unit and growing polyketide chain catalysed by the ketosynthase (KS) domain (Fischbach and Walsh, 2006).

1.3.2.1 Acyltransferase (AT) domain

The acyltransferase (AT) domain, originally identified due to its homology with the malonyl acyltransferase (MAT) in FAS, is responsible for substrate selection and is suggested to possess high substrate specificity (Khosla *et al.*, 1999; Walsh, 2008; Liou *et al.*, 2003). The reaction catalysed by an AT domain is shown in Figure 1.5. The site of acylation is a conserved serine (Dunn and Khosla, 2013). There are two types of AT domains known to be utilised by PKS, a *cis* AT domain which associates with a single specific ACP, as illustrated in the DEBS PKS (Figure 1.8) (Dunn *et al.*, 2014). Alternatively, there are *trans* acting AT domains, these domains are stand-alone and associate with multiple ACPs for loading of the extender unit as occurs with leinamycin biosynthesis (Piel, 2010; Cheng *et al.*, 2003).

A highly conserved arginine at position 222, or a tryptophan in the DEBS and avermectin PKS, has been identified as controlling substrate specificity (Liou *et al.*, 2003; Rangan and Smith, 1997). Substrate specificity of AT domains within

initiating modules varies; in contrast AT domains in the extending modules are more substrate specific acting as a gate-keepers, stopping chain elongation in the absence of the correct extender unit (Keating and Walsh, 1999; Khosla *et al.*, 1999). The AT domain provides an excellent opportunity for engineering to introduce different extender units for novel product biosynthesis.

1.3.2.2 Ketosynthase (KS) domain

The ketosynthase (KS) domain, approximately 45 kDa in mass, catalyses the decarboxylative condensation of the extender unit and growing chain. A conserved cysteine in the active site enables the extender unit to be briefly transferred onto the KS domain before being added to the growing polyketide chain and contributes to the substrate specificity of the module in some way (Fischbach and Walsh, 2006; Chen *et al.*, 2006). The KS domain is either inactive or missing completely in the initiating modules (Chen *et al.*, 2011). The KS domain has also been pinpointed to be responsible for the wealth of stereochemistry observed in polyketides (Lau *et al.*, 1999).

1.3.2.3 Acyl carrier protein

Acyl carrier proteins (ACP) are essential non-catalytic domains of less than 100 amino acids in length. They are the site of chain elongation and are known to interact with more than 30 partners (Fischbach and Walsh, 2006; Gully and Bouveret, 2006; Gully *et al.*, 2003; Butland *et al.*, 2005). Phosphopantetheinylation of carrier proteins is unconditionally required for biosynthesis of not only polyketides, but nonribsomal peptides and fatty acids. This post translational modification is mediated by a phosphopantetheinyl transferase (PPTase). Each enzyme complex; PKS, NRPS and FAS, is associated with a PPTase capable of this modification. Phosphopantetheinylation occurs on a conserved serine within the ACP, with the phosphopantetheine group from Coenzyme A being covalently attached by a phosphodiester link (Figure 1.6). The phosphopantetheine modification creates a flexible arm of approximately 20 Å in order to pass the growing chain from one module to another (Walsh *et al.*, 1997; Lambalot *et al.*, 1996; Quadri *et al.*, 1998; Staunton and Weissman, 2001).

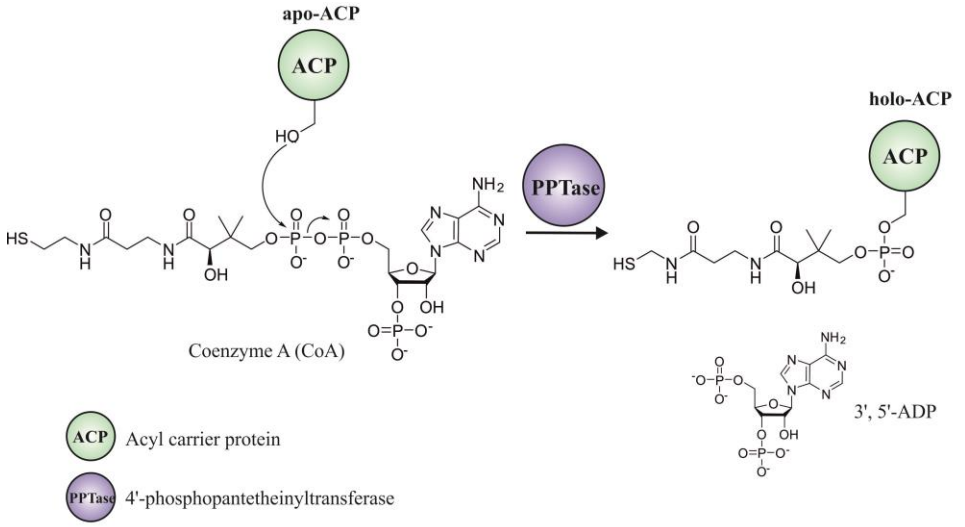


Figure 1.6- Mechanism by the ACP becomes phosphopantetheinylated. The hydroxyl group of a conserved serine of an acyl carrier protein (ACP) becomes phosphopantetheinylated by a phosphopantetheinyltansferase mediating the formation of a phosphodiester link, converting *apo*-ACP to *holo*-ACP (Walsh *et al.*, 1997; Lambalot *et al.*, 1996; Fischbach and Walsh, 2006).

1.3.2.4 Chain initiation and termination

As well as individual modules responsible for extender unit incorporation, there are two additional features, a loading module composed of an AT domain and ACP responsible for initiation and a thioesterase domain responsible for chain termination.

More often than not the module responsible for polyketide chain initiation lacks the KS domain. In the initiation module, the AT domain is sufficient for loading the ACP with a starter unit before being handed off to the next module, this occurs in the DEBS PKS (Figure 1.8) (Keating and Walsh, 1999; Caffrey *et al.*, 1992).

The thioesterase (TE) domains are responsible for chain termination and are located at the C-terminus of the PKS. Chain termination occurs using a Ser-His-Asp catalytic triad in an analogous manner to serine proteases. Depending on the nature of the TE domain, macrocyclisation can occur of the polyketide product being released. If the acyl-bound intermediate is attacked by an external nucleophile a linear product will be released, attack by a nucleophile within the polyketide chain will lead to a cyclised product (Du and Lou, 2010).

1.3.2.5 Linker regions

Assembly line synthesis by a multimodular protein requires the modules to be tethered together in some way, and additionally for loading and catalysing condensation between two acyl moieties, so will the globular domains. Therefore linker regions between both the individual domains and entire modules exist. Linkers between domains are up to 100 residues long and composed mainly of alanine, proline and charged residues, the composition of the linkers dictating their flexibility and thus the range of motion and interaction with other domains (Staunton and Wilkinson, 1997; Bevitt *et al.*, 1992).

1.3.2.6 Accessory domains: KR, DH and ER

The structural diversity observed in polyketides is also achieved by the addition of accessory domains within a module. These domains control the level of reduction of the β-keto group of the extender unit added, and are termed a ketoreductase (KR), dehydratase (DH) and enoylreductase (ER). Unlike FAS, which have all three domains present achieving full reduction of all β-keto groups, PKS modules display varying levels of reduction at the β-keto group dependent on which of the accessory domains are present. Figure 1.7 illustrates how the β-keto carbon is processed in PKS and how it is comparable to FAS.

1.3.2.7 Biosynthetic pathways of polyketide synthases and fatty acid synthases

The idea of small monomeric units being condensed together by an enzyme template-directed mechanism is utilised by multimodular enzymes to produce structurally diverse compounds. This mechanism is common to fatty acid synthases (FAS), polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS) (Meier and Burkart, 2009). FAS and PKS utilise the same monomeric units and Figure 1.7 illustrates the units used by FAS and PKS and the differences in levels of reduction the acyl-CoA units undergo. FAS acyl-CoA monomers undergo reduction, dehydration and a second reduction; the acyl and malonyl CoA units are condensed together in a head-to-tail manner to create fatty acids. Polyketide acyl-CoA monomers in a polyketide chain can undergo varying levels of reduction contributing to the structural diversity observed in polyketides, dependent on the presence of the

accessory domains within a PKS module (Meier and Burkart, 2009; Staunton and Weissman, 2001).

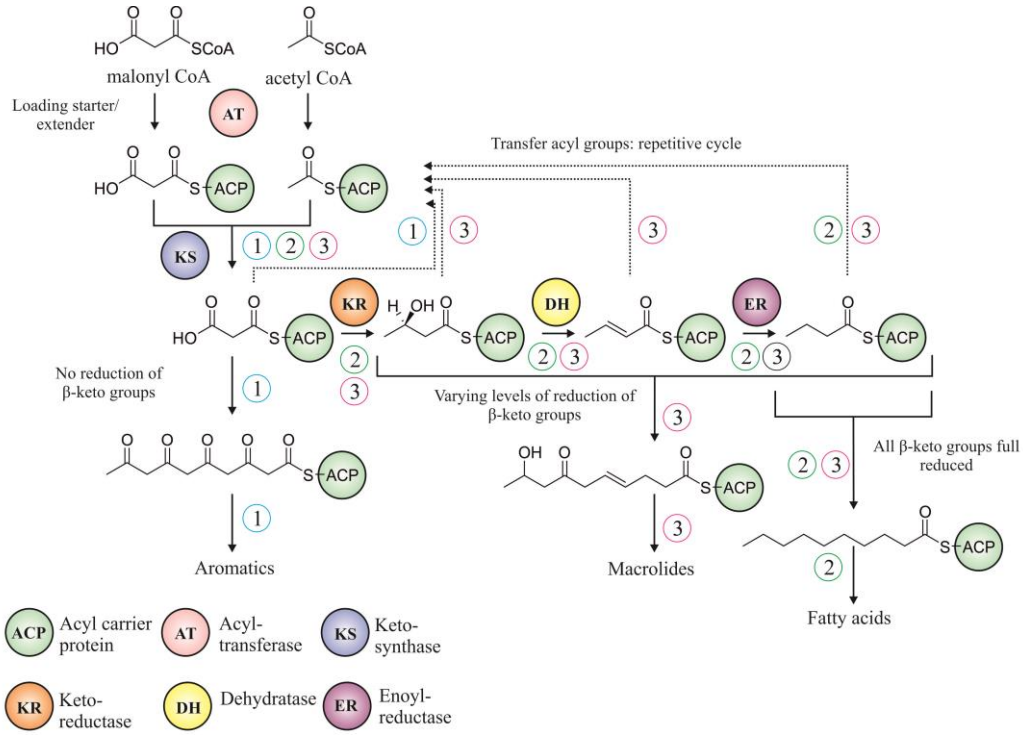


Figure 1.7- The biosynthetic pathways of FAS and PKS. FAS and PKS acyl-CoA monomers are condensed together in an identical manner. An extender unit is selected by acyltransferase (AT) domains, which catalyse the transfer onto an acyl carrier protein (ACP), the site of chain elongation. The ketosynthase (KS) domain then catalyses carbon-carbon bond formation between the extender unit and the growing acyl chain. FAS (2) then utilises a ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER), to fully reduce the acyl unit added. PKS acyl units undergo varying levels of reduction at the β-keto position (3); this is dependent on the presence of KR, DH and ER domains present. No reduction results in aromatic polyketides being formed (1) (Moore and Hertweck, 2002; Meier and Burkart, 2009).

In fatty acid biosynthesis (Figure 1.7 (2)) (as with PK biosynthesis) the FAS AT domain selects the extender unit, and a ketosynthase domain catalyses the carbon-carbon bond formation between the selected unit and the growing fatty acid. The β-keto carbon of the malonyl CoA or acetyl CoA unit added then undergoes reduction by a ketoreductase (KR) domain to a β-hydroxy group, this is followed by dehydration by a dehydratase (DH) yielding an α,β-enoyl product. Finally, a second round of reduction is carried out by an enoylreductase (ER) domain, to give a fully reduced product. In bacterial and mammalian FAS, more units are added in an

iterative fashion giving full length fatty acids (Meier and Burkart, 2009; Staunton and Weissman, 2001). Type II PKS are iterative enzymes which produce aromatic PKs (Figure 1.7 (1)), this occurs when the accessory domains are absent (Das and Khosla, 2009). In type I modular PKSs differential levels of reduction occur dependent on which accessory domains are present. In addition to the essential domains within a module, in type I modular PKS, varying levels of accessory domains can exist (Figure 1.7 (3). The level of reduction the extender units undergo in each round of elongation is dependent within that module and is independent of preceding and subsequent modules, producing compounds composed of extender units differing at the β-keto position (keto, hydroxyl, enoyl and fully reduced) (Staunton and Weissman, 2001; Meier and Burkart, 2009). These domains provide another possible dimension for engineering PKS to produce novel compounds.

Figure 1.8 shows the biosynthetic gene cluster of the PKS DEBS and gives some idea as to the arrangement of domains within modules in PKSs, illustrating the combinations of accessory domains within different modules (Caffrey *et al.*, 1992; Bevitt *et al.*, 1992; Cortes *et al.*, 1990; Tuan *et al.*, 1990).

1.3.2.8 A model system

Extensive research has been carried out on the PKS responsible for erythromycin production, 6-deoxyerythronolide (DEBS). Both structural and mechanistic studies have provided great insight into how this PKS works and engineering has provided tantalising glimpses into the potential for novel compound production. DEBS has been the model system and so the biosynthetic gene cluster is shown in Figure 1.8 (Bevitt *et al.*, 1992; Cortes *et al.*, 1990; Caffrey *et al.*, 1992; Staunton and Weissman, 2001; Staunton *et al.*, 1996; Marsden *et al.*, 1998; Oliynyk *et al.*, 1996; Khosla *et al.*, 2007; Khosla *et al.*, 2014).

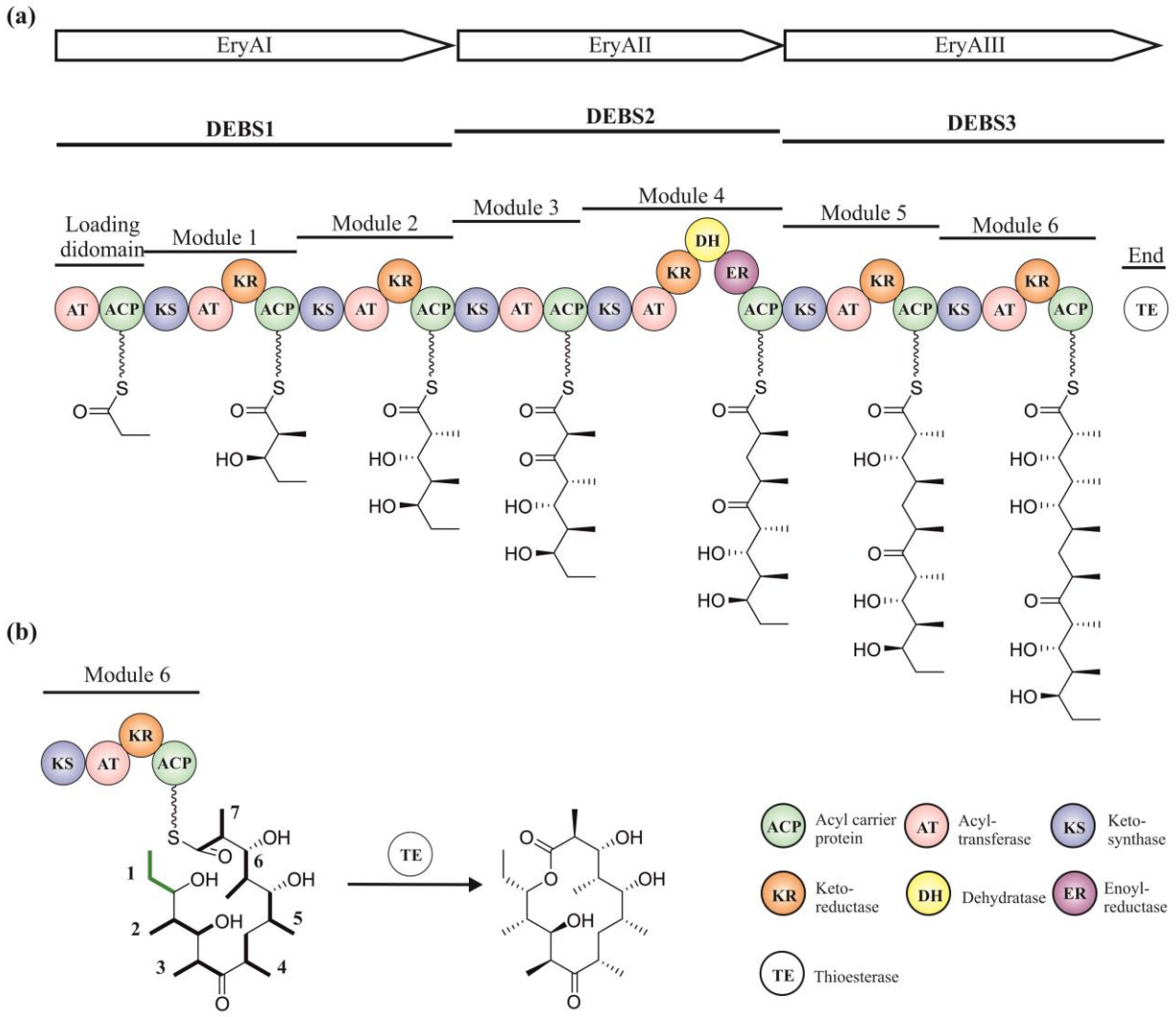


Figure 1.8- Biosynthetic gene cluster of the erythromycin precursor, 6-deoxyerythronalide, producing PKS, DEBS. (a) eryAI, II and III open reading frames encode three large polypeptides, DEBS 1, 2 and 3. DEBS 1 is composed of a loading didomain, module 1 and module 2, DEBS2 is composed of modules 3 and 4 and finally DEBS3 is composed of modules 5 and 6 and the thioesterase domain. (b) The thioesterase domain is not only involved in chain termination but macrocyclisation of the linear polyketide chain. Units incorporated are highlighted in bold on the structure on the left (Caffrey *et al.*, 1992).

DEBS forms the macrolide intermediate 6-deoxyerythronolide B (6-dEB) (Figure 1.8). Three open reading frames, eryAI, eryAII and eryAIII, encode for the polypeptides DEBSI, DEBSII and DEBSIII respectively. DEBSI contains the loading module and two extender modules, DEBS II contains two extender modules and DEBS III contains two extender modules and the thioesterase. The molecular mass of this PKS is over 600 kDa (Cortes *et al.*, 1990; Tuan *et al.*, 1990; Kao *et al.*, 1994; Caffrey *et al.*, 1992). The macrolide intermediate is formed by 6 modules responsible for incorporation of ethyl malonyl CoA as extender units, a loading di-domain which lacks a KS domain and utilises propionyl CoA as the starter unit and finally a thioesterase domain which instigates chain release and aids in cyclisation of 6-dEB (Tuan *et al.*, 1990; Cortes *et al.*, 1990; Caffrey *et al.*, 1992).

1.3.3 Structure determination of PKSs

The modularity of PKSs makes them an attractive target for combinatorial biosynthesis to produce novel compounds. Structural characterisation and mechanistic understanding will aid engineering of PKS to produce novel compounds.

There have been significant efforts to structurally characterise the domains within a PKS module, including di-domains and docking domains (between modules). Several structures exist of KS-AT didomains from DEBS, the KS-AT didomain from module 5 (PDB code 2HG4) (Figure 1.9) and the KS-AT didomain from module 3 (PDB code 2QO3). Structural information revealed well-structured inter- and intradomain linkers composed of highly conserved residues and a loop region in the KS domain homodimer interface which potentially plays a part in substrate specificity. Additionally, an 80 Å distance between the catalytic cysteine in the KS domain and the conserved serine in the AT domain was observed indicating that the DEBS PKS must undergo structural rearrangement for chain elongation to occur (Tang *et al.*, 2007; Tang *et al.*, 2006).

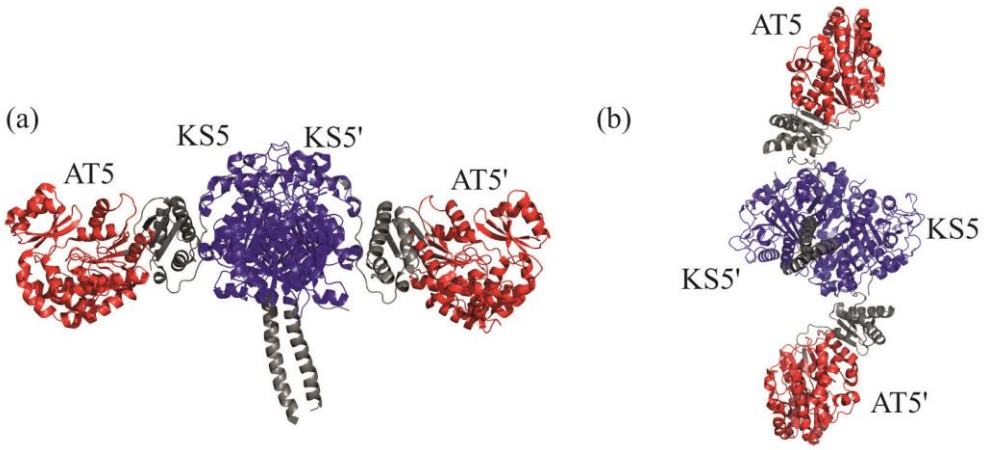


Figure 1.9- The structure of the KS-AT didomain of module 5 from DEBS. Structure solved to a resolution of 2.7 Å. The KS domain (blue) and AT domain (red) are connected by highly structured linker regions (grey). (a) Side view of didomain, showing the extension of the N terminal docking domain (grey). (b) Front view illustrating the homodimer interface is mediated by the KS domains and N terminal docking domain (Tang *et al.*, 2006).

NMR and crystal structures of DEBS and pikromycin docking domain or linkers have been solved. Linkers between modules are structurally well defined and are composed of highly conserved residues which not only promote dimerisation of domains but create unfavourable interactions with other modules so as to prevent misdocking (Broadhurst *et al.*, 2003; Buchholz *et al.*, 2009).

The ketoreductase, dehydratase and thioesterase domains have been well characterised with crystal structures from a number of PKSs (Zheng *et al.*, 2012; Keatinge-Clay, 2008; Keatinge-Clay, 2007; Keatinge-Clay and Stroud, 2006; Tsai *et al.*, 2001; Akey *et al.*, 2006; Scaglione *et al.*, 2010; Zheng *et al.*, 2010; Zheng and Keatinge-Clay, 2011; Gehret *et al.*, 2011; Tsai *et al.*, 2002). Less well characterised, the enoyl reductase, has been solved as a di-domain (ER-KR) (Zheng *et al.*, 2012). Finally, a structure of an ACP from DEBS has been solved by NMR spectroscopy (Alekseyev *et al.*, 2007). Details of the structure of carrier proteins will be covered later in this thesis. Even though the structures of individual domains are well documented there is no high resolution structure of an entire PKS module.

In 1996 Staunton, Leadlay and colleagues proposed an overall model for the DEBS PKS referred to as the "Cambridge model" (Figure 1.10) (Staunton *et al.*, 1996). The foundations of this model were structural studies utilising limited proteolysis and analytical ultracentrifugation (AUC), which showed that KS-AT and TE domains are homodimeric whilst those domains involved in reduction are monomeric. In this model the PKS forms a "double-helical" structure with proteins intertwined forming a homodimer with both chains in the same orientation ("head-to-head") and the essential domains (AT-KS-ACP) forming the core of the complex. The KS domain of one module was shown to have a favourable interaction with the ACP of the opposite module; however, all domains appear to have access to the ACP. The accessory domains then loop out from the main body of the PKS. These models satisfy the ability to vary the accessory domains within different modules, and allow communication between subunits but prevent cross-talk of ACPs with the incorrect domains. This model also supported the seemingly more successful engineering of accessory domains. A similar model was outlined two years later by Cane and Khosla (Cane *et al.*, 1998).

Until recently structural studies involving entire PKS modules have been hindered by their sheer size and flexibility (Tsai and Ames, 2009; Staunton and Weissman, 2001). However, in 2008 advances in the field of X-ray crystallography enabled elucidation of the crystal structure of the mammalian FAS (mFAS) (PDB code 2V28). The mFAS displayed an intertwined homodimer in an X orientation with KS, DH and ER domains forming the dimer interface (Maier *et al.*, 2008). The bottom portion of the X was composed of the essential catalytic domains (KS and MAT) and the top was composed of the domains responsible for reduction (KR, DH and ER) (Maier *et al.*, 2008). As mentioned previously PKS are homologous to FAS and so this gave the best indication of the arrangement of domains within a module until recently (Dutta *et al.*, 2014).

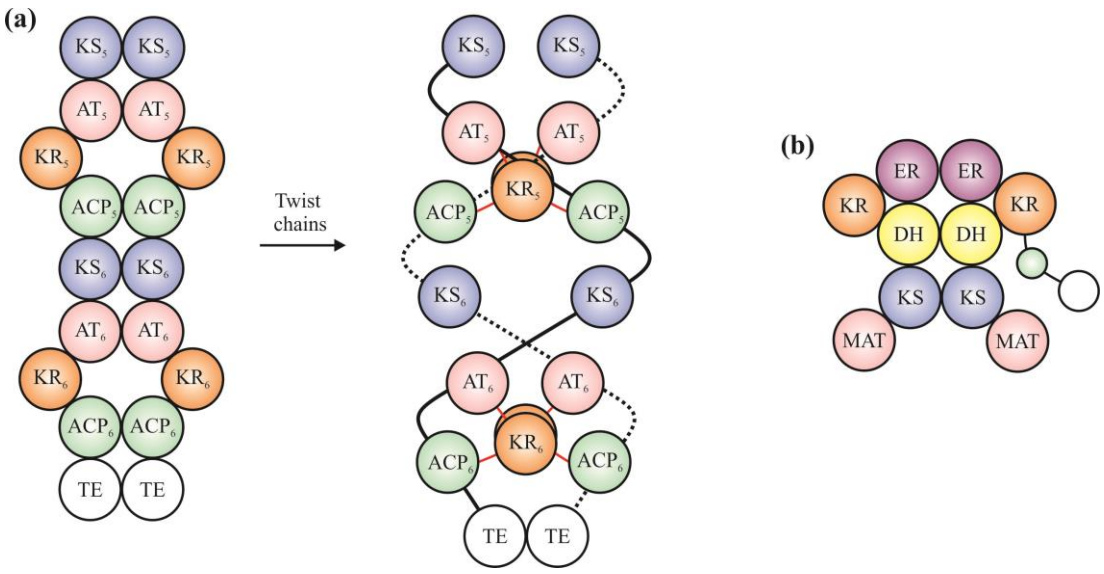


Figure 1.10- Proposed structure of PKS and mFAS (a) Schematic of modules 5 and 6 of DEBS illustrating the "Cambridge model" proposed by Staunton and Leadlay. The two chains are twisted together in a "head-to-head" orientation forming a helix. This helical conformation allows interactions between the KS domain and ACP from each chain. The KR domains loop out between the AT domain and ACP within the same monomer (red linkers) (Staunton *et al.*, 1996; Staunton and Weissman, 2001). (b) Schematic of the structure of mammalian FAS. This iterative enzyme has a core formed of KS, DH and ER homodimers. The bottom of the mFAS is required for chain elongation and substrate selection and the top for reduction of the acyl-CoA. The ACP and TE domains do not form the body of the enzyme (Maier *et al.*, 2008).

As mentioned previously, there had been no structures of an entire PKS module, however, recently electron cryo- microscopy (cryo-EM) was used to elucidate the structure of a dimer of module 5 of the pikromycin PKS (PiKAIII). The dimeric module was 328 kDa and had a fully functional *holo*-ACP associated. Figure 1.11 (a) shows the solid rendered map of the PKS domain solved to between 7.3 and 9.5 Å. The KS dimer is at the top, with the AT domain below supported through a large interaction surface area with the KS domain of the other chain. This is followed by the KR domains forming the base and interface with the next module. The relative orientations of the KS and AT domains observed in the EM structure are different to those in the di-domain crystal structures, having a larger interaction surface than originally shown, but supporting the "Cambridge model". Finally the arch shape observed provides a "reaction chamber" for the ACP. (Dutta *et al.*, 2014; Tang *et al.*, 2007; Tang *et al.*, 2006; Staunton *et al.*, 1996). Recently, high resolution data has been combined with small-angle X-ray scattering (SAXS) to investigate the structure

of modules 5 and 6 from DEBS, providing further structural evidence (Edwards *et al.*, 2014).

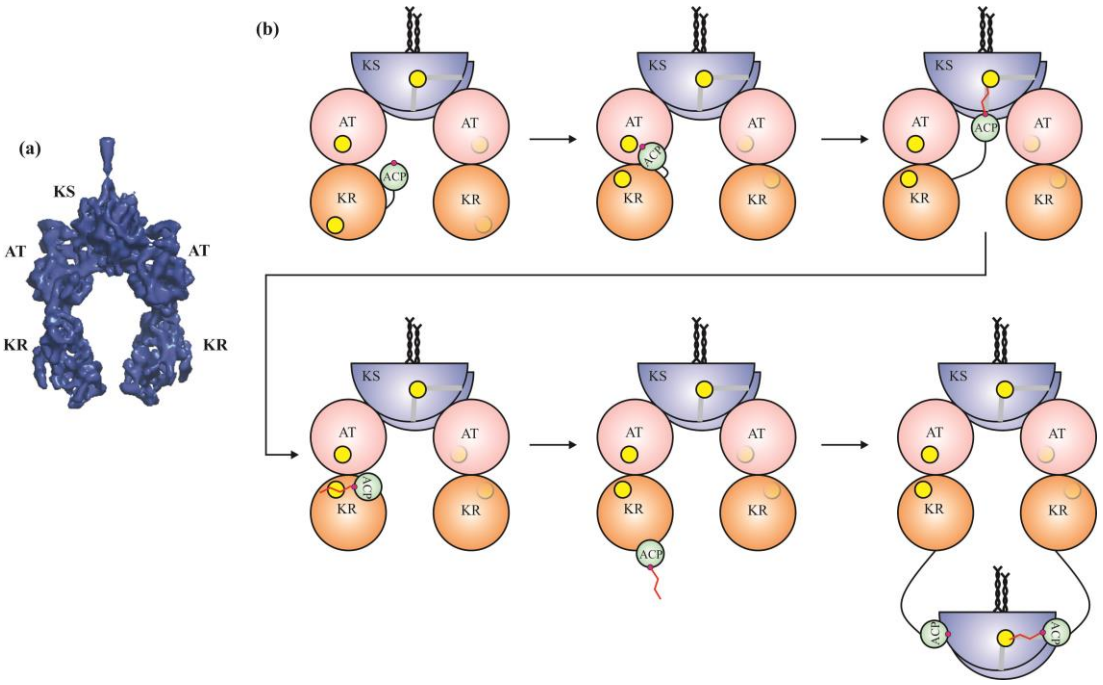


Figure 1.11- Structure and chain elongation mechanism of a module of PikAIII (a) Cryo-EM structure of module 5 of PiKAIII composed of KS-AT-KR-ACP domains with a resolution of 7.3-9.5 Å. The KS forms a dimer at the top, stabilised by AT domains from the other monomer chain. The base of the structure is formed by the KR domains. (b) Dynamics of the module during chain elongation as observed in EM structures. The ACP domain resides within a cavity, then associates with the AT domain to be loaded with the extender unit. The ACP then moves to the KS domain and chain elongation is catalysed, the ACP then ferries the extended chain to the KR domain to undergo reduction. Finally, the ACP then passes the polyketide chain onto the KS domain of the following module (Dutta *et al.*, 2014; Whicher *et al.*, 2014).

The elucidation of the structure of module 5 of the pikromycin PKS by EM also provided information on how chain elongation occurs (Figure 1.11 (b)). The ACP appears to travel to each module for them to carry-out their jobs, specific protein-protein interactions with the catalytic domains are dependent on the substrate attached to the phosphopantetheine arm. There is evidence of domain cross-talk, with the KR domain flipping when the ACP interacts with the AT domain, to be in place for reduction of the growing chain. Conformational changes of the PKS module show the AT domain also restricts side substrate entry in the KS domain, the

route of the extending chain. The mechanism indicated by this data shows a cascade of conformational changes, dependent on the previous module, in an assembly line-like fashion (Whicher *et al.*, 2014).

A combination of structure determination of individual domains, how they dock onto one another, and now a structure determination of an entire module and dynamics in combination with biochemical characterisation may aid in a more knowledge-based reasoning behind engineering PKS, and thus an increased chance in producing functional hybrid enzymes to produce novel compounds.

1.3.4 Engineering of PKSs

Polyketide derived drugs are said to generate worldwide revenues of £10 billion per annum and account for 10% of the top selling drugs (Weissman and Leadlay, 2005). Traditionally, polyketide drugs have been isolated from natural sources such as the soil or by modifying existing drugs. More recently, the idea of creating novel polyketides by engineering polyketide synthases has become an attractive option for the production of novel compounds that have pharmaceutical properties (Weissman and Leadlay, 2005). It was this idea that has motivated decades of research to understanding the structure and function of PKSs.

Polyketide synthases provide a versatile system for engineering due to their modular nature. As described previously, each individual module is composed of domains with discrete active sites responsible for extender unit incorporation and catalysing chain extension. It was suggested that the structure and organisation of these multimodular enzymes would lend themselves to engineering by "cutting and pasting" domains and modules, leading to what is now termed "combinatorial biosynthesis" (Khosla and Zawada, 1996; Weissman and Leadlay, 2005). Combinatorial biosynthesis describes the act of creating novel products by mixing features of known structures, and utilising and manipulating the enzymes that naturally produce these features (Staunton and Weissman, 2001).

There are several approaches for engineering PKS to produce novel compounds, however engineering of PKS in the natural host is not amenable for a number of reasons, and therefore heterologous expression of PKS proteins is often the first

obstacle in engineering PKSs. This is discussed in some depth in chapter 3. Strategies for engineering PKSs and their successes are covered below.

There are several strategies that can be considered for PKS engineering, focussing on either individual domains, whole modules, using synthetic molecules for incorporation and finally alteration of enzymes involved in modification of the final polyketide chain. The most practiced strategy of PKS engineering is engineering of domains within a PKS biosynthetic gene cluster (Staunton and Weissman, 2001).

Structural diversity observed in polyketides can, for some part, be attributed to the types of starter and extender units incorporated (Moore and Hertweck, 2002; Chan *et al.*, 2009). One area of PKS engineering has considered the incorporation of alternative starter units. In 1998 Marsden *et al*. exchanged the promiscuous starter module of the avermectin PKS, able to incorporate C2 branched carboxylic acids, with the DEBS starter module. The resulting chimeric PKS was able to produce novel products. This success highlighted the self-sufficient nature of a PKS module and the ability for downstream modules to accept the new polyketide chain (Marsden *et al.*, 1998; Dutton *et al.*, 1991). Research into the incorporation of different starter units in the DEBS PKS has also included exchanging the DEBS loading di-domain with the oleandomycin and tylosin starter modules, creating truncated products with acetate or propionate as exclusive starter units (Long *et al.*, 2002).

Additionally, a strategy called chemobiosynthesis has been used to incorporate novel unnatural starter units. To do this the loading KS domain of DEBS was inactivated, and the PKS was able to incorporate diketide-*S*-*N*-acteyl cysteamine (SNAC) or diketide-*S*-*N*-proionyl cysteamine (SNPC) to initiate novel product biosynthesis (Jacobsen *et al.*, 1998; Frykman *et al.*, 2001; Dutton *et al.*, 1994). The use of SNACs and SNPCs is quite costly, however it has been shown that the DEBS can also use more cost efficient thioesters, such as methyl thioglycolate, as a means of chemobiosynthesis to create novel products (Murli *et al.*, 2005).

In the same vein, engineering of AT domains within starter and extender modules has altered substrate specificity. There are two suggested ways to do this; the more commonplace method is AT domain exchange, selecting an AT domain with the desired substrate specificity or the less common way, by mutagenesis of amino acids

determined to be involved in substrate specificity (Dunn and Khosla, 2013). The AT domains of both full length and truncated DEBS have been switched with AT domains from the rapamycin PKS, incorporating malonyl CoA as extender units as opposed to methylmalonyl CoA extender units, with varying levels of product yields observed (Lau *et al.*, 1999; Oliynyk *et al.*, 1996; Ruan *et al.*, 1997). AT domain substitution has not been limited to rapamycin, with other PKS AT domains from the pikromycin PKS and the FK520 PKS being exchanged (Ruan *et al.*, 1997; Kato *et al.*, 2002).

The biggest pitfall of the production of novel polyketides by altering substrate selection by AT domain exchange appears to be the decrease in product titres, suggested to be due to the disruption of protein-protein interactions or the inability to process the extender chain downstream (Liou and Khosla, 2003; Hans *et al.*, 2003). In addition to this there doesn't appear to be any specific guidelines for successful AT domain substitutions, with some domains more difficult to replace, and in some cases being dependent on the location of the domain (Reeves *et al.*, 2001; Lau *et al.*, 1999; Ruan *et al.*, 1997; Dunn and Khosla, 2013). To circumvent the issues related to decreased product titres, sections of AT domains responsible for substrate selectivity have been switched, creating hybrid AT domains (Lau *et al.*, 1999). Mutagenesis of the YASH motif (key to methylmalonyl CoA specificity) to the HAFH motif (selective for malonyl CoA) has produced promiscuous DEBS AT domains with the ability to also incorporate unnatural extender units (Reeves *et al.*, 2001; Sundermann *et al.*, 2013; Del Vecchio *et al.*, 2003).

There are two types of AT domains, *cis*- and *trans*- acting. *Cis*-acting AT domains reside within a particular module, *trans*-acting AT domains are detached single enzymes able to load multiple ACP with extender units. It has been suggested that an inactivated *cis*-AT domain could be substituted by a *trans*-acting AT domain (Piel, 2010; Dunn *et al.*, 2014).

Combinatorial biosynthesis has also explored the possibilities of altering the β-keto carbon processing, by the introduction or removal/inactivation of the accessory domains. The bulk of this research has been carried out with the DEBS and rapamycin PKSs (Kao *et al.*, 1997; Kao *et al.*, 1998; McDaniel *et al.*, 1997). Exchange of the KR domains has also shown effects on the stereochemistry of the

product. The KR domains located in the second and fifth modules of DEBS were exchanged with the KR domain from the rapamycin PKS altering the stereochemistry observed in the product (Kao *et al.*, 1998).

The scope of combinatorial biosynthesis is not limited to single domain exchanges, McDaniel *et al*. reportedly substituted multiple AT domains and added the accessory β-processing domains in order to generate a novel macrolactone library, they otherwise claim would not be possible to create by chemical synthesis (McDaniel *et al.*, 1999). Entire modules have been able to be linked together to create novel PKS products, this has been shown by fusing a module from the rifamycin PKS at the end of a truncated DEBS PKS and replacing module 2 in a full length DEBS PKS (Gokhale *et al.*, 1999).

Functional hybrid PKSs have been expressed using multiplasmid approaches. One approach involved cloning native and variant (domain exchanges) DEBS genes into compatible vectors. These vectors were introduced into a *Streptomyces* strain consecutively, 70% of the strains containing a full PKS (3 plasmids) yielded products (Xue *et al.*, 1999). This approach was repeated using pikromycin, erythromycin and oleandomycin PKSs, successfully producing hybrid PKSs (Tang *et al.*, 2000). More recently, design of PKS genes for heterologous expression of PKS modules in *E. coli* has shown huge successes, with synthetic genes of 14 modules from 8 different PKS able to participate in bi-modular interactions in at least one combination (Menzella *et al.*, 2005).

Engineering PKSs is not without its difficulties. As mentioned previously, PKS domains and modules are joined together by linkers and it is well-recognised that interference with these linkers can disrupt PK formation or skip the newly introduced module (Gokhale *et al.*, 1999; Thomas *et al.*, 2002). A greater understanding of linkers between domains and modules could enable design of universal linkers allowing any number of PKS domain and module combinations (Weissman, 2004). In addition to this, some PKSs, such as DEBS, are composed of multiple polypeptides, which need to associate for chain elongation. Highly structured docking domains are responsible for this (Broadhurst *et al.*, 2003). Therefore, in order to create a hybrid PKS a greater understanding of the docking domains is

required. This may enable the design of an "orthogonal" pair of docking domains which could be utilised in engineering hybrid PKS (Weissman, 2004).

The modular nature of polyketide synthases and the applications for creating novel pharmaceutical products are mirrored in nonribosomal peptide synthetases described below.

1.4 Nonribosomal peptide synthetases

Nonribosomal peptide synthetases (NRPS), the producers of non-ribosomal peptides, are large multimodular enzymes on the scale of MDa. NRPS construct NRPs in an analogous manner to the way PKS produce PKs. Like PKSs, the modules within an NRPS are each responsible for the incorporation of a single extender unit, constructing a nonribosomal peptide in an assembly line-like fashion utilising the same enzyme-template mechanism as PKS and FAS. In an NRPS as module also describes a group of domains responsible for incorporation and modification a single extender unit to the growing peptidyl chain (Meier and Burkart, 2009). Also, as with PKSs, the number, type and order of modules within the NRPS dictates the product formed (Schwarzer *et al.,* 2003). NRPS are also known to be able to utilise a vast number of substrates (Marahiel *et al.*, 1997).

1.4.1 NRPS biosynthetic gene clusters

As with PKS, probing for NRPS biosynthetic gene clusters is carried out in a similar manner to PKSs, by exploiting the homologous nature of these multienzymes (Marahiel *et al.*, 1997). Figure 1.16 shows the biosynthetic gene cluster of tyrocidine biosynthesis, the tyrocidine biosynthetic gene cluster was identified by cloning sections of 10-22 kb of *B. brevis* genomic DNA into a bacteriophage vector and screening for homologous DNA using polyclonal antibodies raised against Gramicidin S synthetase 2 (GrsB). This led to the discovery of TycA and TycB (Mittenhuber *et al.*, 1989).

Identification of domains within the surfactin synthetase were identified using oligonucleotide probes based on highly conserved sequences within the gramicidin S, tyrocidine and ACV synthetases and carried out through hybridisation and *in situ* sequencing (Borchert *et al.,* 1992). NRPS operons appear to be between 18 and 45 kb in length, with the operons containing anywhere between one and six modules,

such as TycA and TycC, respectively (Marahiel *et al.*, 1997; Mootz and Marahiel, 1997). Other genes associated with NRPS are located at either the 5' or 3' ends of the operons, these proteins show high sequence homology to the type II fatty acid thioesterase (Mootz and Marahiel, 1997; Marahiel *et al.*, 1997).

Comparison of NRPS DNA sequences and heterologous expression of DNA fragments revealed how NRPs were formed, with particular proteins involved in substrate recognition and activation. It also showed each module was involved in incorporation of a single unit and this dictated the length and size of the peptide synthetase, and *vice versa* (Marahiel *et al.*, 1997). The greater the number of biosynthetic gene clusters that have been fully characterised could conceivably increase the molecular toolbox that could be used for engineering NRPS modules in order to produce novel pharmaceutical compounds.

1.4.2 Biosynthesis of nonribosomal peptides and module organisation

As with PKS's, NRPS modules are composed minimally of three core domains, responsible for the substrate selection of the extender unit, catalysis and housing the growing chain, these are the condensation (C), adenylation (A) and peptidyl carrier protein (PCP) or thiolation (T) domains, respectively (Fischbach and Walsh, 2006). Figure 1.12 illustrates the domain organisation and chain elongation within a minimal module.

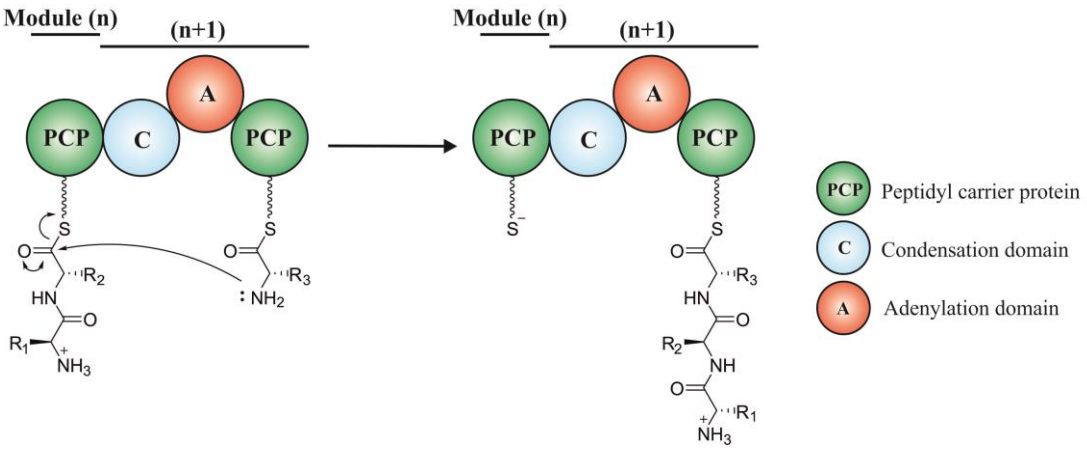


Figure 1.12- Domain organisation within an NRPS minimal module. The condensation domain catalyses the peptide bond formation between the growing peptidyl chain housed on the peptidyl carrier protein and an extender unit, covalently attached to the PCP domain in the next module (Fischbach and Walsh, 2006; Weber *et al.*, 1990).

Synthesis of NRPs are carried out in a number of steps. The A domain is responsible for selection and activation of the correct substrate, converting a carboxylic acid to an aminoacyl adenylate (Figure 1.13) (Dieckmann *et al.*, 1995). The activated amino acid is then transferred onto the terminal thiol of the phosphopantetheine co-factor of the PCP domain within the same module (Ehmann *et al.*, 2000a; Stachelhaus *et al.*, 1996). Formation of the peptide bond then occurs between the electrophilic thioester of the aminoacyl-S-PCP (extender unit) and the nucleophilic amine group of the upstream peptidyl-S-PCP (growing chain), the C domain catalyses this reaction. This continues, with each module adding a single extender unit before reaching the final module, which, more often than not contains a thioesterase domain for chain termination (Fischbach and Walsh, 2006).

Another factor contributing to the diversity of non-ribosomal peptides is the presence of accessory domains within the module; these can include an epimerase domain, an *N*-methyl transferase or a cyclase domain. The epimerase domain (section 1.4.2.5) catalyses the racemisation of the most recently added substrate to the growing peptidyl chain. The cyclase domain (section 1.4.2.6) catalyses the formation of small heterocyclic rings and *N*-methylation (section 1.4.2.7), often observed in NRPs produced by fungi, is carried out by a methyltransferase domain which uses an *S*-adenosylmethionine co-factor to add a methyl group to the N-terminus of the amino acids (Fischbach and Walsh, 2006).

1.4.2.1 Adenylation domain

The adenylation domain initiates the incorporation of a substrate, typically an amino acid, by selection and activation. The A domain, sometimes called "The Gatekeeper" enzyme, is specific for its cognate amino acid. A domains are usually approximately 550 amino acids and activates the amino acid through Mg^{2+}-dependent hydrolysis of ATP, producing an aminoacyl adenylate and pyrophosphate as a by-product. The aminoacyl adenylate can then undergo nucleophilic attack by the free thiol of the phosphopantetheine group attached to the PCP (Figure 1.13) (Sieber and Marahiel, 2005; Hur *et al.*, 2012).

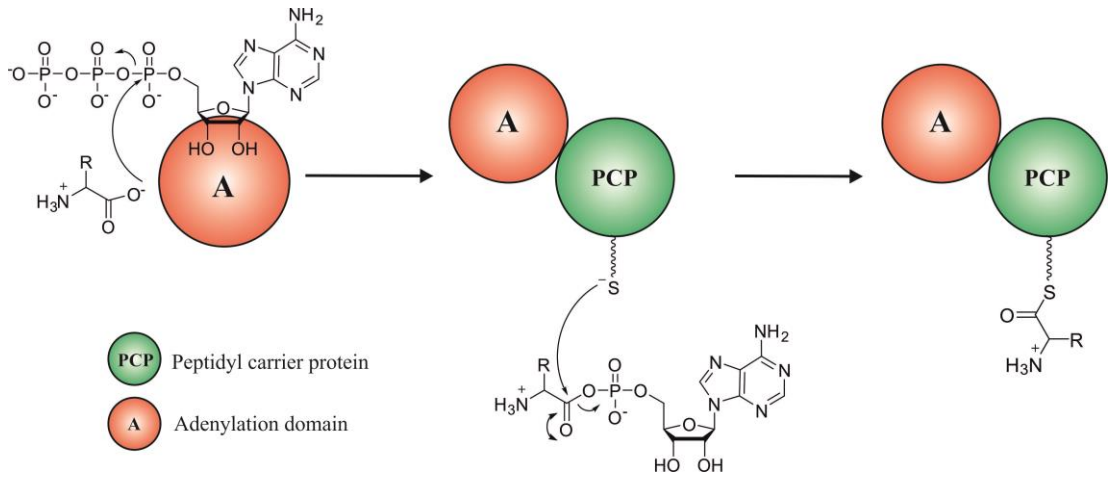


Figure 1.13- Amino acid activation and loading of substrates by adenylation domains in NRPS. The selected amino acid is activated by formation of an aminoacyl adenylate in a Mg^{2+} dependent manner. This then becomes the substrate for nucleophilic attack of the free thiol of the phosphopantetheine arm of the PCP (Sieber and Marahiel, 2005).

Amino acid activation in NRPSs by the adenylation domain is mechanistically similar to amino acid activation by aminoacyl-tRNA synthetases however there is little or no sequence and structural similarities between the two. Additionally, although the adenylation domain has substrate specificity, it is more promiscuous than its mechanistically similar counterpart aminoacyl-tRNA synthetases (Sieber and Marahiel, 2005).

A domains show approximately 30-60% sequence identity, and through sequence alignments an AMP binding domain has been identified. This site is conserved throughout the adenylate forming enzyme superfamily, including in acetyl CoA synthetases (Turgay *et al.*, 1992; Starai and Escalante-Semerena, 2004). Sequence alignments of the PheA adenylation domain from the gramicidin S synthetase with 160 other A domains also highlighted 10 residues responsible for substrate specificity, in this case for phenylalanine (Stachelhaus *et al.*, 1999).

Several A domains have been heterologously expressed in *E. coli*, purified and subsequently their activity and substrate specificity have been assayed *in vitro* by ATP-PPi exchange with radiolabelled substrates, for example the EntE and EntF adenylation domains from the enterobactin NRPS (Linne and Marahiel, 2004; Ehmann *et al.*, 2000a). Several crystal structures of A domains have also been solved (section 1.4.3).

1.4.2.2 Condensation domain

Condensation (C) domains catalyse the formation of a peptide bond between the newly recruited extender unit and growing peptidyl chain housed on the PCP (Stachelhaus *et al.*, 1998). Condensation domains also possess the ability to catalyse the formation of a bond between an acyl group from an ACP-bound PK from a PKS module and an aminoacyl group bound to a PCP. This occurs in PKS/NRPS hybrid enzymes (section 1.5), such as in the rapamycin PKS/NRPS (König *et al.*, 1997; Hur *et al.*, 2012). Some C domains also have the ability to catalyse the formation of ester bonds, in fact a C domain from an NRPS in *S. globisporus* is known to be able to do both (Lin *et al.*, 2009; Hur *et al.*, 2012). In an analogous manner to some PKSs, the domain responsible for catalysis, the condensation domain, is sometimes missing in the initiation module in some NRPS, such as in TycA (Figure 1.16) (Fischbach and Walsh, 2006; Mootz and Marahiel, 1997). Catalysis in C domains is mediated by electrostatic interactions, aided by the structure of the C domain (Keating *et al.*, 2002; Hur *et al.*, 2012).

It has been shown that in addition to the substrate specificity of the adenylation domain, the condensation domain also displays some level of substrate specificity for the aminoacyl adenylate to be added to the growing NRP chain, being selective for the D-enantiomer in addition to selectivity observed for particular side chains. This selectivity only seems to pertain to the aminoacyl adenylate and not the growing chain (Belshaw *et al.*, 1999). This specificity of the C domain for the aminoacyl adenylate, once the peptide bond has been formed and the aminoacyl adenylate added to the growing chain, is lost and the peptidyl-S-PCP is no longer a substrate, this causes release of the growing chain. This mechanism has been suggested to prevent accidental initiation in an elongation module. This also explains the uni-directional nature of NRPS (Linne and Marahiel, 2000).

1.4.2.3 Peptidyl carrier protein

The peptidyl carrier protein (PCP) is homologous to the acyl carrier protein (ACP) in PKSs (Meier and Burkart, 2009; Crosby and Crump, 2012). As with the ACP, the

PCP is a small non-catalytic protein essential for NRP chain elongation. The PCP is also phosphopantetheinylated on a conserved serine by an auxiliary phosphopantetheinyl transferase (PPTase) (Walsh *et al.*, 1997; Lambalot *et al.*, 1996). Elongation occurs in a synonymous fashion to PKS as illustrated in Figure 1.12. The structure of the PCP from module 3 of TycC has been elucidated by NMR (Weber *et al.*, 2000), structural aspects of PCP domains will be discussed in depth in chapter 5.

1.4.2.4 Chain termination and macrocyclisation

The thioesterase (TE) domain resides at the C-terminus of the final module within the NRPS and is responsible for NRP chain termination (Schneider and Marahiel, 1998) using an active site serine in a Ser-Asp-His catalytic triad (Kohli and Walsh, 2003). Chain termination can have one of two outcomes; a linear NRP is observed when hydrolysis occurs, as occurs in vancomycin biosynthesis (Hubbard and Walsh, 2003; Keating and Walsh, 1999). Macrocyclisation occurs when a nucleophile within the NRP chain attacks itself, this occurs in daptomycin biosynthesis (Hur *et al.*, 2012). The main difference between PKS and NRPS TE domains is the PKS TE domain is a homodimer, forming a hydrophobic channel, and the NRPS domain is a monomer. The NRPS TE domains of surfactin and fengycin have a hydrophobic cavity, with a lid region, which can accommodate the NRP chain and allow cyclisation (Grunewald and Marahiel, 2006; Bruner *et al.*, 2002; Samel *et al.*, 2006).

1.4.2.5 Epimerase domain

The structures of NRPs show the presence of D-amino acids to be ubiquitous, however in Nature, in the producing organisms, there is a predominant deficiency of D-amino acids. In most NRPSs a domain is present responsible for the epimerisation of the L-amino acids that are incorporated, residing at the C-terminus of the PCP within the same module as the newly added aminoacyl substrate. Epimerisation of amino acids occurs at the α-carbon of the newly added aminoacyl adenylate; in fact the amino acid must be tethered to the PCP. Figure 1.14 illustrates epimerisation of

an L-amino acid to a D-amino acid by an epimerase domain (Pfeifer *et al.*, 1995; Stachelhaus and Walsh, 2000; Hur *et al.*, 2012).

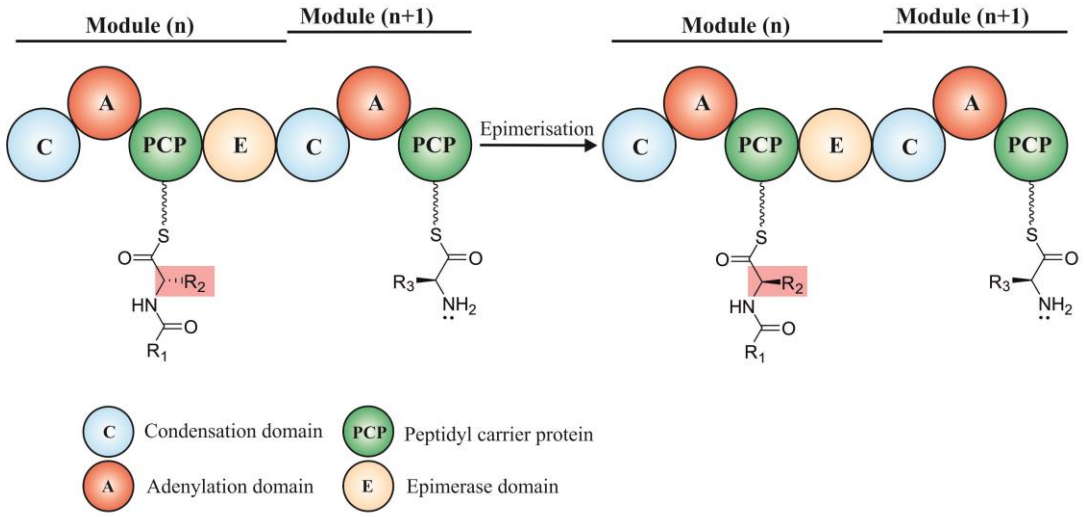


Figure 1.14- Epimerisation of the peptidyl chain mediated by an epimerase (E) domain. The E domain converts the most recently added amino or aryl acid from the L-isomer to the D-isomer before the growing peptide chain is passed to the next module. Figure adapted from Fischbach and Walsh (2006).

Studies were performed on modules 4 and 5 of the tyrocidine biosynthetic gene cluster (shown in Figure 1.16) to investigate substrate specificity of the C domain downstream for the L or D enantiomers. In module 4 of tyrocidine biosynthesis L-phenylalanine is incorporated into the peptidyl chain and is subsequently converted to D-phenylalanine by the E domain located at the C terminus of the PCP, between modules 4 and 5. The subsequent C domain (in module 5) is then responsible for the condensation of the new peptidyl growing chain and L-asparagine. This study showed that the C domain of module 5 is specific for the D-peptidyl growing chain, in addition to being specific for the L-aminoacyl adenylate (L-asparagine) to be added (Clugston *et al.*, 2003). Although the E domain itself has been known to accept alternative substrates (Luo *et al.*, 2001).

1.4.2.6 Cyclisation domain

Another domain that increases the diversity seen in NRPs is the cyclase (Cy) domain. Cy domains catalyse the formation of heterocyclic rings at cysteine, serine

and threonine residues creating thiazoline or oxazoline rings (Hur *et al.*, 2012). The Cy domains also catalyses the peptide bond formation between the extender unit and growing peptidyl chain (Fischbach and Walsh, 2006; Hur *et al.*, 2012). Figure 1.15 illustrates the formation of the heterocyclic ring.

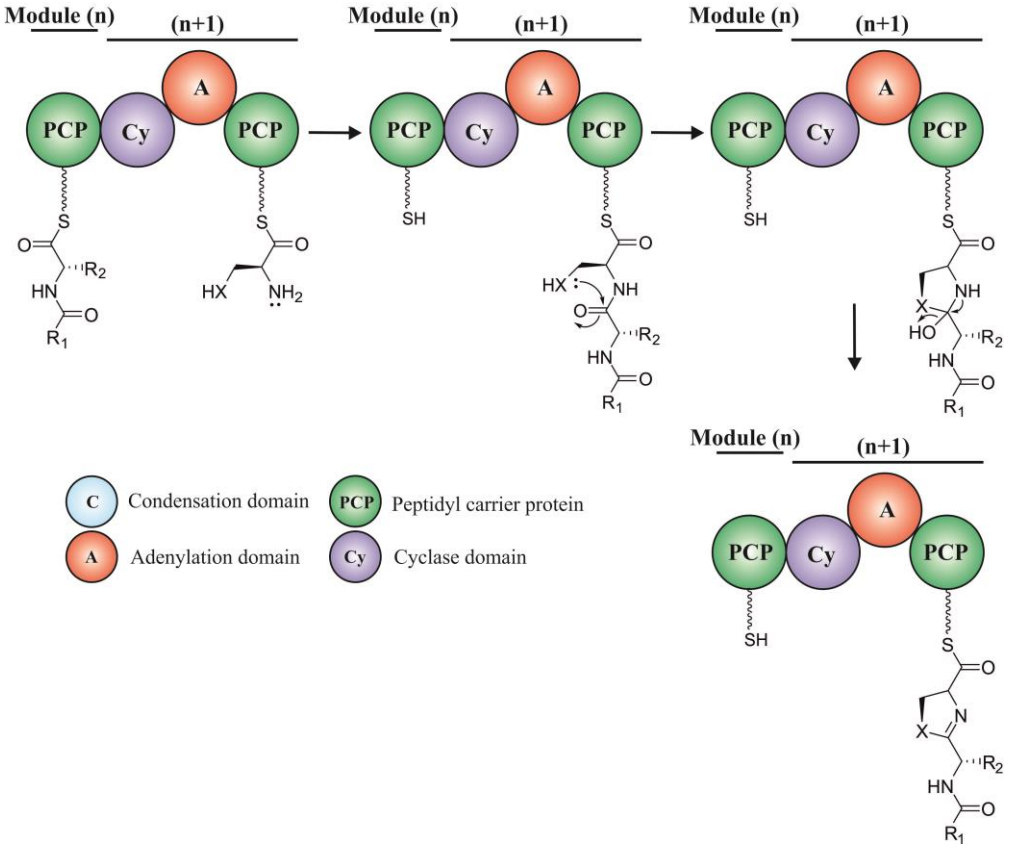


Figure 1.15- Mechanism of heterocyclisation within NRPs. The cyclase (Cy) domain catalyses the formation of the peptide bond between the extender unit and growing peptide chain. The resulting peptide bond then undergoes intramolecular nucleophilic attack by the thiol/hydroxyl side chain, this is followed by dehydration, forming a thiazoline or oxazoline ring (Hur *et al.*, 2012).

During heterocyclisation the thiol/hydroxyl groups of cysteine, serine and threonine residues added to the peptidyl chain can attack the peptidyl backbone creating a thiazoline or oxazoline ring. The newly formed heterocycle can then undergo further oxidation or reduction (Schneider *et al.*, 2003; Reimmann *et al.*, 2001).

1.4.2.7 Additional NRP chain modifications

N-methylation often aids in the biological activity of NRPs and is mediated by a methyltransferase (MT). The MT catalyses the transfer of a methyl group onto the NRP from a (*S*)-adenosyl methionine (Hur *et al.*, 2012). Methylation occurs when the amino acid is loaded onto the carrier protein. There are two types of MT domains, c*is*-MT domains which reside at the C-terminus of the A domains, and *trans*- acting MT domains which can methylate the amino acid extender unit or the growing peptidyl chain (Weber *et al.*, 1994; O'Brien *et al.*, 2000; Fischbach and Walsh, 2006).

Other modifications of NRPs can include formylation and halogenation. Formylation is mediated by a formylation (F) domain, utilising a N^{10}-formyltetrahydrofolate cofactor. Formylation is known to occur in the biosynthesis of linear gramicidins. This is the least studied accessory domain (Kessler *et al.*, 2004; Schoenafinger *et al.*, 2006; Hur *et al.*, 2012). Finally, halogenation can occur on the heterocyclic rings introduced by Cy domains, this is mediated by a flavin dependent halogenase (Dorrestein *et al.*, 2005; Hur *et al.*, 2012).

1.4.2.8 Linking modules in NRPS

For efficient transfer of the peptide chain between modules, small (15-25 amino acid) domains exist between two different NRPS modules. The domains are called communication-mediating (COM) domains. COM domains reside at the C terminus of upstream module and N terminus of the downstream module; these are a cognate pair to ensure the modules are in the correct order for production of the target NRP (Hahn and Stachelhaus, 2004).

1.4.2.9 A model system

Tyrocidine, a cyclic NRP, is produced by *B. brevis*. Figure 1.16 illustrates the biosynthetic gene cluster of tyrocidine.

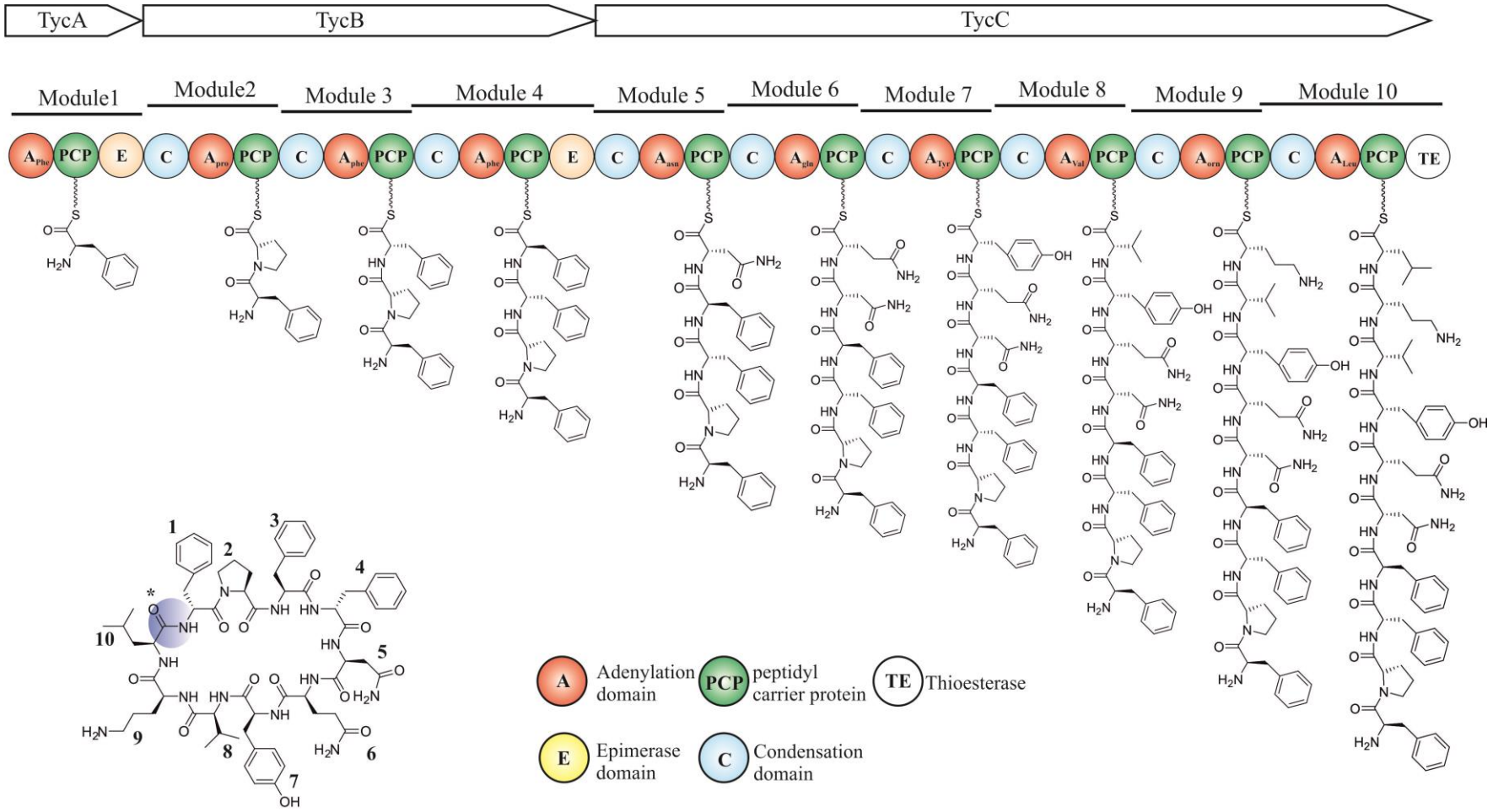


Figure 1.16- Biosynthetic gene cluster of the tyrocidine NRPS. Three polypeptides (TycA, B and C) contain a total of 10 modules. TycA contains module 1, the initiation module which incorporates L-Phe which is converted to D-Phe by an epimerase domain. Module 2 then incorporates L-Pro, followed by L-Phe in module 3. Module 4 then adds L-Phe which is converted to D-Phe. Modules 5 incorporates an L-Asn, followed by L-Gln by module 6. Modules 7 and 8 then incorporate L-Tyr and L-Val. Module 9 incoporates the unnatural amino acid ornithine (L-Orn). Finally, module 10 incorporates a L-Leu before product termination and cyclisation. The blue circle (*) on the tyrocidine structure indicates the join when the decapeptide is cyclised (Mittenhuber *et al.*, 1989; Mootz and Marahiel, 1997).

The tyrocidine NRPS is encoded for by three open reading frames, producing three polypeptides; TycA, TycB and TycC. TycA is composed of the loading domain (module 1), TycB is composed of modules 2, 3 and 4. Finally, TycC, the largest of the polypeptides is composed of modules 5-10, modules 1 and 4 exhibit epimerase activity and module 9 incorporates the unnatural amino acid ornithine (Orn). Tyrocidine is composed of (2) D-Phe, (1) L-Phe, (1) L-Pro, (1) L-Asn, (1) L-Orn, (1) L-Gln, (1) L-Tyr, (1) L-Val and (1) L-Leu (Mootz and Marahiel, 1997). Chemical synthesis of analogues of tyrocidine, performing single amino acid substitutions has yielded analogues with improved antibacterial activities (Marques *et al.*, 2007). It was determined that the starter unit D-Phe and L-Orn are required for cyclisation of tyrocidine (Kopp and Marahiel, 2007).

1.4.3 Structural and mechanistic studies of NRPSs

As with PKS, structural studies of entire NRPS modules are severly hindered by size and flexibility. However, structural understanding may enable combinatorial biosynthesis of these multimodular enzymes. Structure elucidation has been successful for individual domains and di-domains which has given insight into important structural characteristics, especially when considering engineering strategies.

X-ray crystallographic data shows that the adenylation domain is composed of two domains, a larger N-terminal subdomain and a smaller C-terminal subdomain (Marahiel and Essen, 2009). Crystal structures of A-domains responsible for activating 2, 3-dihydroxyxbenzoate (DhbE) from *B. subtilis* (May *et al.*, 2002) and L-phenylalanine (PheA) from *B.brevis* (Conti *et al.*, 1997) display highly conserved folds (Marahiel and Essen, 2009). In fact, although adenylation domains have really low sequence homology they appear to have a conserved, canonical fold (Sieber and Marahiel, 2005). Figure 1.17 shows the crystal structure of PheA, with specific residues for substrate recognition highlighted (Conti *et al.*, 1997).

Figure 1.16 shows the crystal structure of the PheA domain. The conserved A domain structure is composed of two subdomains, a large N-terminal domain contributing 9 out of the 10 residues that compose the hydrophobic binding pocket of the substrate, phenylalanine. The C-terminal domain, a smaller domain of

approximately 100 residues, contains K527, this residue is important for not only the coordination of Phe but also AMP (Conti *et al.*, 1997; Stachelhaus *et al.*, 1999).

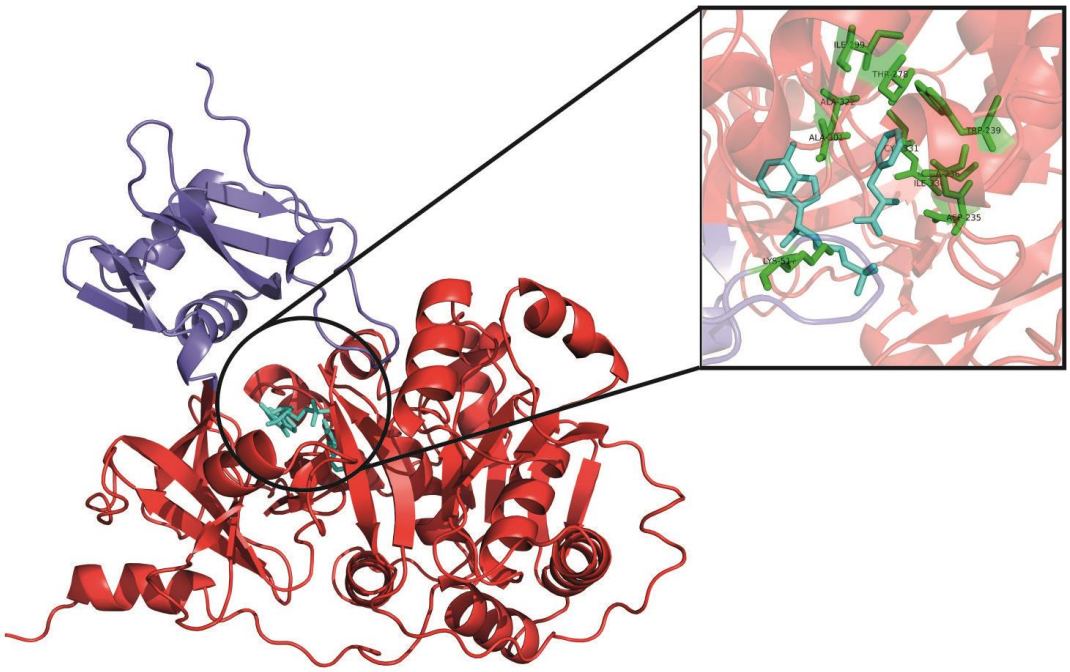


Figure 1.17- Crystal structure of the PheA adenylation domain. N (red) and C (blue) subdomain are highlighted, as are the substrates AMP and Phe (sky blue). The specificity conferring code of PheA which binds phenylalanine is magnified in the inset. This includes residues: D235, A236, W239, T278, I299, A301, A322, I330, C331 and K517 (Conti *et al.*, 1997; Stachelhaus *et al.*, 1999).

Structural studies of DltA, the D-alanine protein ligase, a member of the adenylate activating family, provided some mechanistic insight into the adenylation domain. The small C-terminal sub-domain can be in an open state, allowing entry of the substrates, followed by a closed state, enabling adenylation of the amino acid and transfer onto the PCP (Yonus *et al.*, 2008; Hur *et al.*, 2012).

Structural studies of the condensation domain showed that it is also composed of N and C-terminal subdomains, forming a 'V' shape. Figure 1.17 shows the condensation domain VibH from *Vibrio cholerae* (PDB code 1L5A) (Keating *et al.*, 2002). The 'V' structure of VibH revealed a solvent channel, in which the peptidyl growing chain and aminoacyl adenylate approach from each side and peptide bond formation catalysed (Keating *et al.*, 2002).

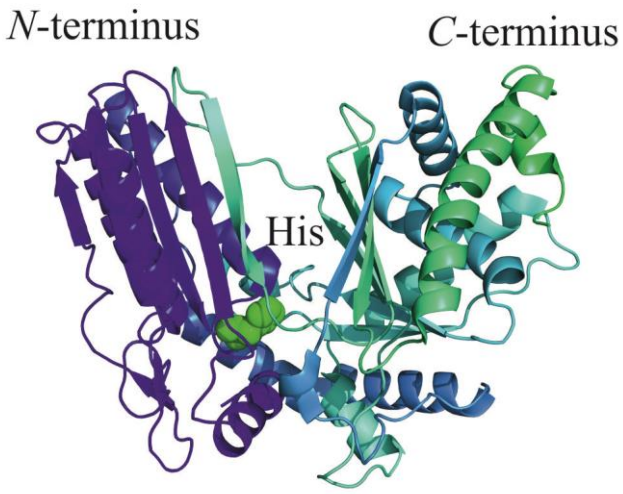


Figure 1.18- Crystal structure of the condensation domain, VibH. The condensation domain is composed of two subdomains, N and C terminal subdomains. These two subdomains form a 'V' shape, providing a trench where the peptide bond is catalysed between the growing peptidyl chain and amino acyl adenylate. His 126 is highlighted as the catalytic residue (Keating *et al.*, 2002).

The structure of a PCP domain has been elucidated by NMR, highlighting multiple conformations in the *apo*- and *holo*- states (Weber *et al.*, 2000; Koglin *et al.*, 2006), this will be further covered in chapter 5.

Crystal structures of the TycC5-6 PCP bi-domain (PDB code 2JGP) (Samel *et al.*, 2007) from the tyrocidine NRPS and termination module from the surfactin NRPS (PDB code 2VSQ) have been solved (Tanovic *et al.*, 2008). The bi-domain structure however did not show a functional interaction as the phosphopantetheine arm would not be able to reach the active site (Samel *et al.*, 2007). The conformation of the PCP domain in the termination module however was in an appropriate orientation for an interaction with the C domain (Tanovic *et al.*, 2008).

The overall module structure of mFAS, and more recently the EM structure of a PikAIII module showed these multienzymes forming a homodimer to function (Dutta *et al.*, 2014; Asturias *et al.*, 2005; Maier *et al.*, 2008). In contrast in NRPSs, the TE domain and VibH C domain, suggested that NRPS are actually monomeric (Bruner *et al.*, 2002; Tsai *et al.*, 2001; Keating *et al.*, 2002). Gel filtration, chemical

cross-linking and AUC showed that the bimodule TycB was monomeric (Sieber *et al.*, 2002; Glinski *et al.*, 2002).

1.4.4 Engineering NRPSs

Engineering of NRPSs has been less extensive than its PKS counterparts. There are two approaches for the engineering of NRPS to yield novel products; editing the peptidyl chain or adding/removing groups on the peptidyl chain (Giessen and Marahiel, 2012). As with PKS however, heterologous expression can assist in easier manipulation for engineering novel product biosynthesis (Challis, 2006).

A target for NRPS engineering is the design of a hybrid NRPS, selecting modules depending on the desired product. This has been achieved with a two module system, creating a hybrid NRPS with a module from surfactin, selecting an aspartate residue and a second module from the tyrocidine NRPS which incorporates a phenylalanine. In addition to this, a termination module was present, this produced a functional NRPS (Doekel and Marahiel, 2000).

Originally, exchanges of A-PCP domains were made to alter the substrates incorporated; novel analogues of surfactin were made by exchanging the A-PCP didomain of module 7 with fungal A-PCP didomains, creating a hybrid NRPS. However, this yielded low product titres, potentially due to the substrate specificity of the downstream C domains (Stachelhaus *et al.*, 1995; Belshaw *et al.*, 1999; Ehmann *et al.*, 2000b). The reduced product titres have also been suggested to be due to disruption of the COM domains, generating an orthogonal pair of COM domains may in fact aid in the creation of hybrid NRPS enzyme (Weissman, 2004). Identification of linkers more tolerant to forming new interactions between A and PCP domains has aided in construction of new hybrid PKS (Doekel and Marahiel, 2000).

The most promising aspect of NRPS engineering has been targeting specific residues within the active site of an adenylation domain, known as the specificity conferring code. This was originally carried out with the adenylation domain of L-Phe from the gramicidin NRPS, where mutagenesis of the PheA adenylation code led to the incorporation of L-leucine. This has also been done in the surfactin PKS, altering the

substrate specificity from L-Glu to L-Gln and L-Asp to L-Asn (Eppelmann *et al.*, 2002). Genetic engineering of the adenylation domain in this way does seem to have some limitations, shape, size and hydrophobicity of the active site appears to control substrate specificity and thus the extent of engineering (Villiers and Hollfelder, 2009).

NRPS thioesterase modules could also be used to initiate macrocyclisation. Experiments carried out with the tyrocidine and surfactin TE domains showed that they were able to initiate cyclisation of synthetic peptides equivalent to the native peptides they cyclise (Tseng *et al.*, 2002; Kohli *et al.*, 2001; Trauger *et al.*, 2000). The tyrocidine TE showed broad substrate tolerance, only really requiring the same N and C terminal residues in the peptidyl chain (Trauger *et al.*, 2000)

The modular nature of these enzymes provides many aspects to engineer to produce novel compounds, some of these engineering strategies, such as the use of a thioesterase domain to initiate macrocyclisation can be used in conjunction with chemical synthesis to produce novel compounds (Sieber and Marahiel, 2003).

1.5 NRPS-PKS hybrid enzymes

Hybrid NRP/PK products can be found in Nature. There are two ways in which NRP/PK hybrid products are synthesised, either they are synthesised by separate enzymes and joined or they are synthesised by NRPS/PKS hybrid enzymes (Du *et al.*, 2003).

Both NRPSs and PKSs work in an assembly line like fashion; NRPS and PKS modules are also found together in complexes in Nature, producing hybrid synthases. There are several well-characterised NRPS-PKS hybrids such as those the produce rapamycin and epothilone. The compounds that NRPS/PKS hybrid enzymes produce, just as with NRPS and PKS, can be predicted from the domains present. Hybrid assembly lines do not have to have a set ratio of NRPS to PKS modules (Schwarzer *et al.*, 2003; Fischbach and Walsh, 2006). As usual growing chains are transferred between NRPS modules and PKS modules, switching from carbon-carbon bond formation to amide bond formation (Fischbach and Walsh, 2006; Schwarzer *et al.*, 2003; Walsh *et al.*, 2006). As mentioned previously, C domains in

NRPS are shown to perform both types of catalysis for bond formation (Hur *et al.*, 2012).

Structurally, the dimeric nature of PKS modules, and the heavily implied monomeric structure of NRPS modules suggests that while the PKS portion of the enzyme dimerises, the NRPS portion remains a monomer (Sieber and Marahiel, 2005). However, investigation into the docking domains between NRPS and PKS modules suggests that NRPS potentially self-associate to aid in the interaction with their dimeric PKS partners (Sieber and Marahiel, 2005; Hur *et al.*, 2012). Understanding the interactions between NRPS and PKS modules will theoretically increase the options when engineering NRPS and PKS modules.

1.5.1 The hybrid NRP/PK indanomycin

In Nature, pyrrole groups fulfil a number of roles, forming of hydrogen bonds, coordination of metal ions and taking part in stacking interactions (Walsh *et al.*, 2006). Indanomycin is a pyrroloketoindane produced by an NRPS/PKS hybrid enzyme found in *Streptomyces antibioticus* NRRL 8167 (Li *et al.*, 2009). Indanomycin displays a number of beneficial properties these include insecticidal, anti-protozoal and antibacterial properties (Liu *et al.*, 1979; Zhang *et al.*, 1997; Li *et al.*, 2009) and works by complexing monovalent or divalent metal ions, using the carboxylic acid and ether oxygen to transport them across a cell membrane and disrupt ion balance (Roege and Kelly, 2009). Of the known pyrroloketoindanes, indanomycin is the only one whose biosynthetic gene cluster has been characterised.

1.5.1.1 Biosynthetic gene cluster and structure of indanomycin

Indanomycin is produced by an NRPS/PKS hybrid enzyme; this multimodular enzyme is composed of one NRPS module at the beginning of the biosynthetic gene cluster, followed by 11 PKS modules. Figure 1.19 illustrates the composition of indanomycin, composed of pyrrole-2-carboxylate, formed from the conversion of L-proline, six malonyl-CoA units, two methylmalonyl-CoA units and two ethylmalonyl-CoA units (Li *et al.*, 2009; Roege and Kelly, 2009).

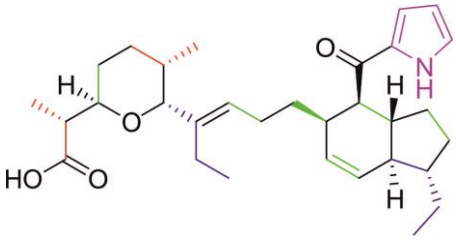


Figure 1.19- Structure of Indanomycin. Indanomycin is composed of six malonyl CoA units (green), two methylmalonyl CoA (red) units and two ethylmalonyl CoA units (blue). Initiation of indanomycin biosynthesis is mediated by a pyrrole (pink) (Roege and Kelly, 2009).

The biosynthetic gene cluster of indanomycin was discovered and characterised by the Kelly Group (Li *et al.*, 2009). As with other NRPs and PKs, identification of the biosynthetic gene cluster involved probing for the presence of a gene suspected to be present. Knowing that indanomycin requires the incorporation of L-proline, primers were designed to probe for a proline adenylyltransferase (Li *et al.*, 2009).

Interestingly, other NRPS such as those for tyrocidine biosynthesis (Figure 1.16) or vancomycin biosynthesis have adenylation domains which incorporate unnatural amino acids such as ornithine or β-hydroxytyrosine (Mootz and Marahiel, 1997; Smith *et al.*, 1975) the NRPS module of indanomycin incorporates L-proline, and modifies when it is attached to the carrier protein as with biosynthesis of aminocoumarins (Garneau *et al.*, 2005).

1.5.1.2 Starter unit biosynthesis

As mentioned previously, indanomycin is formed by an NRPS/PKS hybrid enzyme. The first module, the NRPS module, is responsible for the formation of the starter unit pyrrole-2-carboxylate instigating NRP/PK chain initiation. In starter unit biosynthesis L-Proline is activated and covalently attached to the phosphopantetheine arm of the carrier protein by the adenylyltransferase. The proline is then oxidised to a pyrrole by the flavin dependent prolyl-*S*-dehydrogenase to pyrrole-2-carboxylate. *idm*I encodes the dehydrogenase (IdmI), *idm*J encodes the free standing adenylyltransferase (IdmJ), the activation domain, and *idm*K encodes the carrier protein (IdmK). The starter unit is synthesised then passed off into *idm*L, the first module of the PKS part of the NRPS/PKS (Li *et al.*, 2009).

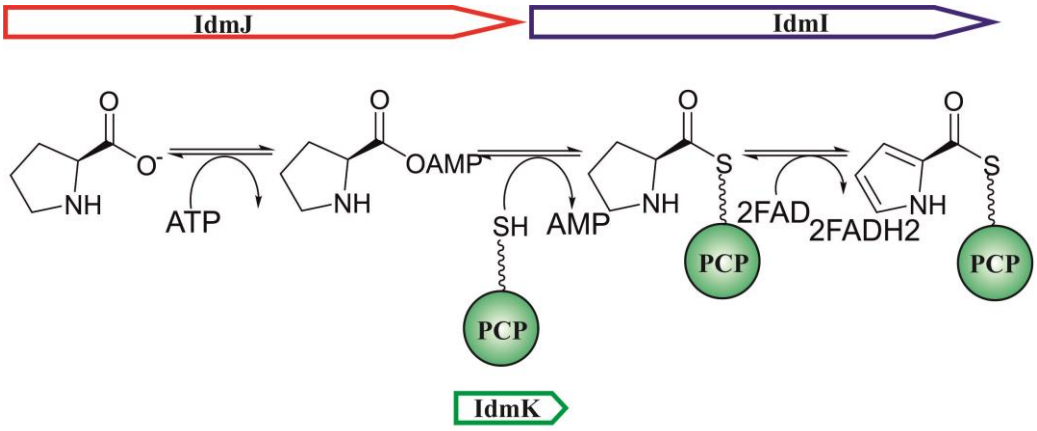


Figure 1.20- NRPS initiation module in indanomycin biosynthesis. IdmJ, the adenylyltransferase selects L-proline as the starter unit and activates and loads it onto IdmK, the carrier protein. The prolyl-*S*-dehydrogenase then oxidises the proline to pyrrole-2-carboxylate (Li *et al.*, 2009).

It is interesting to note that the selection and activation of L-proline and the dehydrogenation to pyrrole-2-carboxylate are also observed in other biosynthetic clusters such as in aminocoumerin biosynthesis (Garneau *et al.*, 2005). Proteins within the clorobiocin and coumermycin biosynthetic gene clusters have been shown to form pyrrole-2-carboxylate in the same manner as indanomycin, however adenylation and dehydrogenase proteins show lower sequence similarity (60-61%) to those found in indanomycin biosynthesis compared to between themselves (91-92%) (Garneau *et al.*, 2005; Walsh *et al.*, 2006). Genes responsible for the formation of the pyrrole group in clorobiocin and coumermycin biosynthetic gene cluster have been cloned and expressed in *E. coli*. ATP[^{32}P]PPi experiments were used to probe the substrate specificity of the adenylation domains, comparing L-proline to proline analogues and L-pipecolinic acid. This showed the adenylation domain already had the ability to accept alternative substrates (Garneau *et al.*, 2005).

The aims of this project were to make some progress towards engineering modular enzymes to produce novel compounds with potential pharmaceutical benefits. The studies with the clorobiocin and coumermycin pyrrole forming enzymes indicated the potential for heterologous expression and expanding the substrate specificity of an L-proline activating adenylation domain.

1.6 Project outlines

This thesis describes the experiments carried out to structurally and biochemically characterise proteins responsible for starter unit formation in indanomycin biosynthesis with the future aim to engineer these proteins as a source of novel polyketide/nonribosomal peptide production. The proposed work had three main goals:

- **Cloning and recombinant expression of proteins required for the conversion of L-proline in starter unit biosynthesis of indanomycin.** This involves cloning of the genes encoding for the adenylyltransferase (IdmJ), dehydrogenase (IdmI) and carrier protein (IdmK) into an appropriate expression vector and expression to be carried out in *E. coli*. Approaches for cloning and soluble protein expression are discussed in chapters 3 and 4.

- **Purification and biochemical characterisation of IdmJ, IdmI and IdmK.** Successful cloning and protein expression will be followed by purification by nickel affinity chromatography. Biochemical characterisation of the system will require phosphopantetheinylated IdmK, therefore investigation into exogenous PPTases with ability to modify IdmK may be required. Discussion of the expression of IdmK and obtaining a *holo*-IdmK species can be found in chapter 4. Characterisation of starter unit incorporation can be carried out *in vitro* with homogenous protein samples.

- **Structural characterisation of IdmJ, IdmI and IdmK.** After establishing a protocol to obtain a homogeneous protein sample, structural characterisation of individual proteins within the whole module can be carried out. A combination of biochemical and structural techniques can be used to investigate protein-protein interactions required for a functional NRPS module. Structural characterisation can provide insights into a structure function relationship, aiding in structure based engineering strategies. Structural characterisation of IdmK is discussed in chapter 5.

2. Materials and methods

2.1 Materials

2.1.1 Chemicals

Analytical grade chemicals and reagents were used throughout. Monobasic sodium phosphate, dibasic sodium phosphate, methanol, glycerol, glycine, sodium chloride, ampicillin, Tris base, glacial acetic acid, ethylenediaminetetraacetic (EDTA), isopropanol and ammonium acetate were purchased from Fisher Scientific Ltd. (Loughborough, UK). Agar, agarose, dithiothreitol (DTT) and isopropyl-β-D-1-thiogalactopyranoside were purchased from Melford Laboratories Ltd. (Suffolk, UK). Ethanol, bromophenol blue, *N,N,N`,N`*-tetramethylethylenediamine (TEMED), nickel (II) chloride, acrylamide:*N,N′*-methylenebisacrylamide 24:1, urea, tryptone and yeast extract were obtained from Sigma-Aldrich (Dorest, UK). Phosphate buffered saline (PBS) was purchased from Oxoid, part of Fisher Scientific Ltd. Hydrochloric acid, 2-mercaptoethanol and imidazole were purchased from Acros Organics, also part of Fisher Scientific. Ammonium persulfate was purchased from Amersham Biosciences Ltd. (Buckinghamshire, UK). Sodium dodecyl sulphate (SDS) was obtained from BDH Chemicals Ltd. (Poole, UK). SYBR® Safe was purchased from Invitrogen (Paisley, UK). 6 × Blue/Orange dye was purchased from Promega (Southampton, UK). Nickel (II) chloride hexahydrate was obtained from Alfa Aesar (Lancashire, UK). Lysozyme was purchased from MP Biomedicals (Cambridge, UK). $^{15}NH_4Cl$ and D-glucose-$^{13}C_6$ were purchased from Cambridge Isotope Laboratories, Inc. (Leicestershire, UK).

2.1.2 Bacterial strains and plasmids

The bacterial strain used in transformations for DNA manipulation was *E. coli* XL10 Gold® Ultracompetent cells (Tet^{r}Δ(*mcrA*)*183* Δ(*mcrCB-hsdSMR-mrr*)*173 endA1 supE44 thi-1 recA1 gyrA96 relA1 lac* Hte [F′ *proAB lacI*^{q}ZΔ*M15* Tn*10*(Tet^r) Amy Cam^r]). Protein expression was achieved in the *E. coli* strain BL21-Gold (DE3) (B F$^–$ *ompT hsdS*($r_B^-m_B^-$) *dcm*$^+$ Tet^r *galλ*(DE3) *endA* Hte). The *E. coli* strains were both purchased from Agilent Technologies, Cheshire, UK.

Both the pKK223-3 and pETDUET plasmids were provided by Prof. Alan Berry (Astbury Centre, University of Leeds, UK). The pMAL™ c5X vector was a kind gift from Dr James Ross (Astbury Centre, University of Leeds, UK). pET23a vectors were either a kind gift from Dr David Brockwell (Astbury Centre, University of Leeds, UK) or purchased from Merck chemical Ltd. (Formerly Novagen) (Nottingham, UK).

2.1.3 Enzymes

Restriction enzymes *Eco*RI-HF®, *Hind*III-HF®, *Kpn*I-HF®, *Pst*I-HF®, *Sac*I-HF®, *Sal*I-HF®, *Xho*I and *Nde*I were purchased from New England Biolabs® (Ipswich, MA, USA). *PfuTurbo*® was obtained from Promega (Southampton, UK). Deoxyribonuclease I from bovine pancreas was purchased from Sigma Aldrich (Dorset, UK).

2.1.4 Antibodies

The primary antibody used was the Monoclonal Anti-polyHistidine antibody produced in mouse at a 1: 3 000 dilution. The secondary antibody used was Anti-mouse IgG conjugated with horseradish peroxidase in a 1: 20 000 dilution, both were purchased from Sigma-Aldrich (Dorset, UK).

2.1.5 Genes

Gene sequences from *Streptomyces antibioticus* NRRL 8167 encoding a prolyl-*S*-dehydrogenase (*idmI*), a proline adenylyltransferase (*idmJ*) and a proline carrier protein (*idmK*) were obtained from GenBank (Accession number FJ545274) (Li *et al.*, 2009) and genes were synthesised and purchased from Genscript (New Jersey, USA). Synthetic gene DNA and amino acid sequences can be found in the appendix.

2.1.6 Oligonucleotides

Oligonucleotides (primers) were synthesised and purchased from Sigma Aldrich (Dorset, UK) or Integrated DNA Technologies (IDT) (Glasgow, UK). Primer sequences can be found in the appendix.

2.1.7 Media

Bacterial cultures were grown in 2 ×TY medium unless stated otherwise stated. This medium contained 16 g tryptone, 10 g of yeast extract and 5 g of NaCl made up to 1 L with distilled water. Bacterial cultures of IdmJ were grown in LB medium, containing 10 g tryptone, 5 g yeast extract and 10 g NaCl made up to 1 L with distilled water. TB medium contained 12 g of tryptone, 24 g of yeast extract and 4 mL glycerol made up to 900 mL, after sterilisation by autoclaving at 121 °C for 20 min, 100 mL of sterile filtered 0.17 M KH_2PO_4 0.72 M K_2HPO_4 was added prior to culturing. ZYP-5052 autoinduction medium contained 925 mL ZY (10 g tryptone, 5 g yeast extract made up to 925 mL with distilled water and 1 mL 1 M NaOH added, followed by sterilisation by autoclaving at 121 °C for 20 min), 1 mL 1 M $MgSO_4$ (filter sterilised), 20 mL 50 × 5052 (25 g glycerol, 73 mL distilled water, 2.5 g glucose and 10 g α-lactose filter sterilised) and 50 mL 20 × NPS (90 mL distilled water, 6.6 g $(NH_4)_2SO_4$, 13.6 g KH_2PO_4 and 14.2 g Na_2HPO_4 sterile filtered). Media was sterilised by autoclaving at 121 °C for 20 min. 1.5 % (w/v) agar was added to 2 × TY for solid phase medium.

Isotopically labelled protein was produced by growing bacterial cultures in minimal medium containing 10 g K_2HPO_4, 10 g KH_2PO_4, 7.5 g Na_2PO_4, 9 g K_2SO_4 and 1 g NH_4Cl made up to 1 L with distilled water. Before culturing, 100 µL 1 M $CaCl_2$, 2 mL 1 M $MgCl_2$ and 20 mL 20 % (w/v) glucose were added. $CaCl_2$ was sterilised by autoclaving and $MgCl_2$ and glucose were sterile filtered. For ^{15}N labelled protein NH_4Cl was substituted with $^{15}NH_4Cl$. For expression of ^{13}C, ^{15}N labelled protein the glucose was also substituted with 10 mL 20 % (w/v) D-glucose $^{13}C_6$.

2.2 Methods

2.2.1 General methods

2.2.1.1 pH measurements

pH measurements were completed using a Jenway 3020 pH meter, calibrated according to the Manufacturer's guidelines

2.2.1.2 Centrifugation

Centrifugation was performed using an Avanti® J-26 XP centrifuge (Beckman Coulter®, Buckinghamshire, UK), at 4 °C unless otherwise stated. Samples of volumes ≤ 1.5 mL were centrifuged using a SANYO MSE Micro Centaur Centrifuge (MSE, London, UK) at room temperature unless otherwise stated.

2.2.1.3 Asceptic technique

Standard asceptic techniques were employed throughout, media and heat resistant materials were sterilised by autoclaving at 121 °C for 20 min. Heat labile materials were sterilised by sterile filtering using a 0.22 µm Millex-GP syringe Tip filter purchased from EMD Millipore (Hertfordshire, UK) or wiped down with 70% (v/v) ethanol.

2.2.1.4 Antibiotic supplements

Ampicillin was made up as a 50 mg mL^{-1} stock in distilled water, sterile filtered and used at a final concentration of 50 µg mL^{-1} in solid phase and solution phase media.

2.2.1.5 Culture growth

Single colonies were picked from 2 × TY or LB agar plates supplemented with the correct antibiotic(s) and used to inoculate 5 mL 2 × TY or LB starter cultures in 50 mL Falcon tubes. These cultures were incubated for 16 hrs at 37 °C, shaking at 200 rpm in an orbital shaker. Unless stated otherwise, 50 mL day cultures were inoculated with 50 µL of the starter culture and incubated for 9 hrs at 37 °C, shaking

at 200 rpm in a 250 mL conical flask. 10 mL of the day culture was then used to inoculate 1 L 2 × TY medium in a 2.5 L conical flask. Protein expression was induced by the addition of IPTG, giving a final concentration of 0.1 mM unless stated otherwise. For optimal protein expression, the optical density (OD) at 600 nm was observed and the culture was induced when the cell density reached an OD of between 0.6 and 1.0.

2.2.1.6 Glycerol stocks

Glycerol stocks were used to streak out on 2 × TY or LB agar plates for culture growth (see 2.2.1.5). Glycerol stocks were made by adding 0.5 mL of starter culture to 0.5 mL sterile glycerol in a Nunc CryoTube™ (Thermo Fisher Scientific, Roskide, Denmark), mixed well and stored at -20 °C.

2.2.2 DNA methods

2.2.2.1 Plasmid purification

Plasmid DNA was purified from 5 mL cultures using Wizard® *Plus* SV Minipreps DNA purification system (Promega, Southampton, UK) as instructed in the protocol provided. Plasmid DNA was purified from 50 mL cultures using HiSpeed® Plasmid Midi Kits (QIAGEN, West Sussex, UK) as per manufacturer's guidelines.

2.2.2.2 Agarose gel electrophoresis

Unless stated otherwise agarose gels contained 0.7 % (w/v) agarose dissolved in 1 × TAE buffer and carried out according to Sambrook *et al.* (1989). A 1 kb ladder was used for determining the size of DNA fragments >2 kb, a 100 bp ladder was used for determining the size of DNA fragments <2 kb, DNA ladders were purchased from Promega (Southampton, UK).

10 × TAE buffer:

10 mM EDTA
200 mM Glacial acetic acid (17.5 M)

400 mM Tris base

2.2.2.3 Extraction of DNA from an agarose gel

DNA was purified from agarose gels by excising the bands and subsequently using a QIAquick® gel extraction kit (West Sussex, UK) according to the manufacturer's guidelines.

2.2.2.4 Polymerase chain reaction (PCR) amplification of DNA

Different PCR extension times were used depending on the target DNA to be amplified. However, *Pfu*Turbo Polymerase was used consistently. All reactions were carried out in thin walled PCR tubes. The reaction mixture was as follows:

5 µL of 10 µM forward oligo
5 µL of 10 µM reverse oligo
5 µL of 0.1-50 ng/µL template DNA
5 µL of 2 mM dNTPs
5 µL of 10 × Cloned *Pfu* DNA polymerase reaction buffer
25 µL of sterile water

The reactions were then incubated at 95 °C for 5 mins, 0.5 µL of *Pfu*Turbo Polymerase (2.5 U μL^{-1}) was added to the reaction and the following cycling conditions were adjusted and used dependent on the length of the target DNA. A list of relevant primers and sequences can be found in the appendix.

Generally, the PCR program after the addition of *Pfu*Turbo polymerase used was as follows:

92 °C, 2 mins

30 × (
92 °C, 30 secs
55-65 °C, 30 secs*
72 °C, 1 min per kb of target
)

72 °C, 10 mins

*The annealing temperature was determined by subtracting 5 °C off of the primer melting temperature (T_m). The T_m of the primer was calculated using the equation 2.1 below:

$$T_m = 64.9°C + 41.0°C \times (G + C - 16.4)/N$$

Equation 2.1- Equation to calculate the T_m of the binding region of primers used in PCR amplification of DNA. Where G + C are the number of G and C bases and N is primer length.

The concentration of DNA used in the reactions ranged from 0.1 to 10 ng μL^{-1}. To assess the success of PCR amplification 5 µL of the reaction was run on an agarose gel (section 2.2.2.2).

2.2.2.5 PCR product purification

PCR products were purified using a QIAquick® PCR purification kit (West Sussex, UK) as per manufacturers guidelines, however sterile water was used to elute the DNA.

2.2.2.6 DNA quantification

DNA was quantified either spectrophotometrically at a wavelength of 260 nm, or by running a 0.7 % (w/v) agarose gel. To quantify DNA on an agarose gel densitometry was used. The intensity of the 500 bp band in the 100 bp ladder (Promega, Southampton, UK) which contains 150 ng per 5 µL of sample was compared to the target DNA band. This can be done using the GeneTools software from Syngene (Cambridge, UK).

2.2.2.7 DNA digests

DNA digests were performed on a 20 µL scale unless stated otherwise, incubated for between 2 and 16 hrs at 37 °C. The reaction set up was as follows:

2 µL compatible reaction buffer supplied
4-7 µL DNA (up to 50 ng)
1 µL of restriction enzyme 1 (20 units)
1 µL of restriction enzyme 2 (20 units)

The reaction was made up to 20 µL with sterile deionised water. For single digests, one restriction enzyme was used and water adjusted accordingly. Where possible High-Fidelity (HF™) restriction enzymes were used.

2.2.2.8 Ligation reactions

DNA ligations were completed between a gene, termed the insert, and a target vector. Reactions were performed in thin-walled PCR tubes in a 10 µL reaction volume. Insert and vector DNA were incubated together in a molar ratio of 3:1 at 4 °C for 16 hrs in a reation containg 1 µL T4 DNA ligase and 1 µL of 10 × T4 DNA ligase buffer. Between 2 and 4.5 µL of the reaction was then transformed into XL10 Gold® Ultracompetent cells (section 2.2.2.11).

2.2.2.9 Ligation independent cloning

Ligation independent cloning was carried out using the FastCloning method as described in Li *et al.* (2011). Primers were designed to amplify the target insert (gene) and vector creating complimentary 5' and 3' ends of DNA fragments (2.2.2.4). An agarose gel was used to confirm the presence of the PCR products (2.2.2.2), which were then mixed in a 1:1 volumetric ratio. The reactions were digested with *Dpn*I for 1 hr and subsequently transformed into XL10 Gold® Ultracomptetent cells (2.2.2.10). Figure 3.28 in section 3.3.6 illustrates the FastCloning strategy.

2.2.2.10 Site directed mutagenesis

Site-directed mutagenesis was performed using a QuikChange® Site-directed mutagenesis kit (Stratagene, Cambridge, UK) as per the manufacturer's instructions. Primers were designed as recommended by the manufacturer. Site-directed mutagenesis primers can be found in the appendix.

2.2.2.11 Transformation into *E. coli* cells

Plasmid DNA was transformed into *E. coli* XL10 Gold® Ultracompetent cells and/or *E. coli* BL21-Gold (DE3) cells (Section 1.1.2) using the heat shock method according to the manufacturer's guidelines.

2.2.2.12 DNA sequencing

DNA sequencing was carried out by Beckman Coulter Genomics (Essex, UK). Sequencing primers used can be found in the appendix.

2.2.3 Protein methods

2.2.3.1 Small scale protein expression trials

Small scale protein expression was performed on a 5 mL scale. Glycerol stocks of *E. coli* BL21-Gold (DE3) cells containing the target plasmid were plated onto 2 × TY agar plates supplemented with ampicillin at a final concentration of 50 μg mL^{-1} (section 2.2.1.4), plates were incubated at 37 °C for 16 hrs. A single colony was then used to inoculate 5 mL 2 × TY starter culture supplemented with ampicillin (section 2.2.1.4), this was incubated at 37 °C for 16 hrs. 5 μL of the starter culture was then used to inoculate a 5 mL culture of TB, 2 × TY, LB or autoinduction medium containing ampicillin (section 2.2.1.4). TB, 2 × TY and LB cultures were incubated at 37 °C until the optical density of the culture at 600 nm reached between 0.6 and 1.0, cells were then induced by the addition of IPTG giving a final concentration of 60 μM. Cultures were then incubated for 16 hr at 37, 21 or 18 °C. Cells were then harvested for analysis of protein expression.

2.2.3.2 Sample preparation for analysis of protein expression

Cells were harvested from 5 mL expression cultures by centrifugation of 1.5 mL of the culture in the microfuge at 13 300 × *g* for 5 min. Harvested cells were then resuspended and incubated with 400 μL of CelLytic™ B lysis reagent for 10 min, followed by centrifugation at 13 300 × *g* for 5 min to separate the soluble and insoluble cellular components. Insoluble cellular components were resuspended in

100 µL of PBS. Samples were then prepared for analysis by SDS-PAGE and N-terminal His_6-tag detection by blotting with an anti-polyHistidine antibody (sections 2.2.3.3 and 2.2.3.6). Samples for SDS-PAGE were diluted 1:4 with sterile H_2O, samples for His_6-tag detection were not diluted.

2.2.3.3 Detection of the N-terminal His_6-tag

Detection of the N-terminal His_6-tag was carried out by blotting 2 µL of soluble or insoluble cellular samples onto a nitrocellulose membrane. The blotted membrane was left to dry for 20 mins, then incubated for 1 hr at room temperature in blocking buffer. The blocking buffer was then removed and the membrane incubated with 10 mL of primary antibody solution for 1 hr. The membrane was then washed with PBS-T for 10 mins; this was repeated 2 more times. The membrane was then incubated with 10 mL of secondary antibody solution for 45 mins at room temperature. The membrane was washed with PBS-T for 10 mins (3 times), followed by a 10 min wash in PBS. The membrane was then visualised by incubation with ECL Prime Western blotting detection reagent (VWR International Ltd, Leicestershire, UK) for 1 min, with increasing levels of exposure using a GeneGnome^5 system (Syngene, Cambridge, UK).

For detection of the N-terminal His_6-tag by western blot an SDS-PAGE gel was run as described in section 2.2.3.6. The SDS-PAGE gel was then placed in transfer buffer for 15 mins. Transfer of proteins from the SDS-PAGE gel to a nitrocellulose membrane was performed using a Mini Trans-Blot® Electrophoretic Transfer Cell (Bio-Rad, Hertfordshire, UK). The transfer stack was assembled as instructed by the manufacturer, transfer was carried out at 100 V for 1 hr. Blocking was performed as described above, however the blocking was carried out at 4 °C and the blocking buffer contained 3 % BSA as opposed to 5 % skimmed dried milk powder. The antibody solution incubations, washing and visualisation was carried out as described above.

PBS-T:

1 × PBS
0.05 % (v/v) Tween-20

Blocking buffer:

PBS-T
5 % skimmed dried milk powder or 3 % BSA

Antibody solution:

Blocking buffer
Antibody at appropriate dilution (section 2.1.4)

Transfer Buffer:

24 mM Tris base
194 mM Glycine
20 % (v/v) methanol

2.2.3.4 Nickel affinity purification of His_6-tagged proteins

E. coli cells were cultured, as described in section 2.2.1.5, containing the plasmid with the target gene for protein expression. Protein purification was performed in batch using Chelating Sepharose™ fast flow resin chelated with nickel ions. *E. coli* cells were harvested by centrifugation (2.2.1.2) at 12 000 *g* for 20 mins. The supernatant was discarded and the pellet re-suspended using a homogeniser in washing buffer, 10 g of cells in 50 mL buffer. The cells were then lysed using a cell disrupter from Constant Systems Ltd (Northamptonshire, UK) at 20 KPSI. The insoluble cell debris was then removed by centrifugation at 40 000 *g* for 45 mins. The supernatant containing the soluble target protein was then subjected to purification by nickel affinity chromatography.

The supernatant was incubated with the 5 mL of nickel resin in a 50 mL Falcon tube for 30 mins with mild agitation at 4 °C. The resin was then centrifuged at 4 000 *g* for 10 mins, the supernatant was discarded. The resin was then washed by the addition of washing buffer to a final volume of 50 mL, re-suspended thoroughly, then centrifuged at 4 000 *g* for 10 mins and the supernatant removed. This wash step was repeated 4 times. Bound protein was eluted by the addition of 15 mL of elution buffer to the resin and incubated for 1 hr at 4 °C with mild agitation; the resin was centrifuged at 4 000 *g* for 10 mins and the supernatant, containing the protein, was

collected. The eluate was then dialysed into 50 mM Tris.HCl pH 7.4 as described in section 2.2.3.5. After dialysis protein was sterile filtered using 0.22 µm Millex-GP syringe tip filters and stored at 4 °C. Sodium dodecyl sulphate polyacrylamide gel electrophoresis (2.2.3.6) and mass spectrometry (2.2.3.7) was used to confirm the presence of the correct protein.

Protein purification washing buffer:

50 mM Tris.HCl pH 7.5
20 mM Imidazole
0.5 M NaCl

Protein purification elution Buffer:

50 mM Tris.HCl pH 7.5
0.5 M Imidazole
0.5 M NaCl

2.2.3.5 Dialysis

Dialysis was performed at 4 °C in volumes 50-100 times larger than the sample, with mild stirring. For proteins smaller than 14 000 Da dialysis tubing with a molecular weight cut off of 3 000 Da was used, otherwise tubing with a molecular weight cut off of 14 000 Da was used. Dialysis tubing was purchased from Fisher Scientific Ltd. (Loughborough, UK). Briefly, samples for dialysis were placed in dialysis bags in the dialysis buffer. Samples were left to equilibrate for 4 hrs, and then the dialysis bag was transferred to freshly made dialysis buffer and again left for 4 hrs to equilibrate. After equilibration, the sample was sterile filtered using 0.22 µm Millex-GP syringe tip filters.

2.2.3.6 Reducing sodium dodecyl sulphate polyacrylamide gel electrophoresis

Reducing sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was carried out using the standard method (Laemmli, 1970). For approximate molecular weight determination the Prestained Protein Marker, Broad Range (7-175 kDa) ladder was used (NEB, Ipswich, MA, USA). These contain the protein

markers: *E. coli* MBP-β-galactosidase (175 kDa), *E. coli* MBP-paramyosin (80 kDa), *E. coli* MBP-CBD (58 kDa), *E. coli* CBD-*Mxe* Intein-2CBD (46 kDa), *E. coli* CBD-*Mxe* Intein (30 kDa), *E. coli* CBD-BmFKBP3 (25 kDa), chicken egg white lysozyme (17 kDa) and bovine lung aprotinin (7 kDa). SDS-polyacrylamide gel electrophoresis was carried out as described in Sambrook *et al.* (2001) using a 15% acrylamide resolving gel and 3.75% acrylamide stacking gel (recipe listed below). Ammonium persulfate (APS) was made up freshly for usage. The two polymerisation initiation components, APS and TEMED, were added to the gel solutions last. Gels were set between two glass plates as *per* manufacturer's guidelines. The resolving gel was poured immediately followed by an isopropanol layer and allowed to set. The isopropanol was removed and the stacking gel poured and a comb inserted. Once the stacking gel was set, gels could be used immediately or wrapped in damp laboratory roll and covered in cling film. Wrapped gels could be stored at 4 °C for up to 1 week.

Resolving gel:

7.5 mL 30 % (w/v) acrylamide
3.75 mL 1.5 M Tris.HCl pH 8.8
3.5 mL H_2O
150 µL 10 % (w/v) SDS
50 µL 25 % (w/v) APS
5 µL TEMED

Stacking gel:

625 µL 30 % (w/v) acrylamide
625 µL 1 M Tris.HCl pH 6.9
3.65 mL H_2O
50 µL 10 % (w/v) SDS
50 µL 25 % (w/v) APS
5 µL TEMED

2 × loading buffer:

154 mg DTT
2 mL 10 % (w/v) SDS

1 mL glycerol
170 µL 1 M Tris.HCl pH 6.9
163 µL H_2O
200 µL 0.2 % (w/v) bromophenol blue in ethanol

Running buffer:

3 g Tris Base
14.4 g glycine
1 g SDS
140 µL β-mercaptoethanol
Made up to 1 L with H_2O

Protein samples were prepared by adding 2 × loading buffer in a 1:1 ratio and boiled at 100 °C for 5 mins. Electrophoresis was performed at 30-60 mA for approximately 1-2 hrs. SDS-PAGE gels were then stained in a methanol/ acetic acid/ water (5/1/1; v/v/v) solution containing 0.1 % (w/v) Coomassie Brilliant Blue for 2 hrs and subsequently destained using a methanol/acetic acid/water (5/1/1; v/v/v) solution. Alternatively, for rapid determination of results, gels were stained using InstantBlue (Expedeon, Cambridgeshire, UK) or Quick Blue Coomassie Stain (Triple Red, Buckinghamshire, UK). InstantBlue was used for in-gel tryptic digests (section 2.2.3.10).

2.2.3.7 Determination of protein concentration

Protein concentration was determined spectrophotometrically at 280 nm and calculated using the Beer-Lambert law (Equation 2.2).

$$A = \varepsilon c l$$

Equation 2.2- Beer-Lambert law, where A is the absorbance, ε is the molar extinction coefficient, c is the concentration of the protein and l is the path length.

The molar extinction co efficient for each protein was calculated using their amino acid sequence and the ExPASy ProtParam tool. The calculated molar extinction coefficient was used throughout this research. The molar extinction coefficient calculated by ExPASy ProtParam for IdmJ was 51130 M^{-1} cm^{-1}, for IdmI was 25940 M^{-1} cm^{-1} and for IdmK was 5500 M^{-1} cm^{-1}.

2.2.3.8 Increasing protein concentration

Protein samples required to be at a higher concentration were concentrated by centrifugation through using a centrifugal concentrator. For sample volumes larger than 0.5 mL Sartorius Stedium Vivaspin® 6 or 20 centrifugal concentrators with a molecular weight cut-off of either 3 kDa or 10 kDa depending on the proteins' molecular weight. For samples smaller than 0.5 mL Sartorius Stedium Vivaspin® 500 centrifugal concentrators were used, with a molecular weight cut-off of 5 kDa (Goettingen, Germany). Concentrators were used in accordance with the manufacturer's guidelines.

2.2.3.9 Sample preparation for mass spectrometry

Buffer exchange or desalting of protein samples was carried out using either Zeba™ Spin Desalting Columns purchased from Thermo Fisher Scientific (Loughborough, UK) or Bio-Spin Exclusion Columns from Bio-Rad (Hertfordshire, UK) according to the manufacturer's guidelines. For analysis by mass spectrometry, protein samples were exchanged into 50 mM ammonium acetate pH 7.4.

2.2.3.10 In-gel tryptic digestion and analysis by LC-MS/MS

In-gel tryptic digestion and analysis LC-MS/MS was carried out by Dr James Ault (Astbury Centre, University of Leeds, UK). Proteins to be analysed by in-gel tryptic digestion were run on a 15 % acrylamide gel as described in section 2.2.3.6. The gel was stained with InstantBlue (Expedeon, Cambridge, UK) to identify the target band. The band/s were excised and split into ~ 1 mm pieces. To remove the InstantBlue stain, gel pieces were immersed in 30 % ethanol in a 0.5 mL Eppendorf tube then incubated at 60 °C for 30 mins with shaking. Gel pieces were incubated in 100 % acetonitrile for 5 mins, followed by 40 µL 20 mM DTT in 25 mM ammonium bicarbonate at 56 °C for 1 hr with shaking. The supernatant was removed, substituted with 40 µL of 55 mM iodoacetamide in 25 mM ammonium bicarbonate, which was then incubated 37 °C for 45 min shaking, in the dark. The supernatant was removed and gel slices washed with 25 mM ammonium bicarbonate for 10 mins at room temperature, this was followed by a 5 min incubation in 100 % acetonitrile. Gel slices were then removed from the acetonitrile and left to dry completely for 1 hr

under a laminar flow hood. Once the gel pieces were dry they were cooled on ice and subsequently incubated with an ice-cold solution of 20 ng/µL trypsin in 25 mM ammonium bicarbonate for 10 mins, on ice. The supernatant was removed and 40 µL of 25 mM ammonium bicarbonate added to the rehydrated gel slices. The gel slices were briefly vortexed and centrifuged then incubated at 37 °C for 18 hrs. The digested peptides were recovered by vortexing and centrifugation of the solution, incubated for 18 hrs and the supernatant transferred to a fresh 0.5 mL Eppendorf tube. 5 µL of wash solution (acetonitrile/water/formic acid (60/35/5; v/v)) was added to the peptide solution. Additional peptide recovery from the gel pieces was performed by incubation of the gel slices with 40 µL of wash solution for 10 mins at 37 °C, the resulting supernatant was combined with the original peptide recovery solution; this was repeated a second time. The peptide solution was then dried by vacuum centrifugation. In preparation for separation by liquid chromatography (LC) and analysis by MS/MS, the dried peptides were resuspended in 20 µL acetonitrile/ water/ formic acid (2/97.9/0.1; v/v/v).

In order to separate the peptides, LC was carried out on the peptide mixture using an Ultimate 3000 nano LC system (Dionex, Amsterdam, Netherlands). 2 µL of the peptide mixture was loaded onto a C18 guard column. The column was then washed with 2 % acetonitrile/ 0.1 % formic acid for 5 min at 25 µL min^{-1}. The peptides were then separated by performing a gradient of 2-40% of 0.1% formic acid in acetonitrile with 0.1% formic acid in water over 60 min at 0.35 µL min^{-1}, on a PepMap C18, 100 µm i.d. x 15 cm analytical column (Dionex, Amsterdam, NL).

The peptides separated on the column were then analysed directly by a quadrupole-ion mobility - orthogonal time of flight mass spectrometer (Synapt G2-S, Waters UK, Manchester, UK) via Nano Flow electrospray ionisation (ESI). The mass spectrometer (MS) was operated in positive mode time of flight (TOF), using a capillary voltage of 3.8 kV and a sampling cone voltage of 20 V, source offset of 15 V and a trap bias of 5 V and backing pressure of 8.47 mbar. The buffer gas utilised was argon at a pressure of 5.0×10^{-4} mbar in both the trap and transfer regions. The source temperature was 80°C. Mass calibration was carried out through a separate injection of sodium iodide (2 µg/µL). A 1 µM solution of the peptide mass standard Glu-1- Fibronopeptide B (Glu-Fib) in acetonitrile/water/formic acid (50/49/1; v/v/v) was introduced as the lock mass calibrant at a flow rate of 1 µL min^{-1} with a one sec

lock-spray scan collected every 30 sec during data acquisition. To determine the lock mass calibration correction factor ten scans were averaged. Data acquisition was carried out using a data dependent analysis. A one sec MS over *m/z* 350-2000 was collected followed by three one sec MS/MS performed over *m/z* 100-2000 of the three most abundant ions in the MS spectrum. Capillary electrophoresis applied to the ion was dependent on its charge state and mass. For increased sequence coverage dynamic exclusion (60 sec) was used.

Data was processed using the MassLynx v4.1 suite of software supplied with the mass spectrometer. Peptide MS and MS/MS data were processed with ProteinLynx Global Server v3.0.2 software (Waters UK, Manchester, UK) and searched against reviewed entries of the UniProtKB/SwissProt database (release 2014_03), the database was modified to include the amino acid sequence of the synthetic IdmJ protein (amino acid sequence located in the appendix).

2.2.3.11 Cysteine alkylation

Alkylation of cysteines was performed using a 20-30 µM protein sample in 50 mM Tris.HCl pH 7.4 containing 8 M urea on a 500 µL scale. To reduce samples, DTT was added to a final concentration of 5 mM from a 0.5 M stock and incubate at 55 °C for 25-45 mins. The protein was allowed to cool to room temperature. 500 mM iodoacetamide was then added to a final concentration of 14 mM and incubated for 30 mins at room temperature in the dark. The reaction was then quenched by either adding DTT to give a final concentration of 10 mM or by preparing the samples for analysis by mass spectrometry by exchanging the buffer (2.2.3.9).

2.2.3.12 Mass spectrometry

Samples to be analysed by mass spectrometry were desalted into 50 mM ammonium acetate as described in section 2.2.3.9. Samples were then analysed by nano-ESI-MS by a quadrupole-ion mobility spectrometry-orthogonal TOF MS (Synapt, HDMS Waters UK Ltd., Manchester, UK) operating in positive TOF 'V' mode. Samples were analysed in acetonitrile : 1% aqueous formic acid (50:50; v/v). The MS was operated with a capillary voltage of 1.2 kV and a cone voltage of 50 V, the

nanoelectrospray nitrogen pressure was was set to 0.1 mbar with a backing pressure of 1.78 mbar. Nitrogen was used as a buffer gas at a pressure of 8.0×10^{-3} mbar in the trap and transfer regions and 3.6×10^{-4} in the ion mobility cell. The source temperature was 80 °C and desolvation temperature 150 °C. Mass calibration was carried out through a separate injection of sodium iodide (2 µg/µL) in 50:50 (v/v) water: acetonitrile. Data was processed using the MassLynx v4.1 suite of software supplied with the mass spectrometer. Analysis of protein samples by mass spectrometry were carried out by Dr James Ault (Astbury Centre, University of Leeds, UK).

2.2.3.13 Protein purification by size exclusion chromatography

For increased homogeneity, protein samples purified by nickel affinity chromatography (section 2.2.3.4) were subject to further purification by size exclusion chromatography. This was performed using an ÄKTA Prime Purifier System and a HiLoad™ Superdex™ S75 prep grade column (GE Healthcare Life Sciences, Buckinghamshire, UK). The column was equilibrated with a selected buffer which had been filtered and degassed. 3 mL of protein purified by nickel affinity chromatography at a concentration of 8 mg/mL was loaded onto the column using a 5 mL sample loop (GE Healthcare Life Sciences). The column was run at a flow rate of 2.0 mL/min. 2 mL fractions were collected and analysed by SDS-PAGE (2.2.3.3), fractions containing the purified protein were pooled and concentrated (2.2.3.5).

2.2.4 Crystallographic methods

2.2.4.1 Sample preparation

A homogeneous protein sample was prepared through purification by nickel affinity chromatography (2.2.3.4) and subsequent size exclusion chromatography (2.2.3.13). The protein sample was sterile filtered using 0.22 µm Millex-GP syringe tip filters purchased from EMD Millipore (Hertfordshire, UK) and concentrated (2.2.3.8) to 12 mg/mL.

2.2.4.2 Screening conditions and crystal trays

Crystallisation condition screens were carried out using 96 well CrystalQuick™ RW sitting drop trays purchased from Greiner Bio-One (Gloucestershire, UK) with a 10 mg/mL protein sample. The commercially available screens used were Crystal screen 1 and 2, Index 1 and 2, Salt RX (Hampton Research, Califronia, USA) and Wizard 3 and 4 (Emerald BioStructure, Washington, USA).

Crystallisation conditions identified for the prolyl carrier protein were pH 8.2, 0.056 M sodium phosphate monobasic monohydrate, 1.344 M potassium phosphate dibasic. Crystals were grown in 24 well plates by hanging drop vapour diffusion. Drops were dispensed in a 1:1, 2:1 and 1:2 crystallisation buffer to protein ratio. Protein samples were of a concentration 10 mg/mL.

2.2.4.3 Data acquisition and processing

Data acquisition and processing was carried out by Dr Chi Trinh (Astbury Centre, University of Leeds, UK). For data acquisition crystals were prepared by immersing in 25% (v/v) glycerol as a cryo-protectant, prior to flash cooling in liquid nitrogen. High resolution diffraction data was collected at 100 K on the I04 beamline Diamond Light Source Ltd (DLS) (Oxfordshire, UK). The data were reduced using the automated processing suit Xia2 at DLS (Winter, 2010). Phasing by molecular replacement was attempted using both Molrep (Vagin and Teplyakov, 1997) and Phaser (Mccoy *et al.*, 2007) with different homologous carrier proteins as the initial search model (PDB codes: 2FAE (Roujeinikova *et al.*, 2007), 2LIU (Busche *et al.*, 2012), 2LKI (Srisailam *et al.*, 2006), 2XZ1 (Guy *et al.*, 2011), 3GZL (Gallagher and Prigge, 2010), 2LPK (Ramelot *et al.*, 2012), 1X3O (unpublished, Riken Structural Genomics/Proteomics), 2EHT (unpublished, Riken Structural Genomics/Proteomics), 2QNW (unpublished, Structural Genomics Consortium) and 2L4B (unpublished, Seattle Structural Genomics Center for Infectious Disease). SWISS-MODELLER (Arnold *et al.*, 2006; Biasini *et al.*, 2014; Guex *et al.*, 2009; Kiefer *et al.*, 2009) and ITASSER (Roy *et al.*, 2010; Zhang, 2008) were also used to generate models from the IdmK amino acid sequence.

2.2.5 Nuclear magnetic resonance (NMR) spectroscopy methods

2.2.5.1 Sample preparation

Uniformly [^{15}N] and [^{13}C, ^{15}N] labelled protein was prepared by growing the bacterial cultures in minimal medium (2.1.7) using $^{13}C_6$ glucose and $^{15}NH_4Cl$ as the sole carbon and nitrogen sources. Labelled protein was expressed and purified as described in sections 2.2.1.5, 2.2.3.4 and 2.2.3.13. Mass spectrometry (section 2.2.3.12) was used to determine the extent of isotopic labelling. Protein samples were dialysed into 50 mM Tris.HCl pH 7.4, 0.02% (*w/v*) sodium azide (section 2.2.3.5). Samples to be used for NMR spectroscopy were concentrated to 0.5 mM and 10% (*v/v*) D_2O was added. Shegemi NMR tubes were used for data acquisition with samples containing 300 µL of the protein sample.

2.2.5.2 Data acquisition

2D and 3D NMR experiments were carried out at 25°C on Varian INOVA spectrometers (Varian Inc., California, USA) at 500 or 600 MHz equipped with room temperature probes or at 750 MHz with a cryogenically cooled probe. Table 2.1 shows the experimental details used to collect each spectrum.

2.2.5.3 Data processing and analysis

All data collected was processed with NMRPipe software (Delaglio *et al.*, 1995). Typical data processing included multiplication by a cosine bell, followed by zero filling to increase the number of points at least by a factor of 2 followed by rounding. Fourier transformation was performed in all dimensions. In the case of the final ^{15}N or ^{13}C dimension, to double the number of data points, mirror image linear prediction was carried out. Phase corrections in the direct dimension were manually adjusted using NMRDraw and NMRPipe (Delaglio *et al.*, 1995). In most cases phase corrections were not required in the indirect dimension due to use of an optimal initial delay time, however, when required they were adjusted in the same manner as the direct dimension. Data were converted to NMRView format by NMRPipe for use with CCPNmr analysis (Vranken *et al.*, 2005).

An initial ^{1}H-^{15}N heteronuclear single quantum coherence spectrum (HSQC) was collected to assess the suitability of NMR for structural determination and further HSQC spectra were collected for sample quality control before or after collection of datasets. Backbone assignments were completed using the following spectra: HNCA, HNcoCA, HNcaCO, HNCO, HNCACB and HNcoCACB. Aromatic residue assignments were completed using 2D aromatic filtered ^{1}H-^{13}C-NOESY-HSQC and 2D hbCBcgcdHDceHE spectra (Marion *et al.*, 1989a; Marion *et al.*, 1989b; Zuiderweg and Fesik, 1989; Yamazaki *et al.*, 1993; Kalverda *et al.*, 2009). A 2D aromatic filtered ^{1}H-^{13}C-NOESY-HSQC spectrum was acquired by using a standard ^{1}H-^{13}C NOESY-HSQC pulse sequence (Biopack) with no increments for ^{13}C and the ^{1}H-^{13}C one-bond transfer time as 14.5 ms. Assignments of aliphatic amino acid side chain were completed using ^{1}H-^{13}C HSQC, HcCH-TOCSY and hCCH TOCSY spectra. Finally distance restraints were measured using ^{15}N-NOESY-HSQC and ^{13}C-NOESY-HSQC spectra.

Backbone assignments were completed by a semi-automated method using MARS (Jung and Zweckstetter, 2004) and checked manually using strip plots. Secondary structure predictions for use with MARS were calculated using the PSIPRED server at http://bioinf.cs.ac.uk/psipred (Buchan *et al.*, 2013; Jones, 1999). TALOS was used to predict Φ and Ψ torsion angles and assigned from $^{1}H\alpha$, $^{13}C\alpha$, $^{13}C\beta$, CO and ^{15}N nuclei from backbone assignments to assign secondary structural preferences (Shen *et al.*, 2009a).

Chemical shift referencing was derived from the chemical shift of the water line assuming a temperature of 25°C and a pH of 7.4 (Cavanagh *et al.*, 2007). Following backbone assignment the chemical shift referencing was further validated using the program CheckShift (Ginzinger *et al.*, 2007). Chemical shift referencing with the referencing compound DSS is still required.

Experiment	Spectral frequency (^{1}H) MHz	SW (Hz)			Number of (complex) data points			nt	t_{mix} (s)
		^{1}H	^{13}C	^{15}N	^{1}H	^{13}C	^{15}N		
^{1}H-^{15}N HSQC	500	8000	-	1800	2048	-	256	16	-
HNCO	600	7620	2170	1940	2048	32	32	4	-
HNcaCO	600	7620	2170	1940	2048	32	32	16	-
HNCA	600	7620	4530	1940	2048	32	32	4	-
HNcoCA	600	7620	4530	1940	2048	32	32	8	-
HNCACB	600	7620	12100	1940	2048	32	32	32	-
CBCAcoNH	600	7620	12100	1940	2048	48	32	8	-
^{1}H-^{13}C HSQC	750	10480	15000	-	2048	400	-	8	-
^{1}H-^{13}C HSQC (aromatic)	750	10480	15000	-	2048	92	-	64	-
HcCH-TOCSY	750	10480	8000	15700	1024	128	54	4	-
hCCH-TOCSY	750	10500	13000	13000	2048	56	72	8	0.14
hbCBcgcdHDceHE	500	8000	3770	-	1024	28	-	2592	-
HccoNH	600	7600	5400	1940	1024	64	32	16	-
HBHAcoNH	600	9620	4750	1940	2048	62	32	16	-
^{1}H-^{15}N J modulated HSQC	600	7620	-	1990	1616	-	128	32	-
Experiment	**Spectral frequency (^{1}H) MHz**	**SW (Hz)**			**Number of (complex) data points**			**nt**	**t_{mix} (s)**
		^{13}C/^{15}N	^{1}H	^{1}H	^{13}C/^{15}N	^{1}H	^{1}H		
^{15}N-NOESY-HSQC	600	12000	8200	2050	2048	144	40	8	0.12
^{13}C-NOESY-HSQC (aromatic)	750	10500	8250	5800	2048	300	0	256	0.15
^{13}C-NOESY-HSQC (aliphatic)	600	9620	6900	6190	2048	182	62	4	0.1

Table 2.1- Experimental parameters of spectra acquired on the Varian INOVA spectrometer (Varian Inc., California, USA). SW-spectral width, nt- number of transients and t_{mix}-mixing time (s).

2.2.5.4 Structure generation

Structure determination by CS-ROSETTA was aided by Dr Gary Thompson. An initial structure ensemble was generated using CS-ROSETTA 3.2 (Shen *et al.*, 2008; Shen *et al.*, 2009b), using H, HN, CA, CB and CO chemical shifts from the backbone assignments and the amino acid sequence of IdmK. A total of 5000 structures were calculated on a cluster of 7-14 CPU processors. Residual dipolar coupling (RDC) Q factor and combined chemical shift and ROSETTA energy scores were used to score the ensemble of structures. RDC Q values were calculated using PALES as described in section 2.2.5.5 and 5.3. Further validation was carried by comparison with known canonical acyl carrier protein folds (PDB codes: 1HY8, 1T8K, 1VKU, 2AVA, 2CGQ, 2GDW, 2KOO, 2KWL and 2L0Q) (Xu *et al.*, 2001; Qiu and Janson, 2004; Zornetzer *et al.*, 2006; Koglin *et al.*, 2006; Ploskon *et al.*, 2010; Barnwal *et al.*, 2011; Chan *et al.*, 2010) and structure homology searches were completed using the Dali server (http://ekhidna.biocenter.helsinki.fi/dali_server/) (Holm and Rosenstrom, 2010).

To validate the ROSETTA fold and generate a high resolution NOE (nuclear Overhauser effect) derived structure of IdmK, refinement of an ensemble of structures from the ROSETTA calculation was carried out using ARIA/CNS (Rieping *et al.*, 2007). The structure ensemble was selected by taking the structure with the best fit for secondary structure elements to the measured ^{1}H-^{15}N RDC data as measured by the RDC Q factor. A set of 20 structures were selected with the lowest RMSD for the backbone secondary structure elements when compared to this structure. ROSETTA structures were regularised using the xplor-nih script (AddAtoms.py) (Schwieters *et al.*, 2006) with all backbone atoms fixed during minimisation (Schwieters *et al.*, 2003; Schwieters *et al.*, 2006).

ARIA structure calculations were carried out by Dr Gary Thompson. ARIA structures were calculated using the best fitting ROSETTA structures as a seed for structure calculation, replacing those generated by ARIA in the first of eight rounds of calculations (it0). The hot and cooling/annealing phases in the ARIA structure calculation were extended by a factor of 4 in length (Fossi *et al.*, 2005b). The final structures were refined in a box of water. Chemical shift restraints were included as Φ and Ψ dihedral restraints for regions of secondary structure showing consistent

TALOS predictions (Rieping *et al.*, 2007). RDC Q values were calculated using Equation 2.3 (Cornilescu *et al.*, 1998).

$$Q = \sqrt{\frac{\sum_i \left(RDC_i^{obs} - RDC_i^{calc}\right)^2}{\sum_i (RDC_i^{obs})^2}}$$

Equation 2.3 - Equation to calculate the agreement between experimental and calculated RDCs (Q factor), where i is the residue (Zweckstetter, 2000).

2.2.5.5 RDC alignment media and measurements

Residual dipolar couplings were measured in two different liquid crystalline alignment media, PF1 filamentous bacteriophage and an alkyl-poly (ethylene-glycol)/*n*-alkyl alcohol mixture.

For alignment of IdmK in Pf1 bacteriophage (ASLA Biotech, Estonia), an initial ^{1}H-^{15}N HSQC was collected (isotropic) then Pf1 bacteriophage was added to a final concentration of 8 mg/mL. A ^{1}H-^{15}N HSQC was collected and processed as described in section 2.2.5.3. This spectrum was used to determine the suitability of Pf1 bacteriophage as an alignment medium.

A second alignment medium was trialled. The alkyl-poly (ethylene-glycol)/*n*-alkyl alcohol mixture was used. Hexanol, the *n*-alkyl alcohol selected, and C12E6 alkyl-poly (ethylene-glycol) (PEG) alignment medium was made at a (2x concentration) by titrating hexanol into a solution of 10% C12E6 using 5μL and 0.5 μL MicroVolume syringes (SGE Analytical Science, Milton Keynes, UK). The magnitude of alignment of the dipolar coupling medium was monitored using the quadrupolar splitting of HDO peak in the D^2 (deuterium) NMR spectrum from the D_2O lock solvent. For the Hexanol/PEG medium the quality of the alignment was followed by monitoring the lineshape in the sample, which showed a clean doublet with a splitting of 29 Hz when complete alignment was achieved at a PEG concentration of 10%. Final samples were prepared by adding 50% of the hexanol/PEG solution to the NMR sample. After alignment of the protein the deuterium splitting was checked again and adjustments to achieve a clean spectrum with a deuterium splitting of ~25Hz were made with hexanol or PEG solution. The quality of the data and deterioration of the spectral quality due to over alignment and

increases in the rotational correlation time from the aligned samples was also monitored using ^{1}H-^{15}N HSQC spectra and ^{1}H coupled ^{1}H-^{15}N HSQC spectra. RDC Measurements were made using ^{1}H-^{15}N J modulated HSQC spectra (Tjandra *et al.*, 1996) with $^{1}J_{HN}$ evolution delays (Δ) of 0.001 (×2), 0.003, 0.005, 0.007, 0.009, 0.011, 0.013, 0.015 (×2), 0.017, 0.019, 0.021 and 0.023s. Data were processed with NMRpipe and peak intensities were measured using NMRView 5.2. Fitting of peak intensity data was achieved using the in house python script fitR_bs (Dr Gary S Thompson and Dr Christopher MacRaild) which carries out non-linear least squares fitting starting with initial fitting parameters estimated using a number of heuristics. Statistical error analysis was carried out via 1000 rounds of Monte-Carlo simulations, with noise derived from differences in peak intensities from the 2 data sets measured with duplicate delays. The data was fitted to the equation 2.3.

$$I_{2\Delta} = I_0[-A + \cos(2\pi {}^{1}J_{NH}\Delta)] \exp\left(-\frac{2\Delta}{T2}*\right)$$

Equation 2.4- With I_0 being the initial peak amplitude, Δ(t) the time delay (s), A the unmodulated component of the peak intensity, $^{1}J_{NH}$ the measured one-bond coupling (with or without the dipolar component) and T2* the effective transverse relaxation rate (Tjandra *et al.*, 1996).

Tensor frames, predicted RDCs, magnitude of alignment (Da), rhombicity (R) and quality of fit (Q) for RDCs were calculated using PALES (Zweckstetter and Bax, 2000). PALES fits the observed dipolar couplings to the alignment tensor frame, D_a and R; values from which predicted dipolar couplings can be back calculated using the Equation 2.4.

$$D = D_a\left[(3\cos^2\theta - 1) + \frac{3}{2}R\sin^2\theta\cos(2\varphi)\right]$$

Equation 2.5-Where θ and φ are polar angles with respect to the alignment tensor frame (D_a and R) (Schwieters *et al.*, 2003; Schwieters *et al.*, 2006).

2.2.5.6 Initial structural investigations of *holo*-IdmK

To investigate structural differences of *holo*-IdmK and *apo*-IdmK experiments were run using ^{15}N *holo*-IdmK. To ensure monomeric-labelled *holo*-IdmK was present experiments were run in the presence of 1 mM DTT. An initial ^{1}H-^{15}N HSQC was

collected of ^{15}N *holo*-IdmK. Conservative chemical shift perturbation mapping was used in order to identify sites of interaction of the post-translational modification without assignment of the *holo* state (Williamson *et al.*, 1997). Combined ^{1}H, ^{15}N chemical shift differences were calculated Equation 2.5.

$$\Delta = \left[\left(\delta^{15}\mathrm{N}_{apo/holo} \right)^{1/2} + \left(5 \times \delta^{1}\mathrm{H}_{apo/holo} \right)^{2} \right]^{1/2}$$

Equation 2.6- Calculation of the differences in chemical shift from the ^{1}H-^{15}N HSQC spectra of *holo*- and *apo*- IdmK (Williamson *et al.*, 1997).

3. Cloning and recombinant protein expression of nonribosomal peptide synthetase proteins required for starter unit biosynthesis of the polyether ionophore indanomycin

The native producers of natural products such as nonribosomal peptides and polyketides include organisms like gram-positive bacteria and fungi (Stachelhaus and Marahiel, 1995). In this thesis the target compound, indanomycin, is produced by a strain of *Streptomyces, Streptomyces antibioticus* NRRL 8167 (Liu *et al.*, 1979). Over-expression of NRPS proteins, required for structural and biochemical characterisation, is problematic in *Streptomyces*. Issues which prevent the over-expression of proteins in the native system include slow bacterial growth during incubation and low product titres (Fujii, 2009; Peirú *et al.*, 2010). An additional disadvantage is the genetic limitations of such an organism, including the occurrence of native PKS/NRPS gene clusters which may contaminate target protein PKS/NRPS expression and purification, the complexity of DNA manipulation with slow growing bacteria, and a lack of tools for genetic engineering (Fujii *et al.*, 2009; Peirú *et al.*, 2010; Cane *et al.*, 1998; Weissman and Leadlay, 2005).

Over-expression of NRPS and PKS proteins has been carried out in some *Streptomyces* strains. Expression of these multimodular enzymes has been preferable due to the already existing abilities of these organisms to produce such macromolecules and the substrates required (Weissman and Leadlay, 2005). Expression is usually carried out in strains that are fully sequenced such as *S. coelicolor* A3 (2) (Bentley *et al.*, 2002) or *S. avermilitis* (Ikeda *et al.*, 2003).

Alternative strategies explored for the expression of NRPS or PKS proteins in *Streptomyces* includes genetic engineering to reduce contamination by native PKS/NRPS gene clusters, by targeted deletion of the native NRPS/PKS gene cluster (such as the deletion of the actinorhodin gene cluster in its producing strain, *S. coelicolor* CH999) (McDaniel *et al.*, 1993). Another approach to reduce the background genetic noise was to perform systematic deletions of the genome to create a genome-minimised bacterial strain. A genome minimised *S. avermilitis* bacterial strain was used to produce the antibiotic streptomycin, at levels above those

found in the native host *S. griseus* (Komatsu *et al.*, 2010; Waksman *et al.*, 2010). However more recently *E. coli* has become a popular choice of host system for recombinant protein expression due to its rapid life cycle, well characterised metabolic pathways, culture times, cost and ease of DNA manipulation (Fujii *et al.*, 1999; Rosano and Ceccarelli, 2014).

Recombinant protein expression of NRPS and PKS modules in *E. coli* is not without its pitfalls, one particular difficulty being the expression of post-translationally modified proteins, which is required for the expression of the PKS/NRPS carrier protein. This issue was addressed by genetic engineering of *E. coli* to introduce the gene encoding the *B. subtilis* phosphopantetheinyl transferase, *Sfp*, to enable production of functional 6-deoxyerythronolide B synthase modules (Pfeifer *et al.*, 2001).

In principle, producing recombinant PKS and NRPS proteins only involves a few, seemingly simple, steps; obtain the gene of interest, clone into a target vector under the control of an inducible promoter, express the target protein in a suitable host followed by protein purification (which can be aided by introduction of an affinity tag when obtaining the target gene). As expected there are a number of viable options at each stage of recombinant protein expression in order to attempt to produce soluble protein, including host, vector, promoter and affinity tag and this is another advantage as multiple combinations of these things can be trialled in *E. coli* (Rosano and Ceccarelli, 2014). The target host selected for recombinant gene expression in this case was *E. coli*. This chapter reviews work carried out to clone the indanomycin NRPS genes, *idm*J and *idm*I and express soluble proteins.

The initial aim of the project was to clone the genes involved in conversion of L-proline to pyrrole-2-carboxylate (Figures 1.20 and 3.1), the starter unit for the polyether ionophore antibiotic indanomycin, into *E. coli* for recombinant gene expression.

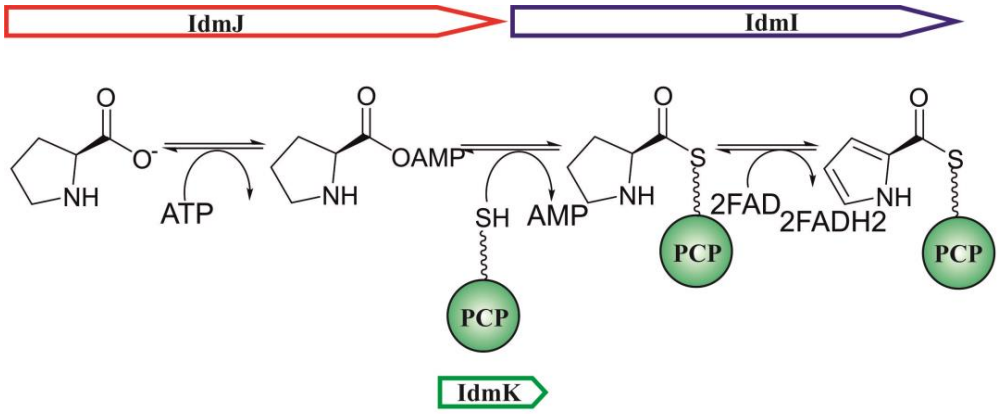


Figure 3.1- Indanomycin starter unit biosynthesis. *idm*J encodes a proline adenylyltransferase which activates L-proline and loads it onto a prolyl carrier protein encoded by *idm*K. The proline is then converted to pyrrole-2-carboxylate by a prolyl-*S*-dehydrogenase encoded by *idm*I (Li *et al.*, 2009).

Prior to this work a number of NRPS proteins have been expressed in *E. coli*, including other NRPS proteins that convert L-proline to pyrrole-2-carboxylate in aminocoumarin biosynthesis (Garneau *et al.*, 2005); however those required for indanomycin starter unit biosynthesis; IdmI, IdmJ and IdmK, had not. In order to characterise these proteins and their enzymatic mechanism for starter unit biosynthesis, and to enable protein engineering experiments, recombinant protein expression was essential.

3.1 Design of synthetic genes for recombinant protein expression in *E. coli*

To express the *S. antibioticus* enzymes responsible for indanomycin starter unit biosynthesis in *E. coli*, genes were designed and subsequently synthesised and purchased from Genscript (New Jersey, USA). Design and purchase of a synthetic gene is now an attractive and accessible option for obtaining the target gene, and preferable to obtaining the gene by PCR from the genomic DNA, since the cost of gene synthesis has significantly decreased (Mueller *et al.*, 2009). Additionally, genes can be designed to contain, or omit, certain features and characteristics such as the inclusion of an affinity tag or codon optimisation for recombinant expression. Amino acid sequences for the adenylyltransferase (*idm*J), prolyl dehydrogenase (*idm*I) and

prolyl carrier protein (*idm*K) were obtained from GenBank (accession number FJ545274) (Li *et al.*, 2009). Genes were designed so they could be cloned and expressed individually or as a synthetic operon.

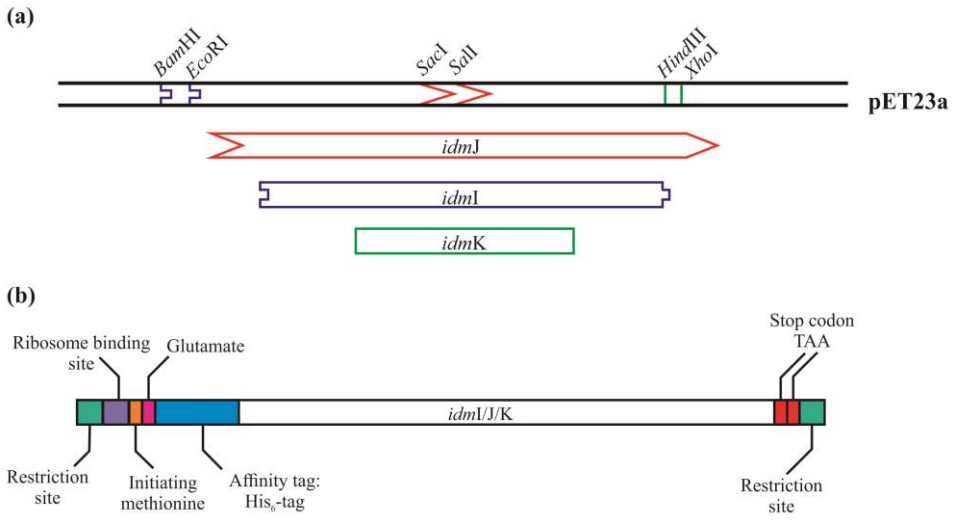


Figure 3.2- Schematic of the design for synthetic genes. (a) Synthetic *idm*J, *idm*I and *idm*K were designed with unique restriction sites at the 5' and 3' ends to be cloned into pET23a individually or altogether. *Bam*HI and *Eco*RI restriction sites were at the 5' and 3' termini of the *idm*I gene. *Sac*I and *Sal*I restriction sites were at the 5' and 3' termini of the *idmJ* gene respectively, and *Hind*III and *Xho*I restriction sites were at the 5' and 3' termini of the *idm*K gene. (b) Synthetic genes contained a ribosome binding site, initiating methionine residue followed by a glutamate, a His_6-tag at the 5' terminus of the gene and a double amber stop codon at the 3' terminus of the gene. These features are flanked by the selected restriction sites.

As illustrated in Figure 3.2, all genes were designed with unique restriction sites at their 5' and 3' ends and the selected restriction sites did not occur within any of the genes. All genes contained ribosome binding sites to permit expression in any vector, and a His_6-tag was included at the N-terminus of each protein to allow protein purification by nickel affinity chromatography. Additional features included a glutamate residue following the initiating methionine in order to prevent cleavage of the methionine therefore preventing premature protein degradation (Tobias *et al.*, 1991; Hirel *et al.*, 1989). A double amber stop codon at the 3' end of the gene was included to prevent a stop codon read-through. As these genes are usually expressed in *Streptomyces*, a bacterial strain whose genome has a high GC content (Muto and Osawa, 1987), genes were designed to be codon optimised for expression of the

protein in *E. coli*. Synthesised genes were supplied in pUC57, an appropriate cloning vector.

3.1.1 Initial cloning strategy for NRPS starter module domains

The initial strategy was to clone the genes individually into pET23a for protein expression. Figure 3.3 illustrates the standard "cut and paste" strategy used for cloning *idm*J, *idm*I and *idm*K into pET23a. Vector and insert DNA were amplified by growth in *E. coli* XL10 Gold Ultracompetent cells. Once amplified, insert and vector DNA was prepared by restriction digestion creating complimentary "sticky ends" for ligation. The resulting construct(s) were transformed into *E. coli* XL10 Gold Ultracompetent cells. Successful pET*idm*I/J/K constructs were indicated by growth on 2 × TY agar plates containing ampicillin and resulting constructs were screened with relevant restriction enzymes, and success was confirmed by DNA sequencing. The correct construct could then be transformed into an *E. coli* expression strain.

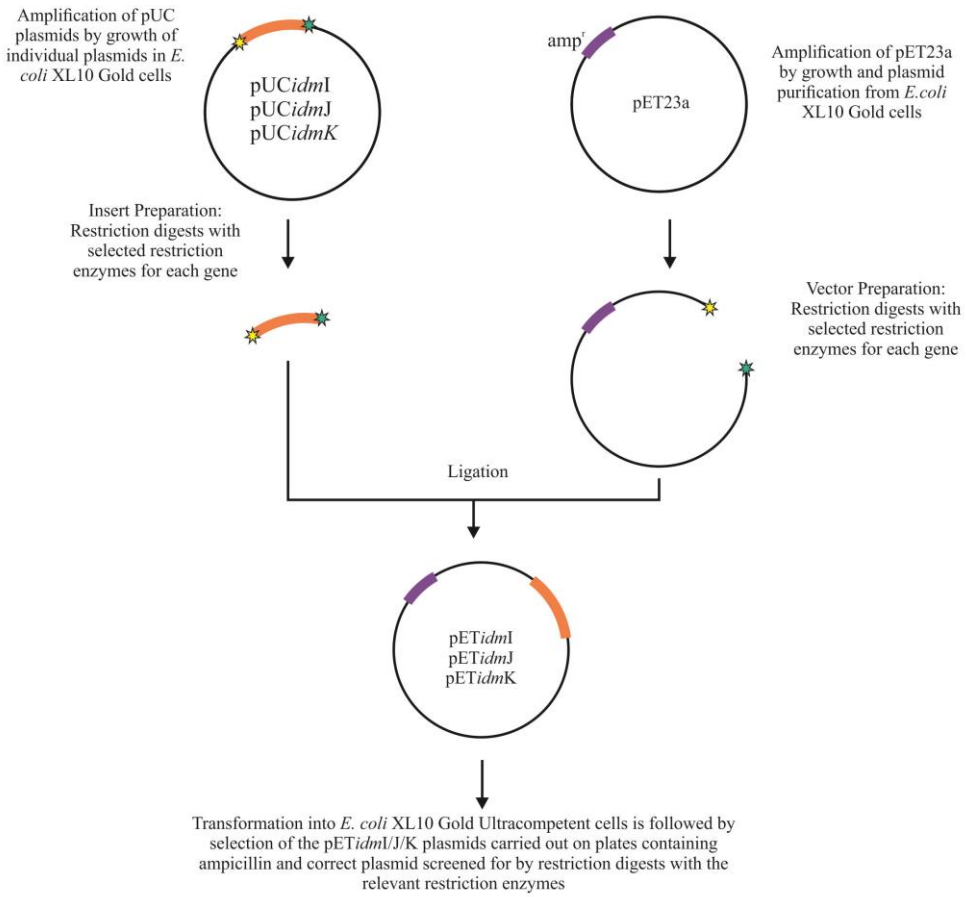


Figure 3.3-"Cut and paste" cloning strategy for cloning *idm*I, *idm*J and *idm*K. Genes were to be cloned individually into pET23a from pUC57 vector yielding pET*idm*I, pET*idm*J and pET*idm*K expression vectors. Both the synthetic gene plasmids and the pET23a vector are amplified by growth in *E. coli* XL10 Gold Ultracompetent cells. Both are digested with the appropriate restriction enzymes and a ligation performed between the insert and vector DNA. Resulting constructs are transformed into *E. coli* XL10 Gold Ultracompetent cells, cells containing the correct constructs will grow in colonies on the agar plates containing ampicillin. Plasmids from the cells can then be purified and screened by restriction digests to identify correct constructs.

3.1.2 Amplification of vector and target gene plasmids

The pET23a vector DNA was purchased from Merck4Biosciences (Nottingham, UK) (kindly gifted from Dr David Brockwell, Astbury Centre, University of Leeds, UK). The empty vector was transformed into *E. coli* XL10 Gold Ultracompetent cells (Section 2.2.2.11). A single colony was used to inoculate a 5 mL 2 × TY starter culture, 0.5 mL of this was subsequently used to make a glycerol stock of *E. coli* XL10 Gold ultracompetent cells containing the empty pET23a vector (Section 2.2.1.6).

pUC57 plasmids containing the individual genes arrived in the form of lyophilised DNA and were resuspended in sterile water as per the guidelines of the supplier of the synthetic genes. The three pUC57 plasmids containing the adenylyltransferase, dehydrogenase and carrier protein were designated pUC*idm*J, pUC*idm*I and pUC*idm*K, respectively. Plasmid DNA was then transformed into Agilent Technologies *E. coli* XL10 Gold Ultracompetent cells (Section 2.2.2.11). Glycerol stocks were made of XL10 Gold cells containing pUC*idm*I/J/K plasmids for storage and DNA production (Section 1.2.1.6). Initial restriction digests of pUC*idm*J with *Sac*I and *Sal*I, pUC*idm*I with *Bam*HI and *Eco*RI and pUC*idm*K with *Hind*III and *Xho*I were performed for each plasmid for a preliminary product check and digested DNA was run on a 0.7 % (w/v) agarose gel as described in Sections 2.2.2.2 and 2.2.2.7.

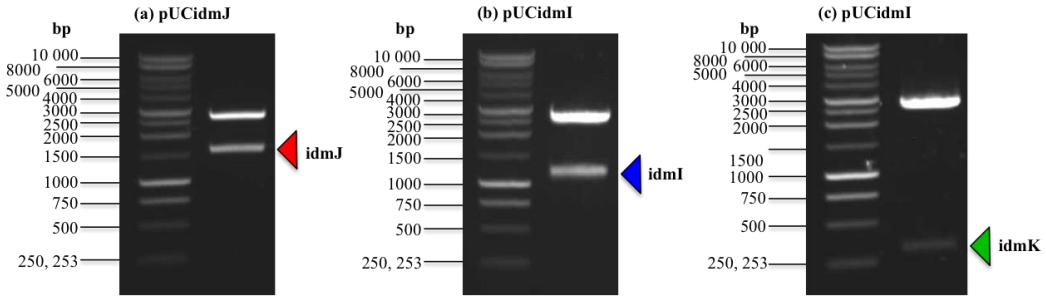


Figure 3.4-0.7 % (*w/v*) agarose gel of a restriction digest of pUC57-gene constructs (a) pUC*idm*J digested with *Sac*I and *Sal*I yielding a band at 1540 bp corresponding to the *idm*J gene (b) pUC*idm*I digested with *Bam*HI and *Eco*RI yielding a band at 1180 bp corresponding to the idmI gene and (c) pUC*idm*K digested with *Hind*III and *Xho*I yielding a band at 325 bp corresponding to the *idm*K gene. All agarose gels contain a band at 2790 bp corresponding to the pUC57 vector. All gels indicate the presence of the correct plasmids after transformation into XL10 Gold Ultracompetent cells.

Once the preliminary digest was analysed the genes were sub-cloned for protein expression as described below.

3.2 The adenylyltransferase (IdmJ)

Conversion of L-proline to pyrrole-2-carboxylate requires the activation of the amino acid for loading onto the prolyl carrier protein (IdmK). The adenylyltransferase (IdmJ) is responsible for selection of the correct substrate, in this case L-proline, and

its activation. The adenylyltransferase (IdmJ) is required to load the carrier protein with L-proline (Li *et al.*, 2009), or any other substrate in an enzyme dependent manner, for future engineering studies.

3.2.1 Cloning the adenylyltransferase (*idm*J) into pET23a

Originally individual genes were to be sub-cloned into pET23a. pET vectors contain T7 promoters for induction by IPTG (Studier *et al.*, 1990) and a range of restriction sites for cloning. To obtain the target gene *idm*J (the insert) the glycerol stock of *E. coli* XL10 Gold Ultracompetent cells containing the pUCi*dm*J plasmid was used to inoculate 5 mL starter culture in 2 × TY medium supplemented with ampicillin. This in turn was used to inoculate a 50 mL culture of 2 × TY medium (Section 2.2.1.5). Plasmid DNA was then purified using a QIAGEN Plasmid Midi kit (Section 2.2.2.1). Vector DNA was produced in a similar manner. The pUC*idm*J and pET23a vector were then digested with the restriction enzymes *Sac*I-HF® and *Sal*I-HF®; high fidelity (HF™) enzymes were used due to reduced star activity (Kamps-Hughes *et al.*, 2013). Restriction digests were performed as described in section 2.2.2.7 and run on a 0.7 % (w/v) agarose gel. Insert and vector DNA bands were then cut out of the agarose gel and a gel extraction performed (section 2.2.2.3). A ligation reaction was performed between pET23a and *idm*J (2.2.2.8), 4 μL of the ligation reaction was subsequently transformed into *E. coli* XL10 Gold Ultracompetent cells (2.2.2.11).

Successful plasmid ligation was screened for by antibiotic selection: bacterial colonies that grew on the 2 × TY agar plates containing ampicillin indicated the presence of a pET*idm*J construct. To confirm the presence of the pET*idm*J construct 10 colonies were picked and used to inoculate 10 individual 5 mL starter cultures. Plasmid DNA from selected colonies was purified using the Promega Wizard® SV Minipreps Purification kit (section 2.2.2.1) and restriction digests using *Sac*I-HF® and *Sal*I-HF® were performed to screen for the correct plasmid. Figure 3.5 shows digested plasmid DNA from 10 colonies screened for the pET*idm*J plasmid. The expected sizes of the DNA bands for a restriction digest using *Sac*I-HF® and *Sal*I-HF® of the pET*idm*J construct are approximately 3.7 kb (3666 bp) and 1.5 kb (1540 bp) corresponding to the vector and insert respectively.

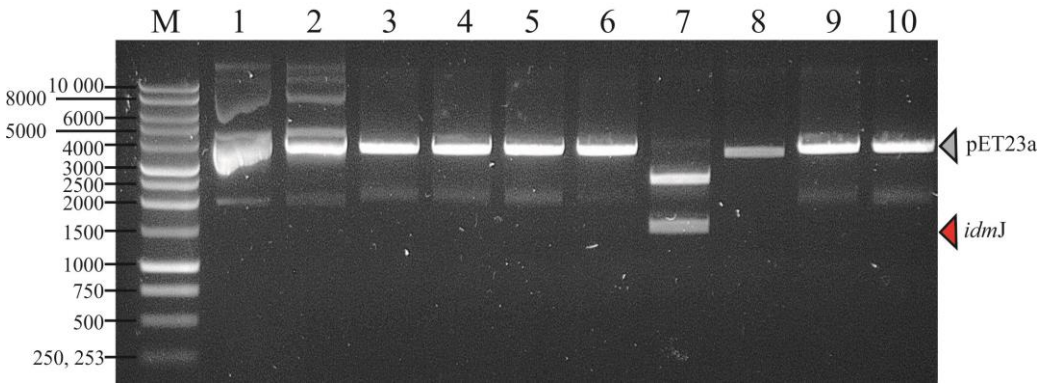


Figure 3.5- 0.7 % (*w/v*) agarose gel of the isolated plasmid DNA from the 10 colonies after digestion with *Sac*I-HF® and *Sal*I-HF® restriction enzymes. The grey arrow indicates the expected size of the vector DNA and the red arrow indicates the expected size of *idm*J. Lanes 1 and 2 appear to be undigested plasmid DNA, plasmids in lanes 3-6, 9 and 10 appear to have yielded two DNA fragments, one at approximately 3700 bp corresponding to the pET23a vector and the second approximately 2000 bp in size, this is too large to be *idm*J. The plasmid in lane 7 contains a DNA fragment at 1500 bp corresponding to *idm*J the vector is 2700 bp in size, corresponding to pUC57. Finally the digested plasmid in lane 8 contains only one DNA fragment at the approximate size for pET23a.

Figure 3.5 indicates that the ligation reaction between pET23a and idmJ was unsuccessful. Plasmids 1 and 2 do not appear to be digested by *Sac*I-HF® and *Sal*I-HF® (lanes 1 and 2), Plasmids 3, 4, 5, 6, 9 and 10 are digested by *Sac*I-HF® and *Sal*I-HF® yielding a DNA fragment corresponding to pET23a and an unidentified insert too large to be *idm*J (lanes 3-6, 9 and 10). Plasmid 7 (lane 7), when digested by *Sac*I-HF® and *Sal*I-HF® produced two DNA fragments, one corresponding to *idm*J and the other pUC57 indicating pUC*idm*J contamination during insert preparation. Finally, digestion of plasmid 8 (lane 8) yielded one DNA fragment corresponding to empty pET23a; this indicates incomplete digestion of pET23a in vector preparations.

Several attempts at cloning *idm*J into pET23a were made. The restriction enzyme stocks were replaced with new enzymes, as well as new vector DNA purchased from Merck4Biosciences (Nottingham, UK), however this strategy was not successful. Inefficient digestion during vector and insert preparations using *Sac*I-HF® and *Sal*I-HF® resulted in plasmids that are incompletely digested which are able to simply "snap shut" in the presence of T4 DNA ligase and will transform efficiently,

contaminating the transformation plates with the incorrect plasmid. Although digests were performed for extended periods of time there may have been incompatibility with the two restriction sites selected. In addition to this, sites selected may have been in too close a proximity to one another in the vector, decreasing the efficiency of activity of the restriction enzyme.

3.2.2 Cloning the adenylyltransferase (*idm*J) into pKK223-3

The next approach to sub-clone *idm*J was to select an alternative vector. The vector selected was pKK223-3 (Pharmacia). This was obtained in the form of pKnanA (Timms *et al.*, 2013) from Prof. A. Berry (Astbury Centre, University of Leeds, UK). Figure 3.6 illustrates the planned cloning strategy. The *idm*J gene was amplified by culturing cells containing the pUC*idm*J plasmid, the plasmid purified and insert cut out of the pUC57 vector using *Eco*RI-HF® and *Hind*III-HF® restriction enzymes. The pK*nan*A plasmid was propagated by culturing of *E. coli* XL10 Gold Ultracompetent cells containing the pK*nan*A plasmid, purifying the plasmid DNA, performing a restriction digest using *Eco*RI-HF® and *Hind*III-HF® restriction enzymes to cut out the *nan*A gene and create complimentary sticky ends to the *idm*J insert. The insert and vector will be ligated together and transformed into *E. coli* XL10 Gold Ultracompetent cells. A successful ligation between pKK223-3 and *idm*J will be indicated by colony growth on 2 × TY agar plates supplemented with ampicillin. Restriction digests with *Eco*RI-HF® and *Hind*III-HF® will be used to screen resulting plasmids from the ligation. Confirmation of the ligation between *idm*J and pKK223-3 can then be confirmed by sequencing and the resulting construct designated pK*idm*J. This strategy exploits the removal of a gene from the target vector to confirm the double digest of the vector has been successful.

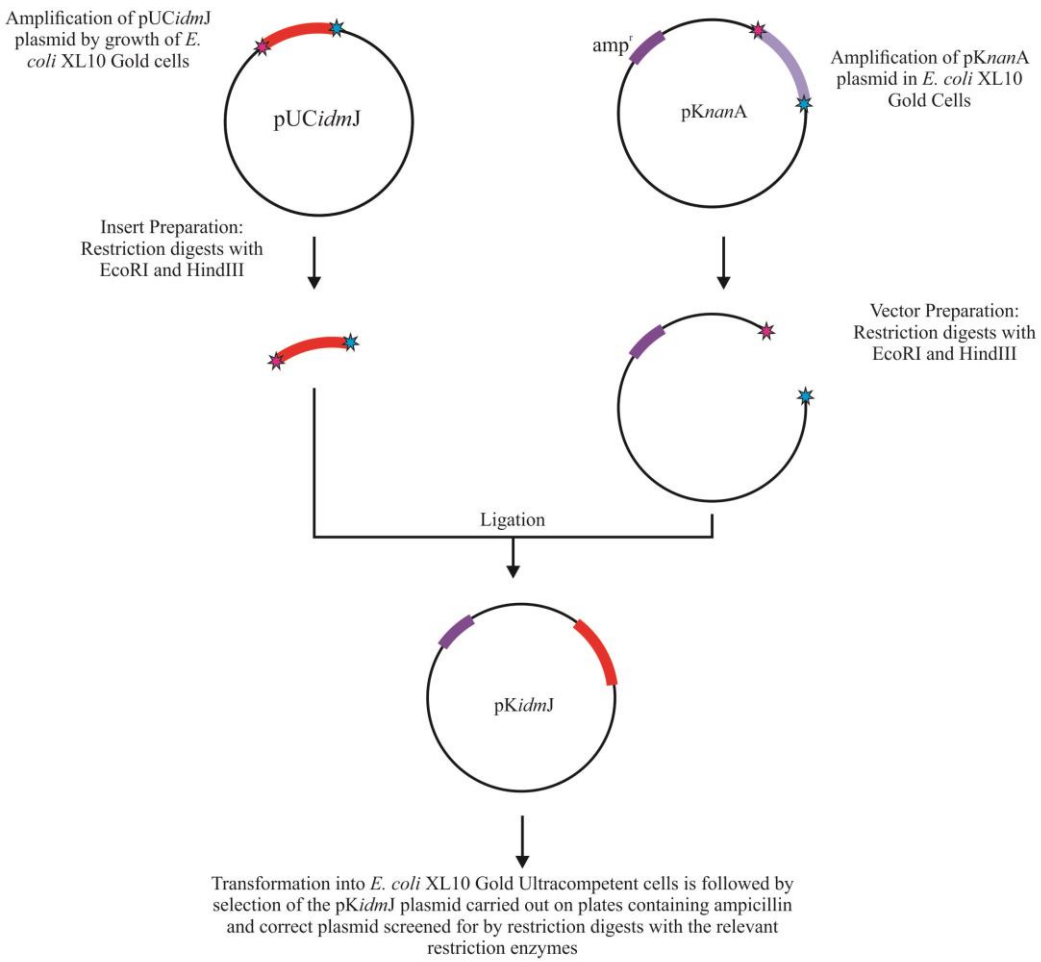


Figure 3.6- Schematic for the method used to clone *idm*J into pKK223-3. *idm*J will be cut out of the pUC57 plasmid using *Eco*RI and *Hind*III sites that reside in the pUC57 multiple cloning site. The pK*nan*A plasmid will be digested with the same restriction enzymes, removing the *nan*A gene creating complimentary sticky ends to the insert. A ligation reaction can then be performed between the two DNA fragments. Resulting plasmids from the ligation can be transformed into *E. coli* XL10 Gold cells and the pK*idm*J plasmid screened for.

Insert and vector DNA were prepared as previously described, using the restriction enzymes *Eco*RI-HF® and *Hind*III-HF® (3.2.1). Figure 3.7 shows the *nan*A gene has been successfully cut out of the pKK223-3 plasmid using the *Eco*RI and *Hind*III sites.

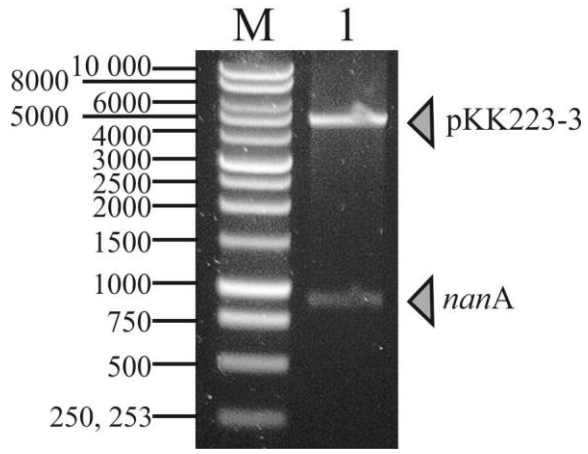


Figure 3.7-0.7 % (*w/v*) agarose gel of restriction digest of the pK*nan*A construct using *Eco*RI-HF® and *Hind*III-HF®.-The *nan*A gene (~900 bp) was successfully cut out of the pKnanA plasmid (~4.6 kb). The vector DNA was then purified from the agarose gel using a Qiagen gel extraction kit (2.2.2.1).

The ligation reaction and subsequent transformation were also performed as before (section 3.2.1). 10 colonies were screened for the correct plasmid by restriction digestion with *Eco*RI-HF® and *Hind*III-HF®; results are presented in Figure 3.8. The pK*idm*J construct digested with *Eco*RI-HF® and *Hind*III-HF® should yield two DNA fragments, one at 4600 bp and the other at 1500 bp.

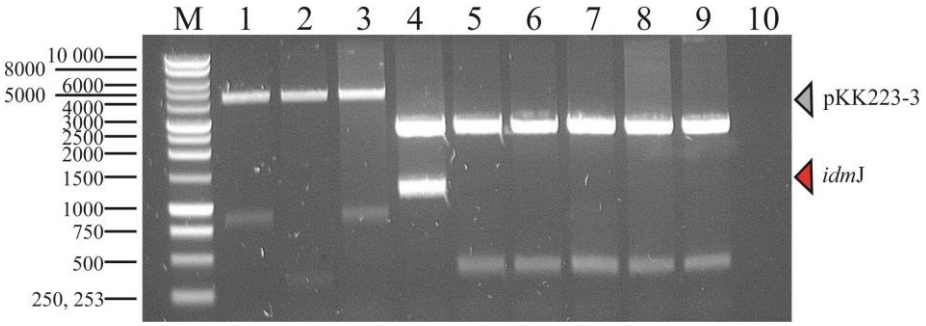


Figure 3.8-0.7 % (*w/v*) agarose gel of restriction digests of the 10 isolated plasmids from a ligation reaction between pKK223-3 and *idm*J. Lanes 1 and 3 appear to contain two bands at 4600 and 900 bp corresponding to pKK223-3 and the *nan*A gene respectively. Lane 2 shows only one band at the correct size for the vector however no insert can be observed. Lane 4 shows DNA fragments corresponding to pUC57 (2.7 kb) and *idm*J. Lanes 5-9 shows DNA fragments corresponding to the pUC57 plasmid and an unknown insert of approximately 500 bp. Plasmid purification for colony 10 (lane 10) appears to have been unsuccessful.

The agarose gel in Figure 3.8 indicates that the ligation between pKK223-3 and *idm*J was unsuccessful. Restriction digests with *Eco*RI-HF® and *Hind*III-HF® of plasmids 1 and 3 (lanes 1 and 3) yielded two DNA fragments, one at 4600 bp and a second at 900 bp corresponding to pKK223-3 and the *nan*A gene respectively. Lane 2 shows a DNA band at 4600 bp corresponding to linearised pKK223-3 with no insert present. Restriction digestion of plasmid 4 (lane 4) yielded two bands, one at 2700 bp and a second at 1500 bp corresponding to pUC57 and *idm*J. Lanes 5-9 show two bands, one at 2700 bp corresponding to the expected size of pUC57 and an unknown DNA fragment approximately 500 bp in length. Finally, there is no DNA present in lane 10 suggesting plasmid purification was unsuccessful. Again, contamination from insert and vector preparations were observed (lanes 1-4) in addition to contamination by an unknown insert (lanes 5-9). This approach did not yield any positive results and was repeated without any success.

3.2.3 Cloning the adenylyltransferase (*idm*J) into pETDUET

After the unsuccessful attempts at cloning *idm*J into pET23a and pKK223-3 a third vector was trialled for sub-cloning, pETDUET. pETDUET is a vector that contains two multiple cloning sites. *idm*J was to be cloned into the first multiple cloning site using the *Sac*I and *Sal*I restriction sites. The *Sac*I and *Sal*I sites are separated by a few more base pairs in pETDUET vector than pET23a. The empty pETDUET vector was transformed into *E. coli* XL10 Gold Ultracompetent cells and a glycerol stock made (sections 2.2.2.11 and 2.2.1.6). Vector and insert DNA were prepared as described in section 3.2.1, using *Sac*I-HF® and *Sal*I-HF® restriction enzymes. A ligation reaction was performed between pETDUET and *idm*J, to create a plasmid designated pET(2)*idm*J, followed by a transformation. Six colonies were picked from the transformation plate and used to inoculate six 5 mL cultures of 2 × TY medium. Plasmid DNA was purified from 4.5 mL of the culture and the pET(2)*idm*J plasmid screened for by restriction digests with *Sac*I-HF® and *Sal*I-HF®. Figure 3.9 shows the results from the restriction digests.

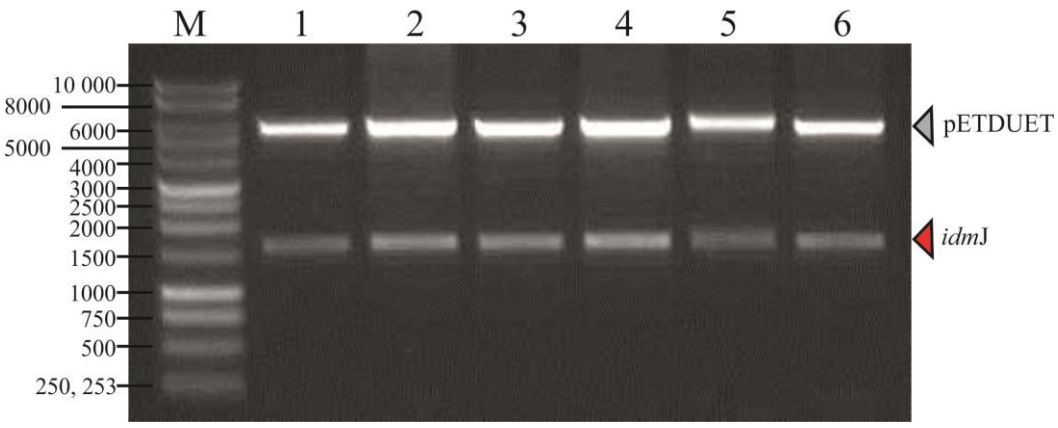


Figure 3.9-0.7 % agarose gel showing restriction digests with SacI-HF® and SalI-HF® of plasmids isolated from a ligation between pETDUET and *idm*J. Lanes 1-6 all show two DNA fragments after digestions with *Sac*I-HF® and *Sal*I-HF®, one band at 5500 bp and a second at 1550 bp corresponding to the expected sizes of pETDUET (5420 bp) and *idm*J (1540 bp).

Initial screening from the ligation reaction indicates the presence of the pET(2)*idm*J plasmid. DNA fragments in lanes 1-6 all showed a band at 5500 bp and another at 1540 bp, corresponding to the correct sizes for vector and insert, respectively. This suggested that all plasmids screened were pET(2)*idm*J. Plasmid 1 was selected for sequencing. A glycerol stock was made of *E. coli* XL10 Gold Ultracompetent cells containing plasmid 1. Purified plasmid DNA from this stock was subjected to DNA sequencing using the primers "pET UPSTREAM" and "DuetDOWN1" (see Appendix) (section 2.2.2.12). The sequencing revealed that pETDUET contained the correct sequence for the *idm*J gene.

The purified plasmid DNA of the pET(2)*idm*J plasmid was then transformed into *E.coli* BL21 (DE3) Gold cells, a single colony was picked and used to inoculate 5 mL of 2 × TY medium. 0.5 mL of this culture was used to make a glycerol stock, which was used for subsequent protein expression trials.

3.2.4 Recombinant protein expression of the adenylyltransferase (IdmJ)

Protein expression of IdmJ was attempted in *E.coli* BL21 (DE3) Gold cells containing the pET(2)*idm*J plasmid. Initial small scale protein expression was attempted in 2 × TY medium as described in section 2.2.3.1. To determine if IdmJ had been expressed as a soluble protein, cells from 1.5 mL of the culture were

harvested, lysed and soluble and insoluble fractions separated for analysis by SDS-PAGE (sections 2.2.3.3 and 2.2.3.6).

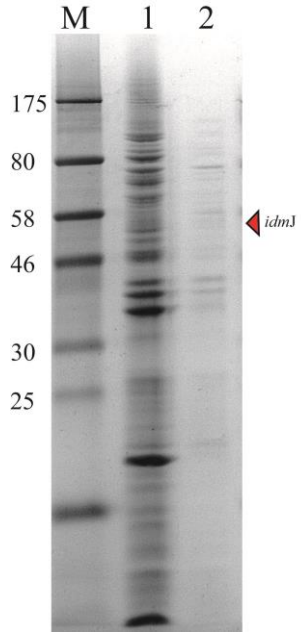


Figure 3.10- Reducing SDS-PAGE gel showing initial overexpression of IdmJ. Lane 1 is the insoluble fraction of induced *E. coli* BL21 (DE3) Gold cells containing the pET(2)*idm*J plasmid. Lane 2 shows the soluble fraction. The red arrow indicates where IdmJ would be expected to appear on an SDS-PAGE gel.

The expected mass of IdmJ, as calculated from the amino acid sequence of the synthetic gene, is 53 522 Da. From the SDS-PAGE gel in Figure 3.10 it was unclear as to whether IdmJ was expressed as a soluble or insoluble protein, if expressed at all.

There are a number of properties in protein expression that can be adjusted in order to produce soluble protein; one aspect is changing the growth medium. Changing the growth medium may help with solubility as different media permit differential cell growth patterns, for example LB medium, promotes cell growth at an early log phase, however prevents high cell densities (Sezonov *et al.*, 2007; Rosano and Ceccarelli, 2014). Other media, such as TB, encourage high cell densities which could mean increased expression; however more cells means less oxygen, which will prompt expression of a large number of proteins creating a metabolic burden, perhaps inhibiting soluble protein expression (Rosano and Ceccarelli, 2014; Unden *et al.*, 1995). For this reason a range of media of varying degrees of richness were trialled. In addition to 2 × TY (the current growth medium), protein expression was

tested in LB, TB and autoinduction media (section 2.1.7). Autoinduction medium supplies glucose as a nutrition source for *E. coli*, protein expression is suppressed in the presence of glucose. When glucose is no longer present protein expression is initiated. Another easily adaptable variable is the temperature at which the induced cells are grown. The temperature can be reduced in an attempt to improve protein solubility. Therefore induced cells will be grown at 37°C, 21°C (room temperature) or 18°C. In order to reduce the rate of protein expression, allowing more time for accurate protein folding, the concentration of IPTG was reduced, giving a final concentration of 60 μM (Tolia and Joshua-Tor, 2006).

Expression trials were carried out on a 5 mL scale. The glycerol stock of *E. coli* BL21 (DE3) Gold cells containing the pET(2)*idm*J plasmid was used to inoculate a 5 mL overnight starter culture, which in turn was used to inoculate 5 mL day cultures of TB, 2 × TY and LB media which were incubated at 37°C. Once the OD_{600} of cell cultures reached between 0.6-1.0 the cells were induced with IPTG to a final concentration of 60 μM. Upon induction the temperature of incubation was either kept at 37°C or reduced to 21°C or 18°C. This method for cell growth and induction was utilised for all media trialled, excluding autoinduction medium. In autoinduction medium, cells were incubated at the same temperature throughout the experiment: due to the slow cell growth at lower temperatures (21 and 18°C) these cultures were incubated for approximately 28 hours in total. Each medium was trialled at all three temperatures. Samples were taken from the soluble cell fraction and analysed by SDS-PAGE (Figure 3.11).

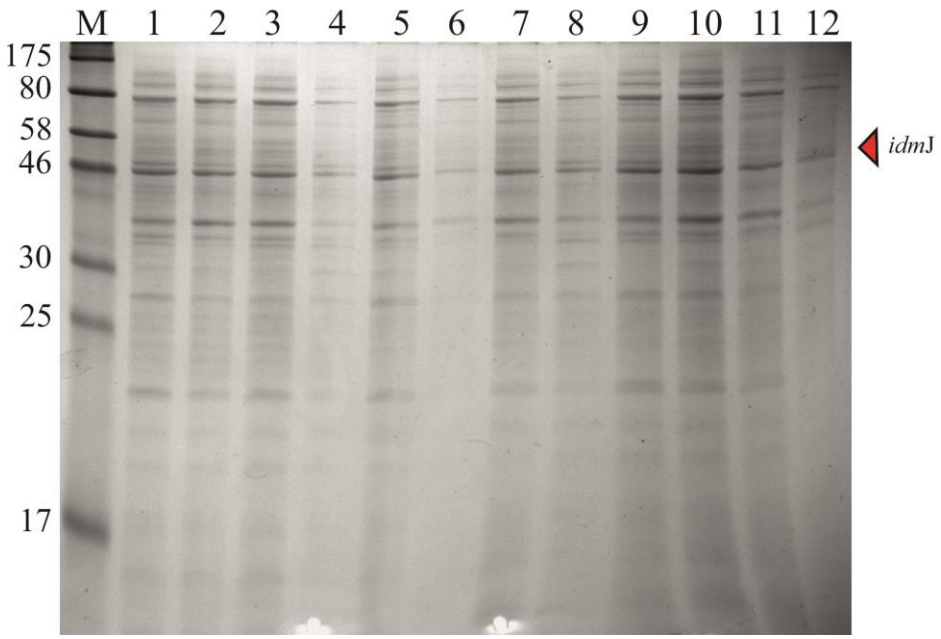


Figure 3.11- pET(2)*idm*J expression trials. Reducing SDS-PAGE gel of the soluble cell fractions from the IdmJ expression trial. Expression of IdmJ was carried out by growing *E. coli* BL21 (DE3) Gold cells containing the pET(2)*idm*J plasmid in four different media at three different temperatures. Lanes 1-3 shows the results of protein expression in TB medium at 37 °C, 21 °C (room temperature) and 18 °C respectively. Lanes 4-6 show the results of expressing protein in 2 × TY at 37 °C, 21 ° C and 18 °C respectively. Lanes 7-9 show the results of expressing IdmJ in LB medium at 37 °C, 21 °C and 18 °C respectively. Finally lanes 10-12 show the results of expressing IdmJ in autoinduction medium at 37 °C, room 21 °C and 18 °C respectively. IdmJ is approximately 53 kDa, the red arrows indicate where a band is expected if soluble protein is expressed.

Analysis of the SDS-PAGE gel of the soluble cell fractions in Figure 3.11 and comparison with untransformed *E. coli* BL21 (DE3) Gold cells (see appendix) suggests that IdmJ was not being expressed as a soluble protein due to a lack of a protein band at the expected mass (indicated by the red arrow). To check for low levels of soluble protein expression an anti His_6-tag antibody was used to detect His_6-tagged protein as a more sensitive method of detection. This was carried out by blotting 2 µL of the soluble fraction or 2 µL of the insoluble fraction resuspended in PBS on a nitrocellulose membrane and the His_6-tag detected as described in section 2.2.3.2. Figure 3.12 below shows the results of His_6-tag detection in samples taken from the pET(2)*idm*J expression trial.

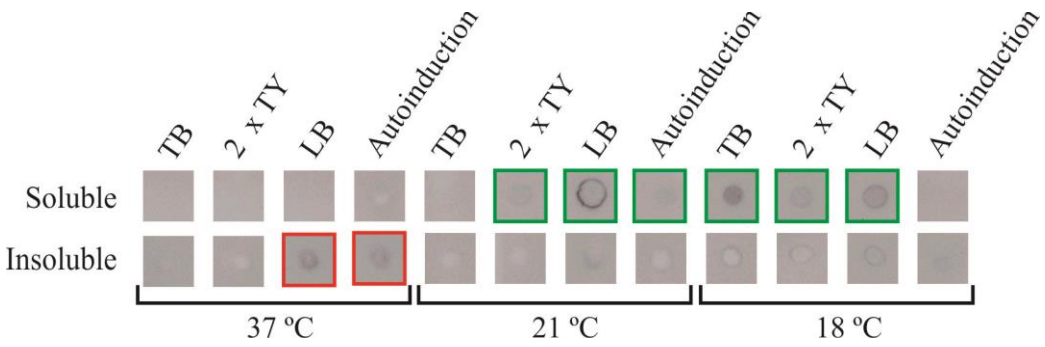


Figure 3.12- N-terminal His_6-tag detection in IdmJ expression trials. Soluble (top row) and insoluble (bottom row) cellular fractions were blotted onto a nitrocellulose membrane and an anti His_6-tag antibody used for detection of a His_6-tag to probe for protein expression. Blots are of *E. coli* BL21 (DE3) Gold cells grown in four different media and expression induced at three different temperatures. At 37 °C His_6-tagged protein was detected in the insoluble fractions of cells grown in LB or autoinduction medium. There does not appear to be any soluble His_6-tagged protein expressed. At 21 °C His_6-tagged protein was detected in the soluble fractions of cells grown in 2 × TY and LB medium. Much lower levels of His_6-tagged protein were detected in the insoluble fraction of cells grown in LB and autoinduction medium. At 18 °C although His_6-tagged protein was detected in the insoluble fractions of cells grown in TB, 2 × TY and LB medium increased levels were detected in the soluble cell fraction. No His_6-tagged protein was detected in cells grown in autoinduction medium at 18 °C. This is a negative display of the image for clarity.

Detection of the His_6-tag utilising an anti-polyHistidine tag antibody identified a number of expression conditions which produced soluble protein. Figure 3.12 shows that His_6-tagged protein was detected in the insoluble fractions of cells grown at 37°C when cells were cultured in LB and autoinduction medium (outlined in red), and no protein expression detected in TB or 2 × TY medium. When the temperature of the cultures were reduced to 21 °C after induction, soluble protein was detected in LB, 2 × TY and autoinduction medium (green). Low levels of His_6-tagged protein were also detected in the insoluble fractions of cells. When cells were incubated at 18°C after induction soluble protein was detected in TB, 2 × TY and LB medium, lower levels of His_6-tagged protein were detected in the insoluble fraction. At lower incubation temperatures, cell growth was slow in autoinduction medium with low levels of protein expression; this therefore suggested that this medium was unsuitable for the expression of IdmJ. The results show that expression of IdmJ is somewhat temperature dependent, with soluble protein being expressed at lower temperature. Although the gene was codon optimised for expression in *E. coli*, fermentation of *S. antibioticus*, the native host, is at 28°C for 72 hours (Li *et al.*, 2009). The most consistent medium for expression of soluble IdmJ was LB medium,

with soluble protein being detected at 21°C and 18°C. The next stage was to attempt protein expression on a larger, 1 L scale, followed by subsequent protein purification.

3.2.5 Purification of IdmJ and analysis

After establishing an expression protocol, the next step was to purify IdmJ for analysis by SDS-PAGE and mass spectrometry. When the genes were synthesised they were designed to contain an N-terminal His_6-tag to enable purification by nickel affinity chromatography. Briefly, for large-scale expression, a glycerol stock of *E.coli* BL21 (DE3) cells containing the pET(2)*idm*J plasmid was used to inoculate a 5 mL starter culture of 2 × TY medium. The starter culture was then used to inoculate 500 mL of LB medium in a 2 L flask. The cell OD was monitored at 600 nm, when cell growth reached an OD of 0.6 the temperature was reduced to 18°C and protein expression induced by the addition of IPTG at a final concentration of 60 μM. The cultures were then incubated for a further 16 hr at 18°C. IdmJ was subsequently purified using batch nickel affinity chromatography as described in section 2.2.3.4. Samples were taken at each stage of the purification and subjected to analysis by SDS-PAGE.

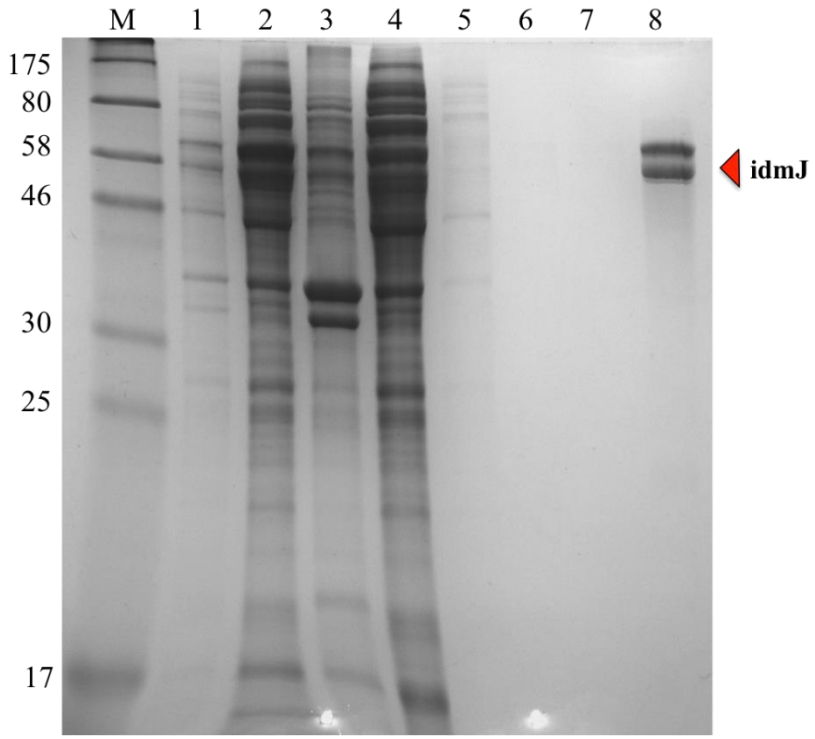


Figure 3.13- Reducing SDS-PAGE gel showing stages of the nickel affinity purification of IdmJ. Broad range pre-stained protein markers are in lane (M). Lane 1 shows the whole cell lysate of induced BL21 (DE3) Gold cells containing the pET*idm*J construct expressed in LB media at 18 °C, cells were harvested by centrifugation and lysed, the soluble (lane 2) and insoluble fractions (lane 3) were then separated by centrifugation. The soluble fraction was then incubated with resin chelated with nickel chloride for 1 hour then the post-load supernatant (lane 4) was discarded. The resin was washed with a low imidazole washing buffer (5, 6 and 7) to remove non-specifically bound proteins. Lane 8 shows purified protein after elution with a high imidazole buffer. There appears to be two protein species eluted from the resin, these species are present in the whole cell lysate (1) soluble fraction (2) and in post-load (4), there are lower levels in the insoluble fraction (3).

Analysis of nickel affinity purification of IdmJ by SDS-PAGE (Figure 3.13) revealed two bands in lane 8, suggesting two His_6-tagged protein species were purified. The approximate molecular masses of these protein species appear to be approximately 58 and 53 kDa. Approximately 15 mg of protein was purified from **1 L** of culture. For a more in-depth and accurate analysis of the molecular mass, a 70 µL sample of eluted protein was desalted into 50 mM ammonium acetate pH 7.4 (section 2.2.3.9) and analysed by ESI-MS (section 2.2.3.12). Figure 3.14 shows the mass spectrum of the eluted proteins.

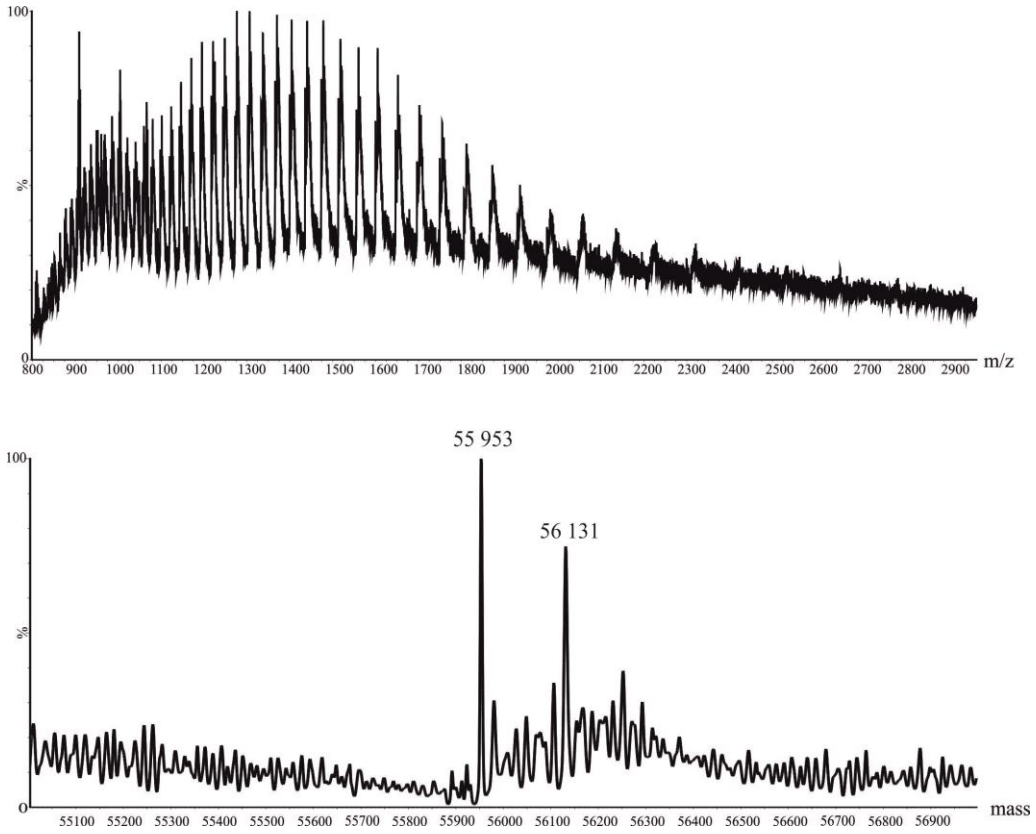


Figure 3.14- ESI-MS of eluted protein from the purification of IdmJ. (a) mass spectrum raw data (b) deconvoluted mass spectrum of raw data identifying two species with masses of 55 953 Da and 56 131 Da. The expected mass for IdmJ is 53 522 Da.

The mass calculated from the amino acid sequence encoded by the synthetic gene for IdmJ is 53 522 Da. Although SDS-PAGE analysis (Figure 3.13) showed two protein species eluted from the resin with approximate masses of 58 and 53 kDa, the mass spectrum above shows the mass of two protein species at 55 953 Da and 56 131 Da. The smaller of these masses (55 953 Da) shows a difference from the expected mass of IdmJ of 2 431 Da and is clearly outside the error margin for mass spectrometry, this protein will be denoted as IdmJ†. There is a second peak at 56 131 Da, 178 Da larger than that of IdmJ†. After considering known post-translational modifications and their masses a suggested source of this additional mass could be due to a spontaneous post-translational modification of the His_6-tag, alpha-N-6-phosphogluconylation (Geoghegan *et al.*, 1999), although no further evidence for this modification was obtained.

3.2.6 Characterisation of IdmJ†

In order to identify the root of the additional mass the first aspect examined was the DNA sequence. Plasmid DNA was purified from the *E.coli* BL21 (DE3) Gold cell glycerol stock used for expression of IdmJ and transformed back into *E. coli* XL10 Gold Ultracompetent cells for DNA sequencing using the sequencing primers "pET UPSTREAM" and "DuetDOWN1" (see Appendix) (section 2.2.2.12). Sequencing results showed no change in the original plasmid, the gene or vector, indicating that IdmJ should have the expected amino acid sequence and thus the correct mass.

The next step was to eliminate the possibility that additions or changes to the amino acid sequence had occurred, to do this LC-MS/MS was utilised to evaluate peptide fragments generated from an in-gel tryptic digest for protein identification (section 2.2.3.10) this was carried out by Dr James Ault (Astbury Centre, University of Leeds, UK). The schematic in Figure 3.15 below illustrates the in-gel tryptic digest method.

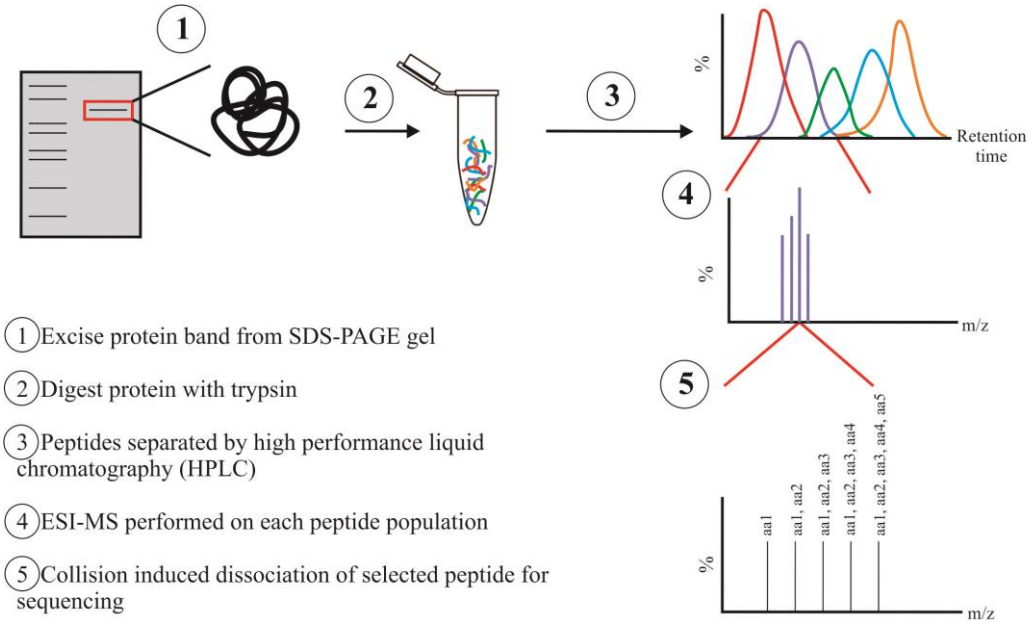


Figure 3.15- Schematic of method for an in-gel tryptic digest followed by LC-MS/MS for protein identification. A sample of eluted protein from the IdmJ purification was run on an SDS-PAGE gel, after staining the gel, protein bands were excised. The protein was reduced and alkylated followed by a digest using trypsin. Peptides were separated by high performance liquid chromatography (HPLC). Identical pools of peptides were subject to ESI-MS. Peptides were then sequenced by collision-activated dissociation followed by MS/MS. Peptide fragments were identified using a database edited to contain the synthetic IdmJ amino acid sequence (Henkin *et al.*, 2004; Fujii, 2009).

The in-gel tryptic digest followed by LC-MS/MS was carried out for both eluted species seen in lane 8 of the SDS-PAGE gel in figure 3.13. Both bands were separately excised from the SDS-PAGE gel, reduced and alkylated and digested with trypsin. Subsequently peptides were separated by high performance liquid chromatography (HPLC) (2.2.3.10). Trypsin cleaves on the C-terminus of either arginine or lysine residues, therefore the fragments produced from a known amino acid sequence can be predicted. Figure 3.16 shows the HPLC traces of both proteins that were purified in the adenylyltransferase purification (Figure 3.13, lane 8) digested with trypsin.

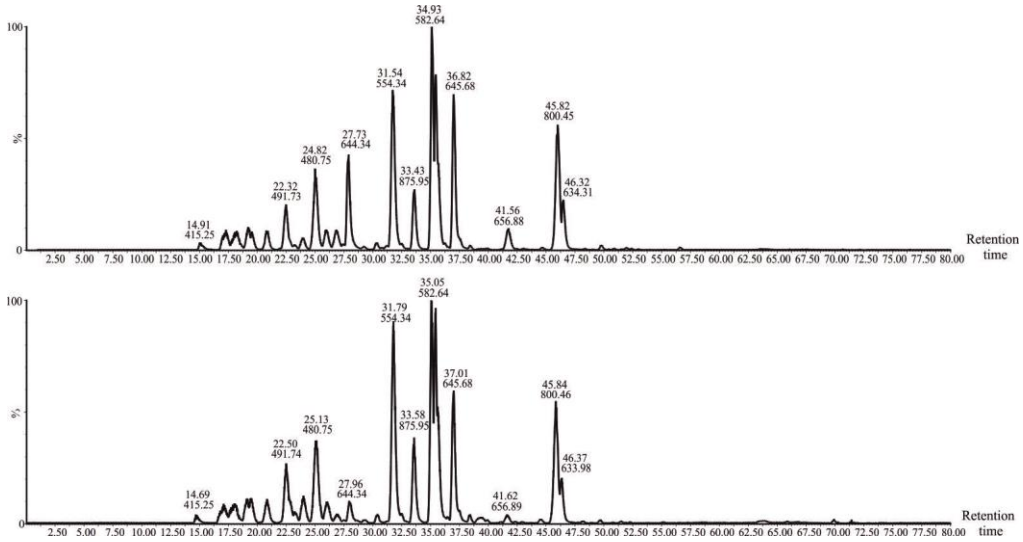


Figure 3.16- HPLC trace of purified IdmJ† digested with trypsin. A comparison of peptides present of bands excised from lane 8 of the SDS-PAGE gel in figure 3.13. The top trace shows peptide fragments present in the protein species in the top band on the SDS-PAGE gel. The bottom trace shows the peptide fragments present in the protein species in the bottom band of the SDS-PAGE.

Figure 3.16 shows the HPLC traces of peptide fragments from the two protein species eluted during purification of IdmJ. Both the top band of approximately 58 kDa (top trace) and bottom band of approximately 53 kDa (bottom trace) in Figure 3.13 display identical HPLC traces in Figure 3.16. As mentioned previously trypsin will cleave the amino acid sequence on the C-terminal side of arginine and lysine residues. This will create a pool of peptides, which can be predicted from the amino acid sequence alone. Identical HPLC traces therefore indicate that the protein species in the top and bottom bands of the SDS-PAGE gel are composed of the same peptides suggesting they are identical proteins.

Collision induced dissociation (CID) of peptides caused further fragmentation of the peptides. This enables sequencing of individual populations of peptides. The peptides that were identified from the tryptic digest and sequenced accounts for approximately 75 % of the amino acid sequence of IdmJ; sequence coverage is illustrated in Figure 3.17.

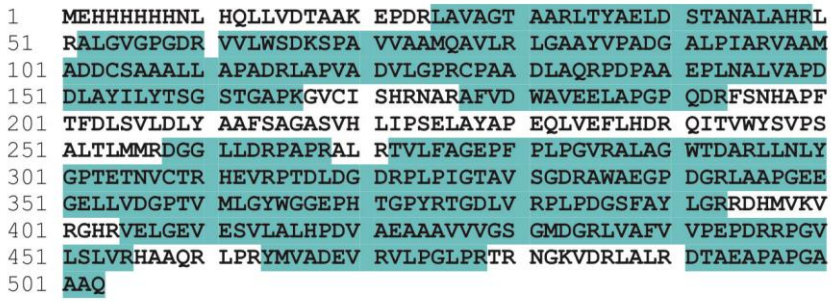


Figure 3.17- Summary of the peptide coverage of the IdmJ synthetic gene amino acid sequence from the tryptic digest. An in-gel tryptic digest generated pre-defined peptide fragments, LC-MS/MS was then used to identify the peptide fragments using databases of known protein sequences. The amino acid sequence of the synthetic IdmJ gene was added to the database. Approximately 75 % of the peptides were identified within the IdmJ amino acid sequence.

Purification by nickel affinity chromatography followed by an in-gel tryptic digest and LC-MS/MS identified peptide fragments that were expected from a tryptic digest of IdmJ. This suggests that both bands in lane 8 of Figure 3.13 were modifications of IdmJ. Identification of the C-terminal peptide fragment indicated the additional mass does not come from a stop codon read-through and purification by the His_6-tag indicated that the N-terminus of IdmJ is correct. Additionally, peptide fragmentation can also highlight any amino acids that have been post-translationally modified. During sample preparation, prior to the tryptic digest, the cysteine residues were alkylated to prevent disulphide bond formation as it may reduce peptide yields and hinder identification. Aside from cysteine alkylation, and other modifications such as oxidation that are caused by the method, the identity of the remaining post-translational modification could not be determined.

DNA sequencing and the in-gel tryptic digest followed by analysis using LC-MS/MS has indicated that the source of the additional mass observed for IdmJ† is not due to changes in the DNA or amino acid sequences. Peptide sequencing by MS/MS was also unable to identify the post-translational modification. Reduction and alkylation

of cysteine residues prior to the tryptic digest could remove a covalent post-translational modification at one or more cysteine residues, therefore removing the source of the additional mass.

To investigate if the cysteine residues were playing a role in acquiring the additional 2 431 Da observed for IdmJ†, the cysteine residues were alkylated and the protein mass obtained by analysis using ESI-MS. 25 µM of IdmJ†, purified by nickel affinity chromatography, was unfolded by 8 M urea. The protein was reduced by the addition of DTT and the cysteine residues alkylated by the addition of iodoacetamide (section 2.2.3.11). The reaction was incubated in the dark for 30 min, and then quenched by buffer exchange into 50 mM ammonium acetate pH 7.4 for analysis by ESI-MS (section 2.2.3.12). Figure 3.18 below shows the mass spectrum of alkylated IdmJ†.

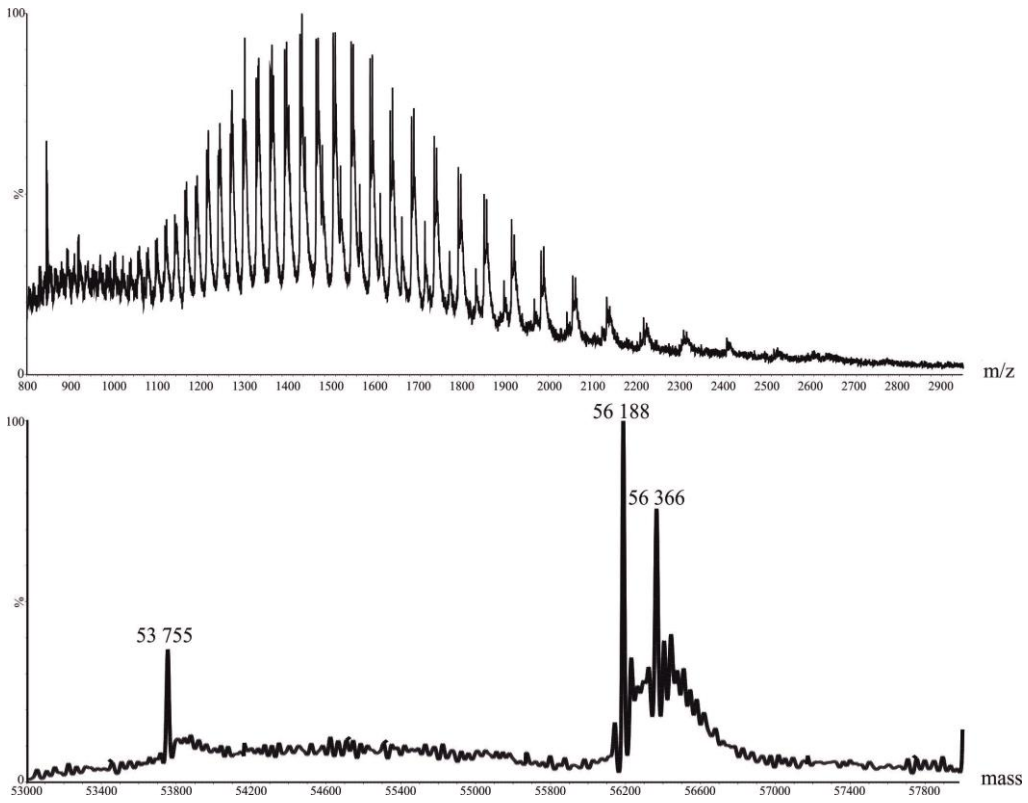


Figure 3.18- ESI-MS analysis of alkylated IdmJ. 25 µM of purified IdmJ was unfolded by the addition of urea to a final concentration of 8 M. The protein was then reduced with DTT and alkylated by the addition of iodoacetamide. (Top) mass spectrum raw data (bottom) deconvoluted mass spectrum of raw data above. 3 peaks are can be seen in the spectrum, at 53 753 Da, 56 186 and 56 368 Da.

There are four cysteine residues in the amino acid sequence of IdmJ at positions 104, 127, 169 and 308, therefore the potential for alkylation at four positions. Alkylation with iodoacetamide introduces a carbaminomethyl group onto the cysteine residue, increasing the mass of the protein by 57 Da per group added. Table 3.1 below shows and compares the expected and observed masses for the alkylation of IdmJ and IdmJ†.

Sample	Expected mass (Da)	Observed mass (Da)
IdmJ (Figure 3.14)	53 522	55 953 (IdmJ†)
Fully alkylated IdmJ	53 753	53754
Fully alkylated IdmJ†	56 181	56 186 56 365

Table 3.1- Table showing the expected masses for full alkylation of all cysteine residues in both IdmJ and IdmJ† compared to the masses observed by ESI-MS.

The mass spectrum in Figure 3.18 shows the alkylation of IdmJ† yielded three protein species; 53574 Da, 56 186 Da and 56 365 Da. Table 3.1 shows the comparison of expected masses for fully alkylated IdmJ and IdmJ†, i.e. alkylation of all four cysteine residues. Table 3.1 shows IdmJ of the correct mass (53 522 Da) with all four cysteine residues alkylated has an expected mass of 53 753. The mass spectrum in Figure 3.18 shows a peak at 53 754, corresponding to IdmJ with the addition of four carbamidomethyl groups. The expected mass of IdmJ† with all four cysteine residues alkylated is 56 181 Da, a protein species was observed with as mass of 56 186 Da, this is indicative of IdmJ† alkylated on all four cysteine residues. Finally, the peak at 56 368 Da indicates the presence of 7 carbamidomethyl groups, suggesting IdmJ† has been alkylated at 7 positions. Although excess iodoacetamide can eventually cause non-specific alkylation of other groups such as the amine group of lysine residues, the thioester of methionine residues, imidazole groups of histadines and carboxylate groups of aspartate or glutamate (Yang and Attygalle, 2007), the source of the extra three carbamidomethyl groups in IdmJ† is unlikely to be a non-specific alkylation. If this were the case more alkylated species would be observed e.g. +5, +6, +7 carbamidomethyl group etc. This could indicate that the source of the additional mass in IdmJ† contains four thiol groups that are being

alkylated. These results show when IdmJ† is unfolded, reduced and the cysteine residues alkylated the expected mass of IdmJ with four cysteine residues alkylated is observed. This provides further evidence of the involvement of the cysteine residues in acquiring the additional mass.

To further investigate the role of the cysteine residues in the observed increase in mass, to potentially deduce the contributions of each cysteine residue, and in an attempt to express a protein of the expected mass, individual cysteine to alanine mutations were made at all four positions. The pET(2)*idm*J plasmid DNA was purified from *E. coli* XL10 Gold Ultracompetent cells and site-directed mutagenesis carried out using primers designated "C104A For", "C104A Rev", "C127A For", "C127A Rev", "C169A For", "C169A Rev", "C308A For" and "C308A Rev" designed to mutate the cysteine residues to alanine (sequences can be found in the appendix). Mutagenesis was carried out using the QuikChange™ Lightning Site-Directed Mutagenesis kit as described in section 2.2.2.10. 3 colonies were picked from the transformation plates of the mutagenesis reaction and plasmid DNA purified. DNA sequencing was carried out using the sequencing primers "pET UPSTREAM" and "DuetDOWN1". This confirmed the presence of the correct point mutations within the *idm*J gene. Plasmids containing the genes with the single point mutations C104A, C127A, C169A and C308A were transformed into *E. coli* BL21 (DE3) Gold cells for protein expression.

Protein was expressed using the same method for wild-type IdmJ. Protein expression was carried out on a 5 mL scale and IdmJ variants were purified by nickel affinity chromatography. A 70 µL sample of each variant was desalted into 50 mM ammonium acetate pH 7.4 for analysis by ESI-MS (sections 2.2.3.9 and 2.2.3.12). Figure 3.19 below shows the deconvoluted mass spectra of IdmJ C104A, IdmJ C127A, IdmJ C169A and IdmJ C308A.

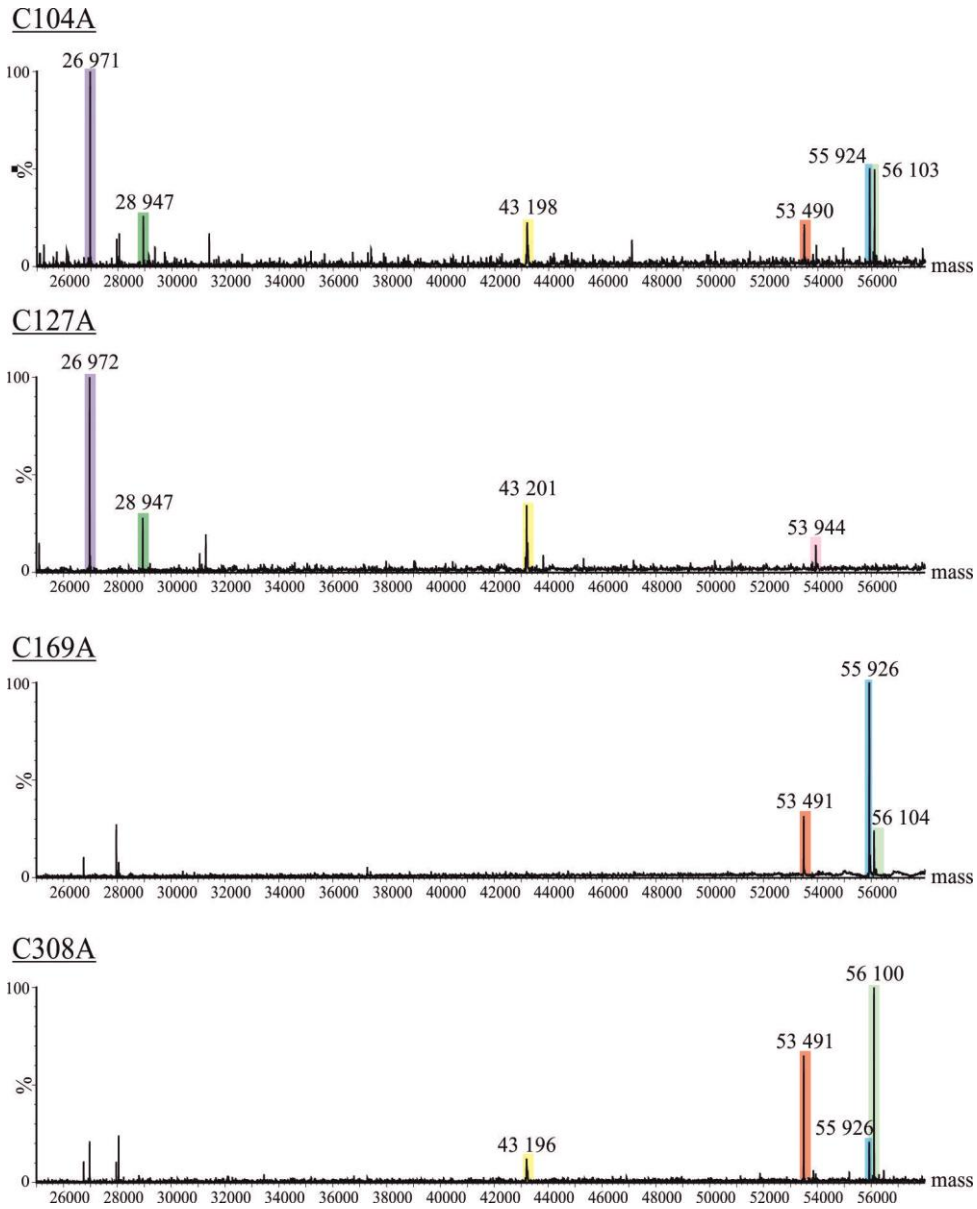


Figure 3.19- ESI-MS analysis of IdmJ variants. From top to bottom deconvoluted mass spectra of IdmJ C104A, IdmJ C127A, IdmJ C169A and IdmJ C308A. A number of protein species can be observed in each spectrum. Peaks suggested to be due to the same species are highlighted in the same colour. Table 3.2 shows the expected masses of IdmJ and IdmJ† containing the cysteine to alanine mutation.

Sample	Expected mass
IdmJ (C104A, C127A, C169A and C308A)	53 492 Da
IdmJ† (C104A, C127A, C169A and C308A)	55 923 Da

Table 3.2-Expected masses for IdmJ and IdmJ† containing the cysteine to alanine mutation.

Analysis of IdmJ variants by mass spectrometry (Figure 3.19) identified potential, consistent, degradation products seen in the four variants (masses ~ 43 200, 28 947 and 26 971 Da (purple, green and yellow)). Analysis indicated that when C104A, C169A and C308A were expressed both IdmJ and IdmJ† were present (red and blue highlighted peaks). MS analysis of these three species also show the ~178 Da adduct seen in Figure 3.14 present (light green). Expression, purification and MS analysis of C127A shows a peak at 53 944 Da (pink), this is a mass difference of 452 Da, compared to IdmJ; this does not correspond to any known particular modification. C127A is the only variant where the larger species, IdmJ†, is not observed. This may indicate that the source of the additional mass may be modification at position C127. Figure 3.20 illustrates what may be occurring in expression of IdmJ in *E. coli*.

Figure 3.20 summarises the results of the cysteine alkylation and mutagenesis experiments and a suggested theory behind what may be occurring. During alkylation (Figure 3.20 (b)) of the cysteine residues IdmJ is fully alkylated (four positions). IdmJ† is still present, however it has been alkylated four and seven times. This may indicate that the post-translational modification is occurring at one position and contains one thiol group and three slightly less reactive groups (marked x) that are able to be alkylated, accounting for the extra masses. Identification of the cysteine residue being post-translationally modified was carried out by creating alanine variants C104A, C127A, C169A and C308A (Figure 3.20 (c)). Expression of C104A, C169A and C308A showed the continued presence of the post-translational modification, as opposed to C127A suggesting this as the residue that becomes modified, potentially during expression. Removal of the modification may create an active enzyme and the additional mass (452 Da) seen in the MS of C127A could be attributed to phosphopantetheine loaded with proline (452 Da) catalysed by IdmJ.

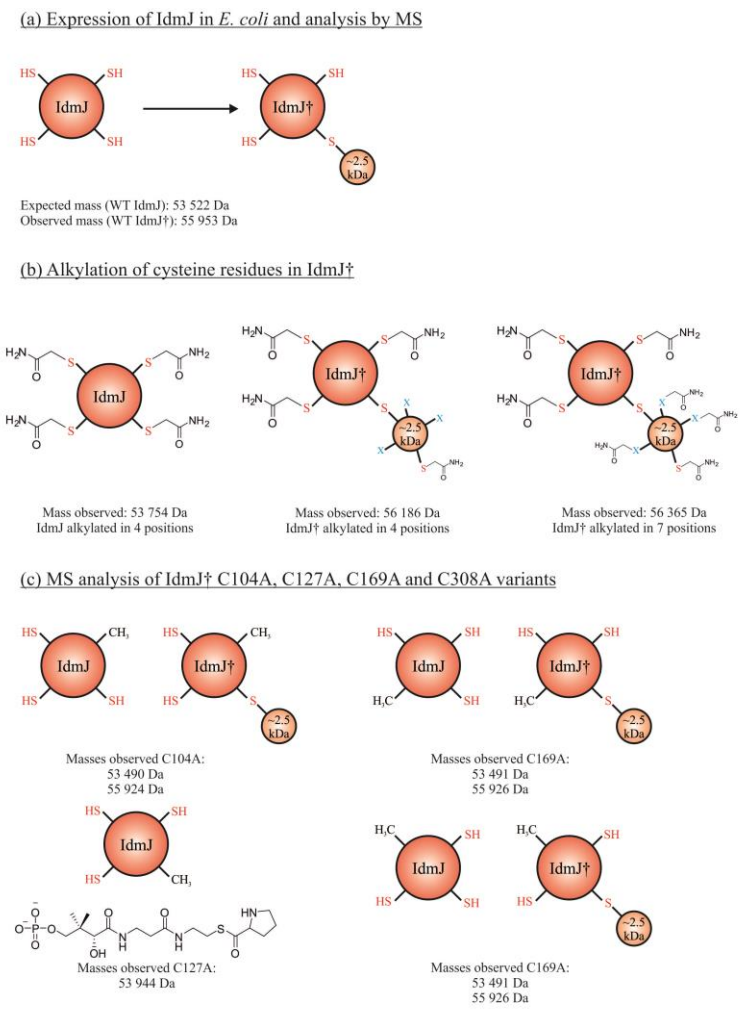


Figure 3.20- Schematic of the expression of IdmJ/IdmJ†. (a) Expression and purification of IdmJ yielded two species. Analysis by ESI-MS showed a protein species which was 2 431 Da larger than expected designated IdmJ†. (b) Alkylation of the cysteine residues yielded three species, one corresponding to IdmJ alkylated in four positions (53 754 Da), the second IdmJ† alkylated in four positions (56 186 Da) and finally IdmJ alkylated in seven positions (56 365 Da). Alkylation of IdmJ† at seven positions could indicate that the post-translational modification contains a thiol able to be alkylated as readily as the other cysteine residues and three additional groups in the post-translational modification able to also be alkylated, but at a slower rate. Systematic mutagenesis of each individual cysteine residue showed that in three variants the modification was still present. However in C127A the 2 431 Da modification was absent indicating this may be the site of modification. The mass of the protein species observed was 452 Da larger than expected for IdmJ, this corresponds to phosphopantetheine loaded with proline.

3.2.7 Summary and conclusions

Original difficulties in cloning *idm*J into pET23a may have arisen from the two restriction sites selected for cloning *idm*J, *Sac*I and *Sal*I, being too close in proximity to one another in the multiple cloning site (11 bp). Attempts at cloning *idm*J into pKK223-3 were also unsuccessful. *idm*J was finally successfully cloned into

pETDUET. Expression trials of the pET(2)*idm*J plasmid was carried out in *E. coli* BL21 (DE3) Gold cells. Soluble protein was not expressed at 37°C. Soluble protein expression was however seen at 25°C and 18°C in TB, LB and autoinduction medium (Figure 3.12). Soluble protein expression was then carried out at 18°C and in LB medium.

Protein purification was carried out using nickel affinity chromatography. Analysis by SDS-PAGE (Figure 3.13) showed purification of two protein species, and analysis by MS (Figure 3.14) showed the protein purified was 2431 Da larger than the expected mass of IdmJ believed to be a post-translational modification, this protein was designated IdmJ†. DNA sequencing and an in-gel tryptic digest indicated that the protein purified was IdmJ (Figure 3.17).

To probe whether a post-translational modification was occurring at one of the cysteine residues the protein was treated with iodoacetamide in order to alkylate the cysteine residues. Mass spectrometry results (Figure 3.18) showed masses equivalent to IdmJ being alkylated four times, equivalent to the number of cysteine residues in the protein, and IdmJ† alkylated four and seven times. Figure 3.20 shows a possible explanation behind how IdmJ† is alkylated in four and seven positions. Potentially, the modification of IdmJ† contains groups that can react with iodoacetamide.

Finally, in order to identify the site at which a potential modification could occur, the cysteine residues in IdmJ were mutated to alanine residues, the variants were then expressed, purified and analysed by MS (Figure 3.19). Purification of C104A, C169A and C308A and analysis by MS showed peaks corresponding to IdmJ and IdmJ† containing the cysteine to alanine mutation. Purification and MS analysis of C127A showed a peak at 53 944 Da, 452 Da larger than expected for IdmJ, while IdmJ† was absent. These results suggest C127 is the site of a 2431 Da modification.

3.3 The prolyl dehydrogenase (IdmI)

The formation of the starter unit, pyrrole-2-carboxylate, requires an enzyme to oxidise the prolyl group (Figure 3.1); this is the role of the prolyl dehydrogenase, IdmI. IdmI carries out the oxidisation of the prolyl group in a flavin-dependent

manner. To successfully reconstitute this system in *E. coli* cloning and expression of IdmI was required.

3.3.1 Cloning the prolyl dehydrogenase (*idm*I) into pET23a

As mentioned previously, the original strategy was to clone each individual gene into the expression vector pET23a. The *idm*I gene was engineered to have *Bam*HI and *Eco*RI sites at the 5' and 3' ends of the DNA respectively. The cut and paste method illustrated in Figure 3.3 was the initial method selected for sub-cloning *idm*I from pUC*idm*I into pET23a, creating a pET*idm*I plasmid. Insert and vector DNA were prepared by digesting the pUC*idm*I plasmid and pET23a vector with *Bam*HI-HF® and *Eco*RI-HF® restriction enzymes for 16 hr. The restriction digests were run on a 0.7 % (w/v) agarose gel and the insert and vector DNA were purified from the agarose gel (section 2.2.2.3). Insert and vector DNA were ligated as described in section 2.2.2.8. Ligated DNA was transformed into *E. coli* XL10 Gold Ultracompetent cells and plated onto 2 × TY agar plates containing ampicillin to select for the pET*idm*I contruct. Plasmid DNA was purified from 10 individual colonies and restriction digests performed using *Bam*HI-HF® and *Eco*RI-HF® restriction enzymes to screen for the correct construct. When digested with *Bam*HI and *Eco*RI the correct construct should yield bands at approximately 3.6 kb (3666 bp) and 1.2 kb (1180 bp) corresponding to the vector and insert respectively.

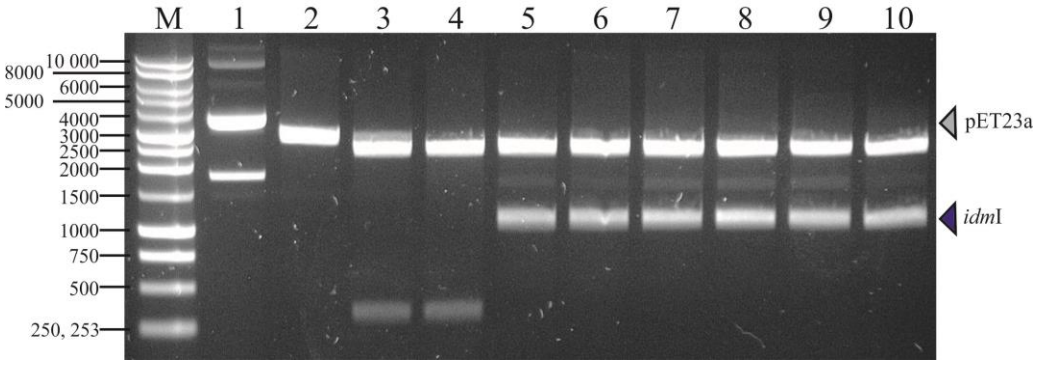


Figure 3.21-0.7 % (w/v) agarose gel of plasmid DNA purified from colonies to screen for the pET*idm*I plasmid. Purified plasmid DNA digested with *Bam*HI-HF® and *Eco*RI-HF®. Plasmid DNA in lane 1 was undigested; lane 2 shows a single DNA band at approximately 3000 bp. Digests of plasmids in lanes 3 and 4 yielded two DNA fragments, one at 2700 bp corresponding to pUC57 and the other an unidentified band at approximately 300 bp. Digestion of plasmid DNA in lanes 5-10 yielded two DNA fragments, one 2700 bp in length and a second 1180 bp this corresponds to pUC57 and *idm*I respectively.

Figure 3.21 indicates that the ligation between pET23a and *idm*I was unsuccessful. Plasmid 1 does not appear to have been digested with *Bam*HI-HF® and *Eco*RI-HF®. There is a single DNA band in lane 2 of approximately 3000 bp, there was no insert of 1180 bp corresponding to *idm*I. Plasmid DNA in lanes 3 and 4 yielded two DNA fragments when digested with *Bam*HI-HF® and *Eco*RI-HF®, one 2700 bp in length and a second 300 bp in length, corresponding to pUC57 and an unidentified insert. Plasmid DNA in lanes 5-10 yielded two DNA fragments when digested with *Bam*HI-HF® and *Eco*RI-HF®, one 2700 bp in length and the other 1180 bp in length corresponding to pUC57 and *idm*I. These results indicated that the ligation reaction was contaminated by the pUC*idm*I plasmid, as well as other unknown contaminants. The preparation of insert and vector DNA was repeated as was the ligation, with no success. Potentially, as with the pET23a digest for *idm*J, the proximity of the two restriction sites being utilised in pET23a may have prevented efficient digestion, even with an extended incubation period. Additionally incomplete digestion of the pUC*idm*I plasmid was contaminating the transformation reaction. An alternative approach of cloning *idm*I into pKK223-3 was then taken.

3.3.2 Cloning the prolyl dehydrogenase (*idm*I) into pKK223-3

In order to clone *idm*I into pKK223-3, creating the plasmid pK*idm*I plasmid, a similar strategy was used as illustrated in Figure 3.6. However, PCR had to be utilised to alter the restriction sites at the 5' and 3' ends amplifying the insert *idm*I. *idm*I was originally designed to have an *Bam*HI site at the 5' end of the gene and *Eco*RI site at the 3' end of the gene, however to use the pKK223-3 vector from the pK*nan*A precursor the *nan*A gene must be cut out using an EcoRI site at the 5' end and *Pst*I site at the 3' end of the *nan*A gene.

Primers designated "IdmI *Eco*RI For" and "IdmI *Pst*I Rev" were designed to change the 5' BamHI site to an EcoRI site and the 3' *Eco*RI site to a *Pst*I site (Sequences can be found in the appendix). A PCR reaction was performed using Pfu Turbo DNA polymerase (section 2.2.2.4). Figure 3.22 shows samples taken from PCR reactions set-up to contain 0.1, 1 and 10 ng of template DNA.

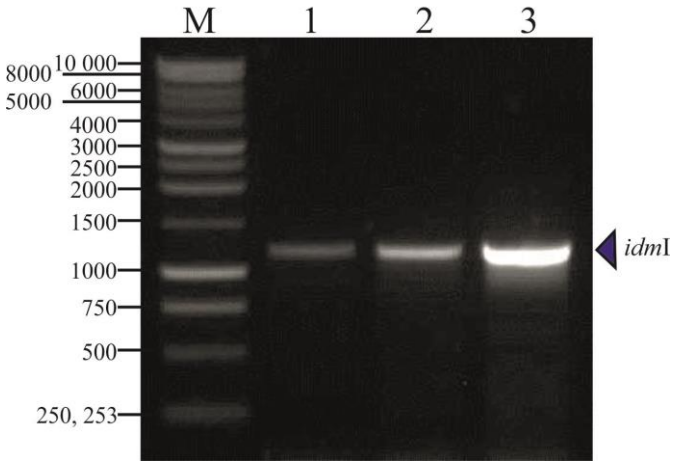


Figure 3.22- 0.7 % (w/v) agarose gel of the PCR reaction to amplify the *idm*I gene. PCR reactions were set up with (1) 0.1 ng (2) 1 ng and (3) 10 ng template DNA. *idm*I was amplified from the pUC*idm*I using primers to change the 5' and 3' restriction sites to EcoRI and PstI respectively.

Insert DNA from the 10 ng template PCR reaction (Figure 3.22, lane 3) was purified using a QIAquick® PCR purification kit (section 2.2.2.5) and digested with *Eco*RI-HF® and *Pst*I-HF®. The subsequent product, the insert, was purified in the same manner as the PCR product. A ligation reaction was then performed between the insert and the pK*nan*A vector digested with *Eco*RI-HF® and *Pst*I-HF®. Subsequent constructs from the ligation reaction were transformed into *E.coli* XL10 Gold cells and the correct plasmid screened for by restriction digests using *Eco*RI-HF® and *Pst*I-HF®. Figure 3.23 shows purified plasmid DNA from the pKK223-3 ligation with *idm*I digested with *Eco*RI-HF® and *Pst*I-HF®.

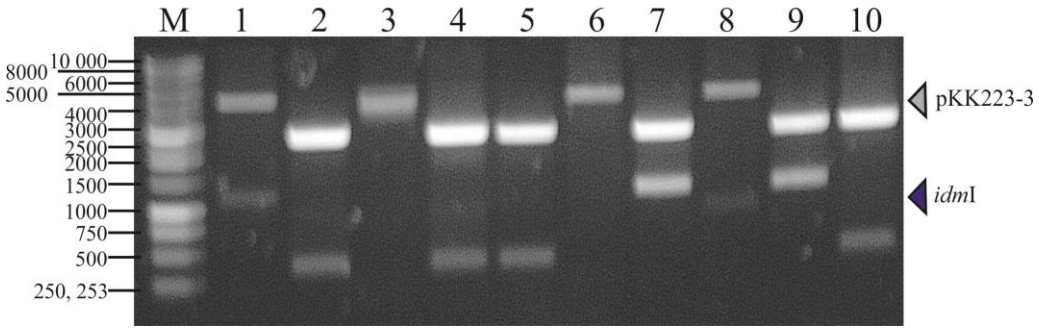


Figure 3.23-0.7 % (w/v) agarose gel of plasmid DNA purified ten colonies from the ligation between *idm*I and pKK223-3 after being digested with *Eco*RI-HF® and *Pst*I-HF®. A pK*idm*I plasmid digested with *Eco*RI and *Pst*I should yield two DNA fragments, one at 4.6 kb and the other at 1.2 kb. Plasmid in lanes 2, 4, 5 and 10 show two DNA fragments, one fragment 2700 bp in length corresponding the correct length of pUC57, and a second unknown DNA fragment approximately 300 bp. Plasmid DNA in lanes 3 and 6 show a single band at approximately 4.6 kb corresponding to pKK223-3. Plasmid DNA in lanes 7 and 9 yielded two bands, one at 2.7 kb corresponding to pUC57 and the second unknown DNA fragment approximately 1.5 kb in length. Plasmid DNA digested in lanes 1 and 8 has yielded two fragments, one at 4.6 kb and the other at 1.2 kb corresponding to pKK223-3 and *idm*I respectively.

A pK*idm*I plasmid digested with *Eco*RI-HF® and *Pst*I-HF® should yield two bands, one at approximately 4.6 kb (4586 bp) and a second at 1.2 Kb (1180 bp). Plasmid DNA in lanes 2, 4, 5 and 10 show bands at 2700 bp corresponding to pUC57 and an unknown DNA fragment 300 bp in length. Plasmid DNA in lanes 3 and 6 digested with *Eco*RI-HF® and *Pst*I-HF® yielded one DNA fragment approximately 4.6 kb corresponding to pKK223-3, without an insert present. Digestion of plasmids 7 and 9 (lane 7 and 9) yielded two DNA fragments, 2700 bp and 1500 bp corresponding to pUC57 and an unknown insert respectively. Finally, plasmid DNA in lanes 1 and 8 digested with *Eco*RI-HF® and *Pst*I-HF® yielded two DNA fragments, one 4.6 kb in length and a second 1.2 kb indicating that plasmids 1 and 8 are the pK*idm*I construct. Subsequent DNA sequencing of plasmid 1 from Figure 3.22 provided further confirmation of obtaining the correct construct, pK*idm*I plasmid (section 2.2.2.12). The pK*idm*I plasmid was transformed into *E. coli* XL10 Gold cells and a glycerol stock made (section 2.2.1.6 and 2.2.2.11).

3.3.3 Recombinant protein expression of the prolyl dehydrogenase (IdmI)

After successful sub-cloning of *idm*I into pKK223-3, the pK*idm*I construct was transformed into *E. coli* BL21 (DE3) Gold cells for protein expression and a glycerol stock made. Small-scale protein expression trials were attempted on a 5 mL culture

scale in TB, 2 × TY, LB and autoinduction medium at 37 °C, 21 °C and 18 °C as described for the pET*idm*J plasmid (section 3.2.4). The soluble cellular components were subject to analysis by SDS-PAGE in order to identify protein expression. The expected mass for IdmI as calculated from the amino acid sequence of the synthetic gene is 41 191 Da. The blue arrow in Figure 3.24 indicates the expected mass of IdmI.

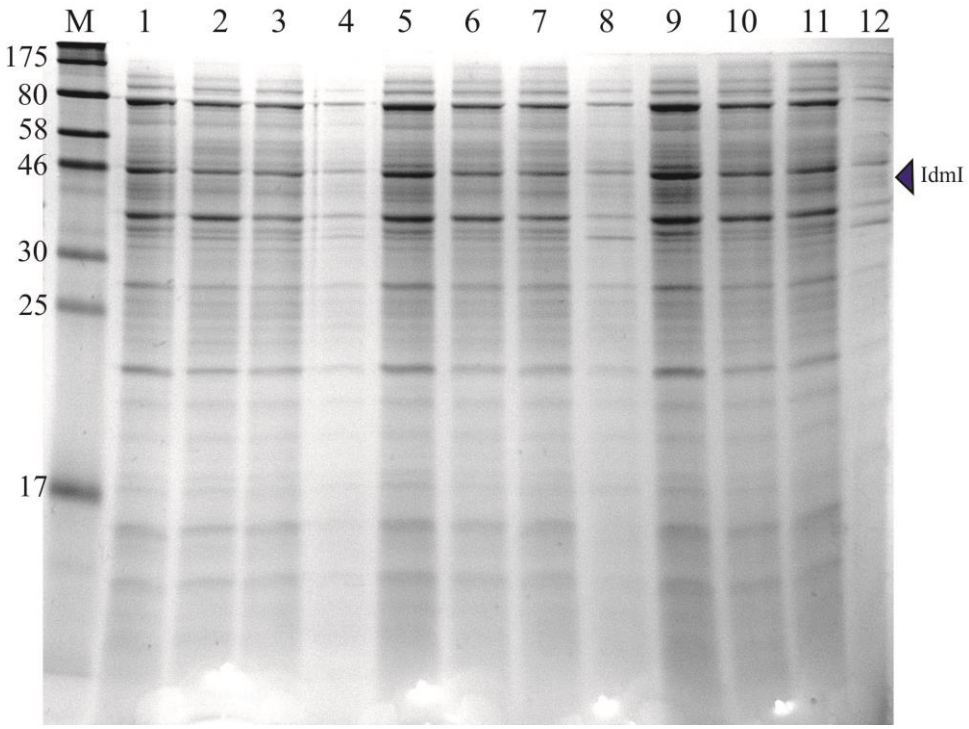


Figure 3.24- Reducing SDS-PAGE analysis of soluble protein produced in the expression conditions trialled for pK*idm*I. Expression conditions for pK*idm*I. Lanes 1-3 is soluble protein expressed in TB medium at 37 °C, room temperature and 18 °C respectively. Lanes 4-6 is soluble protein expressed in 2 × TY medium 37 °C, room temperature and 18 °C respectively. Lanes 7-9 is soluble protein expressed in LB medium at 37 °C, room temperature and 18 °C respectively. Finally lanes 10-12 is soluble protein expressed in autoinduction medium at 37 °C, room temperature and 18 °C respectively. The blue arrow indicates where a protein band for IdmI would occur.

The SDS-PAGE gel in Figure 3.24 shows the soluble protein expressed in *E.coli* BL21 (DE3) Gold cells containing the pK*idm*I plasmid in TB, 2 × TY, LB and autoinduction medium at 37°C, 21°C and 18°C. There does not appear to be a protein band at the expected mass for IdmI in the SDS-PAGE gel. This suggests that IdmI was not expressed as a soluble protein, therefore as with the pET(2)*idm*J plasmid, an anti polyHistidine antibody was used as a more sensitive probe for

protein expression in the soluble and insoluble cell fractions. Figure 3.25 shows the results for probing for the His_6-tag of IdmI in the soluble and insoluble fractions.

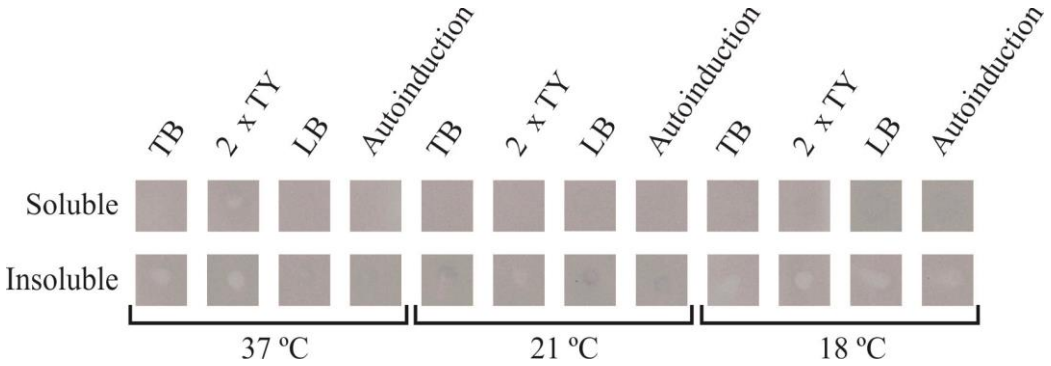


Figure 3.25- N-terminal His_6-tag detection in IdmI expression trials. Soluble (top row) and insoluble (bottom row) fractions were blotted onto nitrocellulose membrane and an anti His_6-tag antibody was used to detect the N terminal His_6-tag of IdmI, probing for protein expression. Blots are of *E. coli* BL21 (DE3) Gold cells containing the pK*idm*I plasmid grown in TB, 2 × TY, LB and autoinduction media at 37°C, 21°C and 18°C. The blot indicates that there was little, if any His_6-tagged protein expressed from the pK*idm*I construct. This is a negative display image for clarity.

In order to identify expression conditions that produced soluble protein an anti polyHistidine antibody was used to detect the N terminal His_6-tag of IdmI from the soluble and insoluble fractions of the expression trial. Figure 3.25 shows no reactivity patches indicating there was too little or no His_6-tagged protein present. Expression of IdmI from the pK*idm*I plasmid in *E. coli* BL21 (DE3) Gold cells was unsuccessful in TB, 2 × TY, LB and autoinduction medium at 37°C, 21°C and 18°C. This therefore led to exploring alternative methods for protein expression.

3.3.4 Cloning the prolyl dehydrogenase (*idm*I) into pMAL-c5X

Since soluble and insoluble IdmI was not expressed from the pK*idm*I plasmid, it was decided to clone IdmI into a pMAL vector, creating pMAL*idm*I. Cloning IdmI into a pMAL vector will create a fusion protein with a maltose binding protein (MBP) tag. Although a clear explanation of why MBP helps with protein solubility does not exist, it has been hypothesised that a large soluble affinity tag will encourage the otherwise insoluble protein to be more soluble (Rosano and Ceccarelli, 2014).

To create an MBP-IdmI fusion protein the restriction sites for *idm*I were changed to *Nde*I and *Eco*RI restriction sites at the 5' and 3' of the gene respectively using the

"IdmI *Nde*I For" and "IdmI *Eco*RI Rev" in a similar manner to the pK*idm*I plasmid, using the same PCR conditions (section 3.3.2) (primer sequences are located in the appendix). The pMAL vector, pMAL-c5X, was purchased from New England Biolabs® (Ipswich, MA, USA) (kindly gifted by Dr James Ross, Astbury Centre, University of Leeds). The insert and vector DNA were prepared by digesting with *Nde*I and *Eco*RI-HF® and a ligation reaction performed as described previously for the pK*idm*I construct (section 3.3.2). Plasmid DNA was purified from 8 individual colonies selected from the pMAL-C5X and *idm*I ligation transformation plate and the pMAL*idm*I plasmid screened for by restriction digest with *Nde*I and *Eco*RI-HF®. Figure 3.26 below shows the results from the restriction digest.

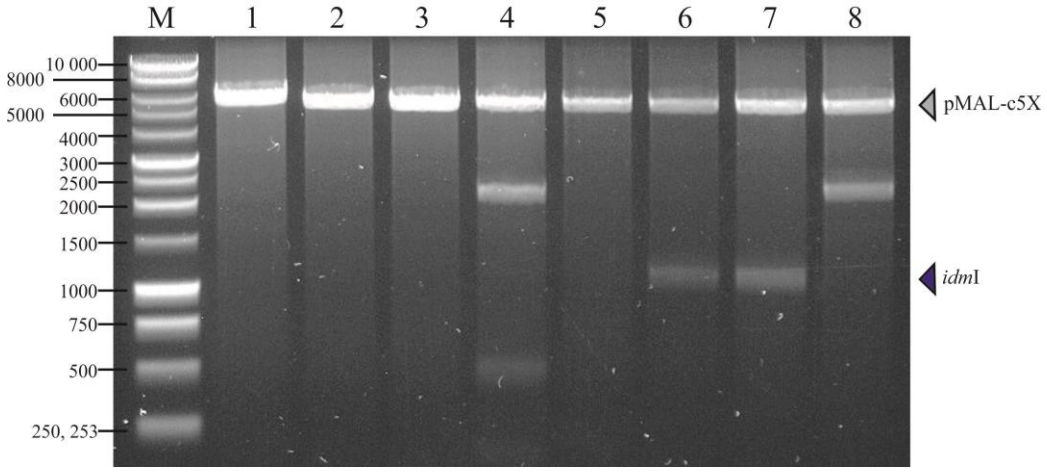


Figure 3.26-0.7 % agarose gel of restriction digests with *Nde*I and *Eco*RI restriction enzymes of purified plasmid DNA from the pMAL-C5X vector and *idm*I ligation. When digested with NdeI and EcoRI a pMAL*idm*I construct would yield two DNA fragments, one 5.6 kb in length (vector) and the other 1.2 kb in length (*idm*I). Plasmid DNA in lanes 1-3 and 5, when digested with NdeI and EcoRI, yielded a single band at approximately 5.6 kb corresponding to pMAL-C5X. Restriction digests of Plasmid DNA in lanes 4 and 8 shows two bands, one corresponding to the vector (5.6 kb), the other approximately 2 kb in length, too large to be *idm*I. Finally, plasmid DNA in lanes 6 and 7 shows two bands, one at 5.6 kb corresponding to the vector pMAL-C5X and the other the correct length for *idm*I (1.2 kb).

Lanes 6 and 7 in Figure 3.26 suggests the successful sub-cloning of *idm*I into the pMAL-c5X. Plasmid 6 (lane 6) was transformed into *E. coli* XL10 Gold cells and was confirmed to be the pMAL*idm*I construct by DNA sequencing (section 2.2.2.12). The pMAL*idm*I plasmid was also transformed into *E. coli* BL21 (DE3) Gold cells for protein expression and a glycerol stock made of these cells containing this plasmid too.

3.3.5 Expression of the MBP/prolyl dehydrogenase fusion protein

A small scale protein expression trial of *E. coli* BL21 (DE3) Gold cells containing the pMAL*idm*I plasmid was carried out as previously described for the pK*idm*I construct, screening TB, 2 × TY, LB and autoinduction medium at 37 °C, room temperature and 18 °C. The soluble cellular components were analysed by SDS-PAGE. As mentioned previously, the expected mass for IdmI is 41 191 Da, however the MBP-IdmI fusion protein will have a total mass of approximately 83 kDa indicated by the blue arrow in Figure 3.27.

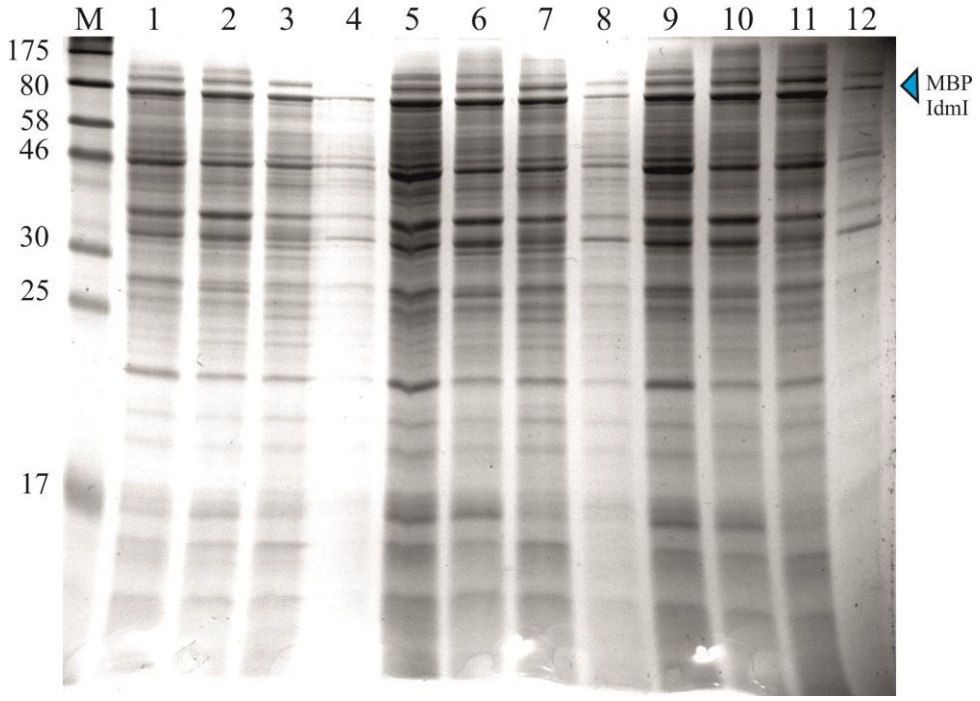


Figure 3.27- Reducing SDS-PAGE analysis of expression trials of the pMAL*idm*I fusion protein. Protein expression trials of MBP-IdmI were carried out with *E.coli* BL21 (DE3) Gold cells containing the plasmid pMAL*idm*I being grown and induced in TB, 2 × TY, LB and autoinduction media at 37 °C, room temperature and 18 °C. Lanes 1-3 is soluble protein expressed in TB media at 37 °C, room temperature and 18 °C respectively. Lanes 4-6 is soluble protein expressed in 2 × TY media 37 °C, room temperature and 18 °C respectively. Lanes 7-9 is soluble protein expressed in LB medium at 37 °C, room temperature and 18 °C respectively. Finally lanes 10-12 is soluble protein expressed in autoinduction medium at 37 °C, room temperature and 18 °C respectively. The blue arrow indicates where a protein band for MBP-IdmI fusion protein would be expected to occur, at approximately 83 kDa in mass.

The expected mass for the MBP-IdmI fusion protein is approximately 83 kDa. There is no clear overexpression of soluble MBP-IdmI in TB, 2 × TY, LB or autoinduction

media in the range of temperatures trialled as illustrated in the SDS-PAGE gel in Figure 3.27 by comparison with uninduced *E. coli* BL21 (DE3) Gold cells (see appendix). In order to determine if the MBP-IdmI fusion protein was being expressed as a soluble or insoluble protein, blotting for the His_6-tag using an anti polyHistidine antibody was performed on both the soluble and insoluble cellular components as carried out for the pET(2)*idm*J and pK*idm*I constructs (sections 3.2.4 and 3.3.3). Figure 3.28 below shows results when blotting for the His_6-tag in soluble and insoluble fractions of the expression of pMAL*idm*I in TB, 2 × TY, LB and autoinduction media at 37°C, 21°C and 18°C.

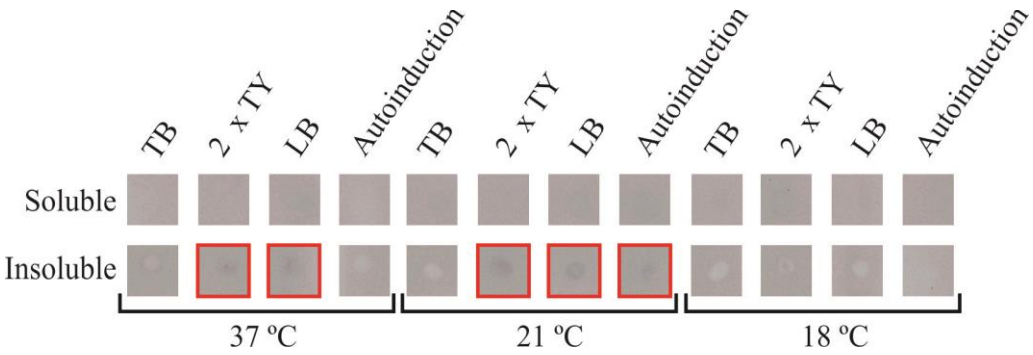


Figure 3.28- Detection of the N-terminal His_6-tag MBP-IdmI fusion protein. Soluble (top row) and insoluble (bottom row) fractions from *E. coli* BL21 (DE3) Gold cells containing the pMAL*idm*I construct being expressed in TB, 2 × TY, LB and autoinduction medium at three different temperatures. The His_6-tag was detected using an anti His_6-tag antibody. The blot above indicates that extremely low levels of MBP-IdmI fusion protein were in the insoluble fractions of cells grown at 21ºC and 37 ºC. This is a negative display image for clarity.

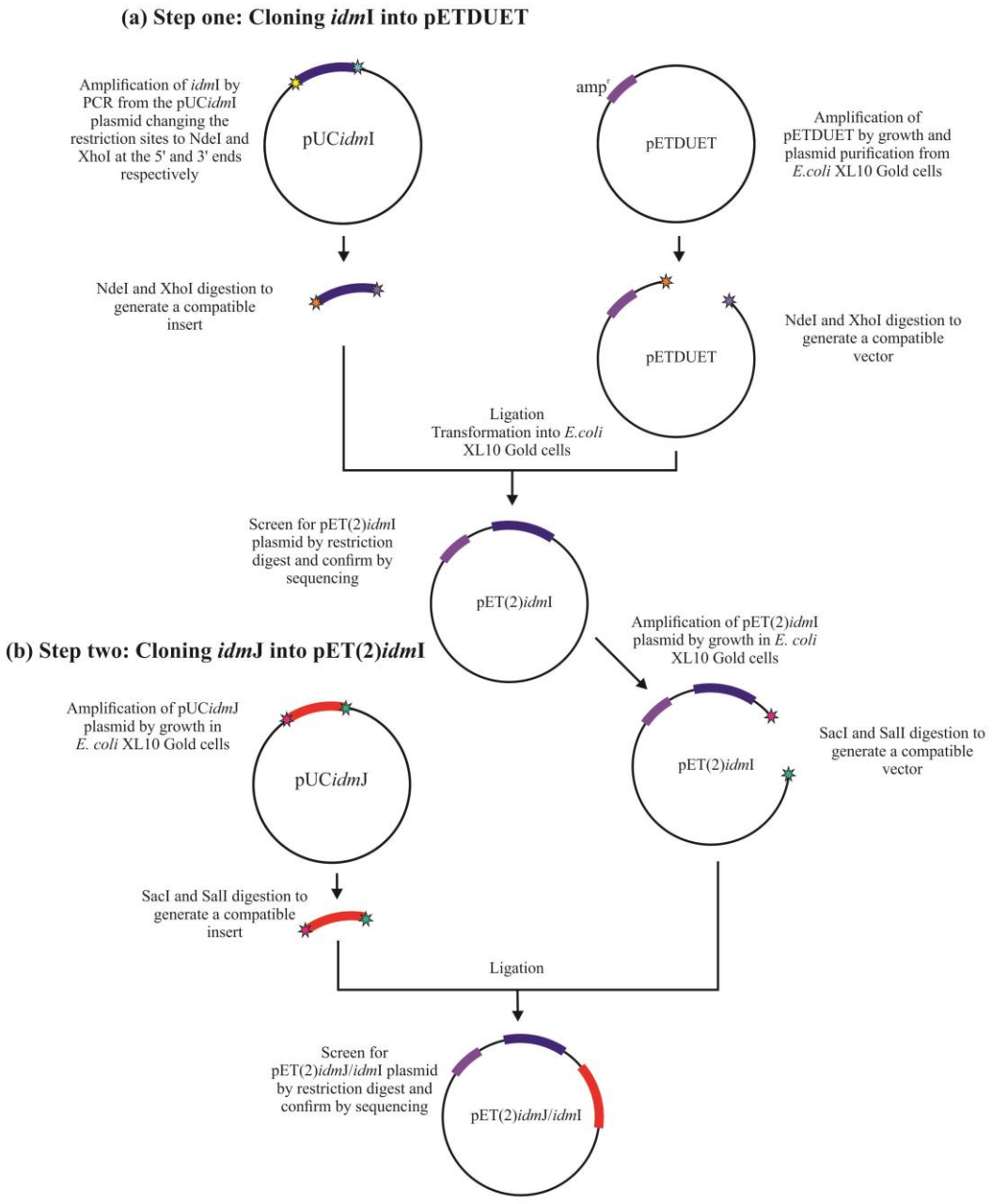
The MBP-IdmI fusion protein was mainly expressed as an insoluble protein, extremely low levels of His_6-tagged protein is present in the insoluble fractions of cells grown in 2 × TY, LB and autoinduction medium at 37°C and 21°C as illustrated by Figure 3.28 (red boxes). Soluble expression of the MBP-IdmI fusion protein was not achieved.

Thus far the expression of soluble IdmI had been unsuccessful, expressing the protein individually and as a fusion protein, however there were still a number of avenues to explore. The next strategy was to clone all three genes into the same vector and express them together. In the natural host all three proteins would be

expressed together and expression of IdmI may dependent upon expression of the other proteins within the same module (Figure 3.1).

3.3.6 Cloning genes responsible for starter unit biosynthesis into pETDUET

In an effort to express the prolyl dehydrogenase (IdmI) as a soluble protein all three protein were to be cloned into the same vector and expressed simultaneously. Figure 3.29 illustrates the strategy for cloning the adenylyltransferase (IdmJ) and proline carrier protein (IdmK) into the same vector as IdmI creating a plasmid designated pET(2)*idm*J/*idm*K/*idm*I.



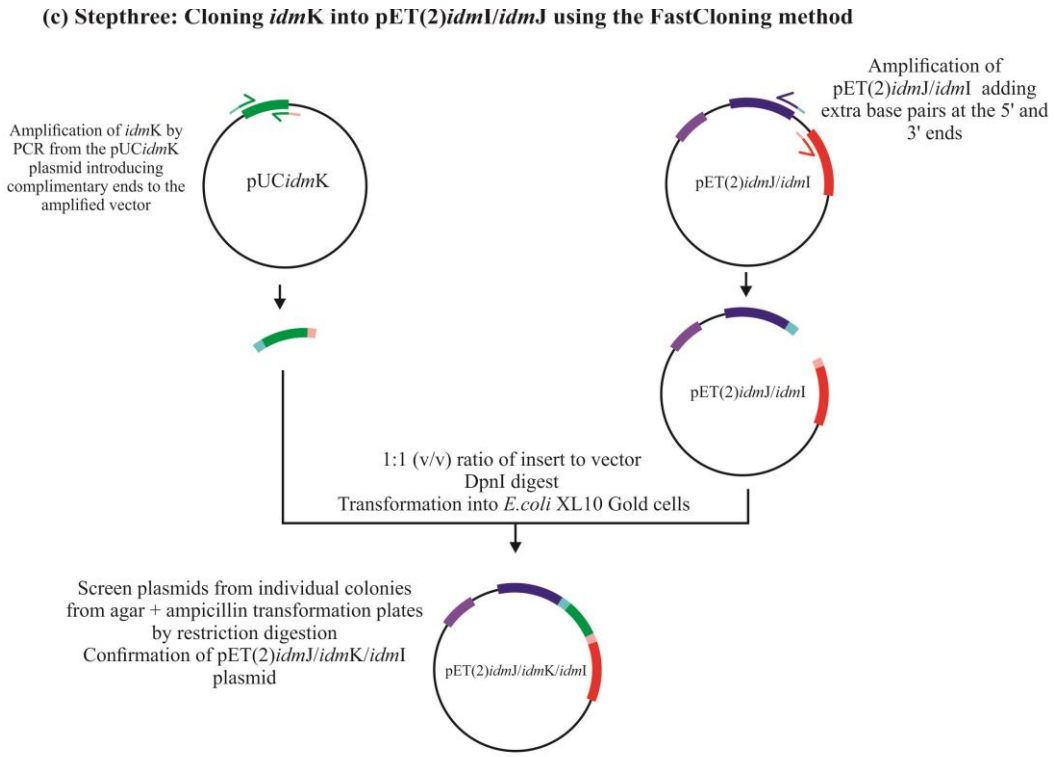


Figure 3.29- Proposed strategy for cloning *idm*J, *idm*I and *idm*K into pETDUET. (a) Step one involves cloning *idm*I into multiple cloning site 2 in pETDUET. PCR amplification will be used to amplify *idm*I from the pUC*idm*I plasmid altering the 5' and 5' restriction sites to *Nde*I and *Xho*I. pETDUET will be amplified by growth of the plasmid in *E. coli* XL10 Gold cells. Both the insert and vector will be digested with *Nde*I and *Xho*I restriction enzymes, and are subsequently ligated together. The resulting constructs will be transformed into *E. coli* XL10 Gold cells and a potential pET(2)*idm*I plasmid screened for by restriction digests with *Nde*I and *Xho*I. Sequencing will confirm the pET(2)*idm*I plasmid. (b) The next step is to clone *idm*J into multiple cloning site 1 of pETDUET. The pET(2)*idm*I plasmid and pUC*idm*J plasmid were amplified by growth in *E. coli* XL10 Gold cells. Both the vector and insert DNA were prepared by digestion with *Sac*I and *Sal*I. A ligation reaction between the insert and vector DNA followed by subsequent transformation into *E. coli* XL10 Gold cells will be carried out. Resulting constructs will be screened by restriction digestion with *Sac*I and *Sal*I, the resulting plasmid pET(2)*idm*J/*idm*I will be confirmed by sequencing. (c) The final step, cloning *idm*K into pET(2)*idm*J/*idm*I, was to be cloned in a ligation independent manner between *idm*J and *idm*I. PCR primers with complimentary ends to each other are to be used to amplify both the insert and vector. Insert and vector DNA can then be digested by *Dpn*I in a 1:1 vector to insert (v:v) ratio and then transformed into *E. coli* XL10 Gold cells. Subsequent constructs can be screened by restriction digestion and DNA sequencing used to confirm construction of the pET(2)*idm*J/*idm*K/*idm*I plasmid.

The first step to create the pET(2)*idm*J/*idm*K/*idm*I is to clone *idm*I into the multiple cloning site 2 of pETDUET. To do this, PCR primers were designed to alter the 5' and 3' restriction sites to *Nde*I and *Xho*I. Insert DNA was amplified by PCR which

simultaneously changed the 5' and 3' end restriction sites, pETDUET was amplified by growth of *E. coli* XL10 Gold cells containing the pETDUET plasmid. Both insert and vector DNA were digested with *Nde*I and *Xho*I and subsequently ligated together. A transformation was performed of resulting constructs, colony growth on 2 × TY agar plates containing ampicillin indicated the presence of the pET(2)*idm*I plasmid. The plasmid was screened for by digestion with NdeI and XhoI restriction enzymes. Figure 3.30 (a) shows an agarose gel of the PCR reaction carried out to amplify *idm*I from the pUC*idm*I using "*Nde*I IdmI For" and "*Xho*I IdmI Rev" primers (primer sequences located in the appendix). Figure 3.30 (b) shows the agarose gel of four individual plasmids purified from four individual colonies picked from the transformation plates of resulting constructs from the ligation between pETDUET and *idm*I, creating a pET(2)*idm*I plasmid.

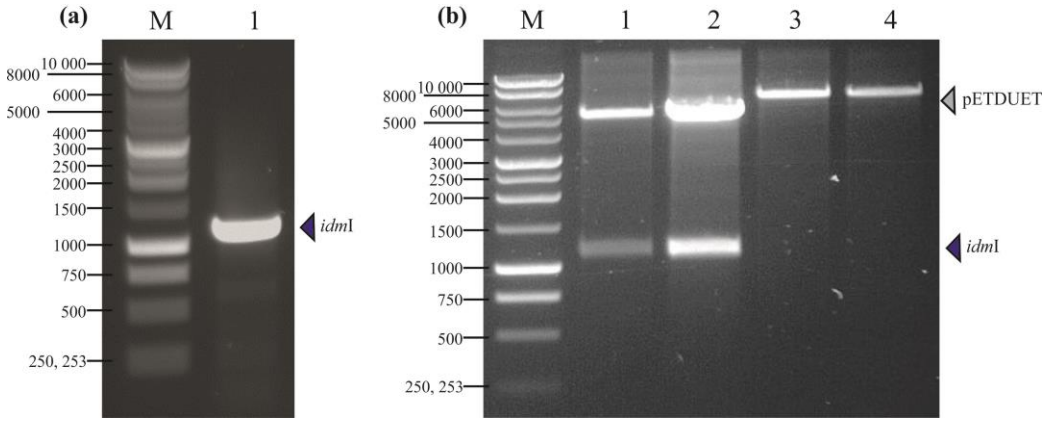


Figure 3.30-0.7 % agarose gels displaying the results of steps taken to clone *idm*I into pETDUET. (a) *idm*I was amplified by PCR using "*Nde*I IdmI For" and "*Xho*I IdmI Rev" primers, the PCR reaction was carried out with 10 ng of template DNA. The Blue arrow indicates the size amplified *idm*I was expected at. (b) Four individual colonies were picked from the transformation plates of the ligation reaction between *idm*I and pETDUET, plasmid DNA was purified and digested with *Nde*I and *Xho*I. Plasmids in Lanes 1 and 2 show DNA fragments at 5.5 kb corresponding to pETDUET and a second fragment at approximately 1.2 kb corresponding to the *idm*I gene. Plasmid DNA in lanes 3 and 4 digested with *Nde*I and *Xho*I yielded a single unknown DNA fragment approximately 8 kb in length.

The PCR amplification of *idm*I was successful (Figure 3.30 (a)). A positive pET(2)*idm*I clone would yield two bands on an agarose gel, one at approximately 5.5 kb (5420 bp) corresponding to pETDUET, a second 1.2 kb (1180 bp) corresponding to *idm*I. Lanes 1 and 2 in the agarose gel in Figure 3.30 (b) show DNA bands at these expected sizes, suggesting the ligation between pETDUET and *idm*I was successful. Plasmid DNA in lane 1 was subsequently transformed into *E. coli* XL10

Gold cells (section 2.2.2.11), a glycerol stock made and subject to DNA sequencing to confirm this plasmid was the pET(2)*idm*I construct (section 2.2.1.6 and 2.2.2.12).

The next stage was to clone *idm*J into multiple cloning site one of the pET(2)*idm*I plasmid. Figure 3.29 (b) illustrates the strategy to be used to create the pET(2)*idm*J/*idm*I plasmid. In brief, the standard cut and paste method was used to cut *idm*J out of the pUC*idm*J plasmid using the *Sac*I and *Sal*I sites and ligated into the pET(2)*idm*I plasmid, also digested with *Sac*I-HF® and *Sal*I-HF® restriction enzymes. A ligation reaction was carried out as described in section 2.2.2.8. 8 colonies were picked from the transformation plate from the ligation reaction between pET(2)*idm*I and *idm*J and the pET(2)*idm*J/*idm*I plasmid screened for by digestion with *Sac*I-HF® and *Sal*I-HF® restriction enzymes. A pET(2)*idm*J/*idm*I plasmid digested with *Sac*I-HF® and *Sal*I-HF® would yield two DNA fragments, one at approximately 6.7 kb corresponding to pETDUET (5420 bp) and *idm*I (1180 bp) and a second DNA fragment at 1.5 kb corresponding to *idm*J (1540 bp).

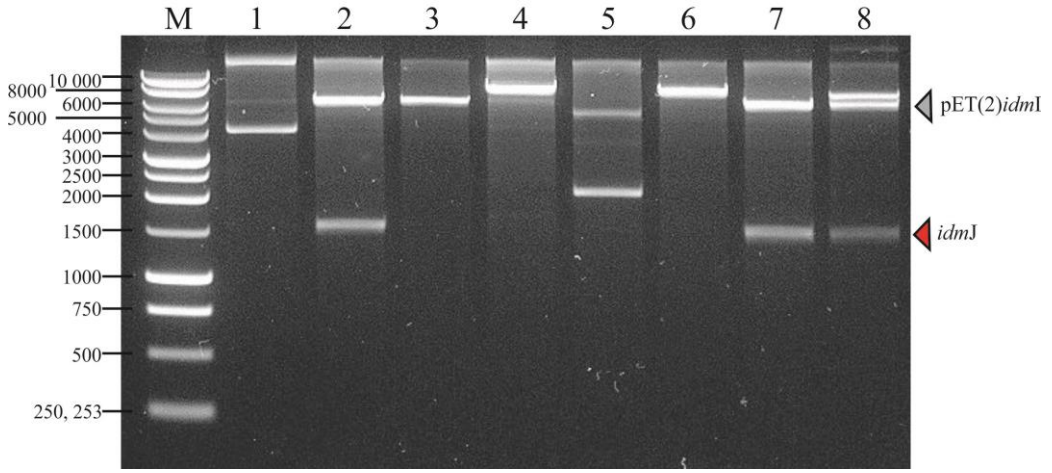


Figure 3.31-0.7% agarose gel of restriction digests with *Sac*I-HF® and *Sal*I-HF® of plasmid DNA from 8 colonies screening for the pET(2)*idm*J/*idm*I plasmid. 8 individual colonies were selected from the resulting constructs from the pET(2)*idm*I and *idm*J ligation transformation. Digestion of plasmid DNA in lane 1 yielded two DNA fragments, both unknown fragments, one fragment >10 kb and the second approximately 4.5 kb. Digestion of plasmid DNA in lanes 2, 7 and 8 yielded two DNA fragments, one at approximately 6.7 kb and a second at 1.5 kb corresponding to pET(2)*idm*I and *idm*J respectively. Plasmid DNA in lane 3, digested with *Sac*I-HF® and *Sal*I-HF® yielded at single band at approximately 5.5 kb, corresponding to the pET(2)*idm*I plasmid. Plasmid DNA in lanes 4 and 6 also yielded a single band when digested with *Sac*I-HF® and *Sal*I-HF®, this time an unknown DNA fragment at approximately at 8 kb. Finally restriction digestion of plasmid DNA in lane 5 yielded two DNA fragments one at ~5 kb and a second at 2 kb from an unknown origin.

To create the pET(2)*idm*J/*idm*I plasmid, *idm*J had to be cloned into the first multiple cloning site of pETDUET using the *Sac*I and *Sal*I restriction sites. Figure 3.31 is the agarose gel of the resulting constructs from a ligation between pET(2)*idm*I and *idm*J digested with *Sac*I-HF® and *Sal*I-HF®. Plasmid DNA in lanes 1, 4, 5 and 6 yielded DNA fragments of varying sizes that were unidentified. Plasmid DNA in lane 3 corresponded to the pET(2)*idm*I plasmid, *idm*J was not present. Although the plasmid in lane 8 appears to display bands corresponding to pET(2)*idm*I and *idm*J there are additional bands above the band corresponding to pET(2)*idm*I, therefore indicating the presence of DNA fragments of unknown origin. Plasmid DNA in lanes 2 and 7 however did yield two DNA fragments, one at 6.7 kb and a second at 1.5 kb corresponding to pET(2)*idm*I and *idm*J respectively. This indicates that the pET(2)*idm*J/*idm*I construct has been made. Plasmid DNA from lane 2 was transformed into *E. coli* XL10 Gold cells, a glycerol stock made. DNA sequencing with sequencing primers "pET UPSTREAM" and "DUETDOWN1" was used to confirm that *idm*J was cloned into multiple cloning site one of pETDUET (primer sequences are located in the appendix).

The final stage in creating the pET(2)*idm*J/*idm*K/*idm*I clone was to utilise a ligation independent cloning method called FastCloning (Li *et al.*, 2011) to sub-clone *idm*K into the pET(2)*idm*J/*idm*I plasmid to express all three proteins together. Figure 3.28 (c) illustrates the FastCloning strategy to be used. In short, insert DNA from pUC*idm*K was amplified using primers designated "IdmK Triple Clone For" and "IdmK Triple Clone Rev". The PCR primers used to amplify the pET(2)*idm*J/*idm*I plasmid were designated "pET(2)*idm*J/*idm*I Triple Clone For" and "pET(2)*idm*J/*idm*I Triple Clone Rev". Amplification with these primers created overlapping complimentary ends between "IdmK Triple Clone For" and "pET(2)*idm*J/*idm*I Triple Clone Rev" and between "IdmK Triple Clone Rev" and "pET(2)*idm*J/*idm*I Triple Clone For" (primer sequences are located in the appendix). Results from the PCR reactions can be seen in Figure 3.32.

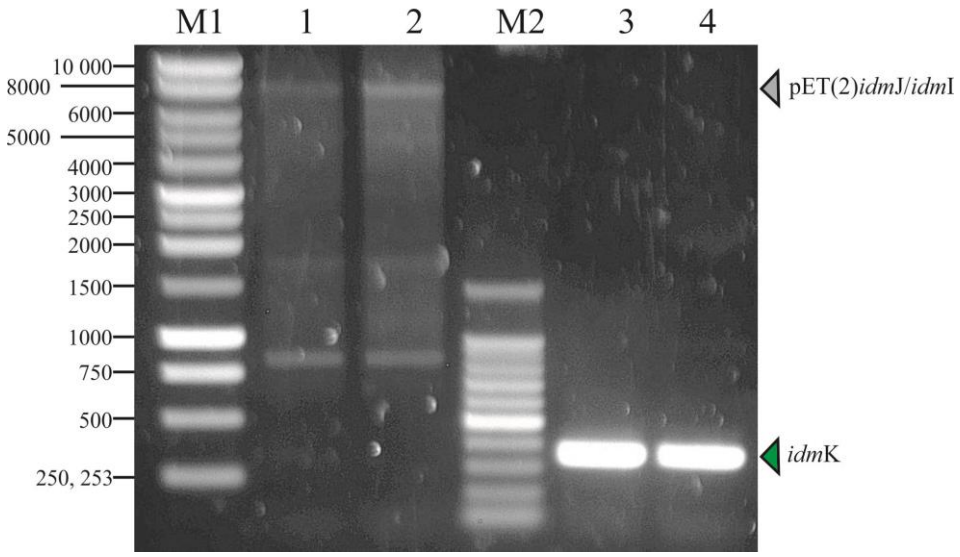


Figure 3.32- PCR amplification of pET(2)*idm*J/*idm*I and *idm*K. Both sets of PCR reactions were carried out using whole cell lysate of *E. coli* XL10 Gold cells containing either the pET(2)*idm*J/*idm*I or pUC*idm*K plasmids. M1 is 1 kb marker and M2 is 100 bp marker. PCR reactions in lanes 1 and 2 are pET(2)*idm*J/*idm*I using 10 and 50 × dilution of the starter culture of . *coli* XL10 Gold cells containing the pET(2)*idm*J/*idm*I plasmid. Presence of a band in both conditions at approximately 8.2 kb corresponds successful amplification of the pET(2)*idm*J/*idm*I plasmid. PCR reactions in lane 3 and 4 shows the successful amplification of *idm*K from 10 and 50 × dilution of the starter cultures of *E. coli* XL10 Gold cells containing the pUC*idm*K plasmid. The band at approximately 350 bp corresponds to *idm*K.

The agarose gel in Figure 3.32 shows the successful amplification of *idm*K (lanes 3 and 4) and pET(2)*idm*J/*idm*I (lanes 1 and 2). *idm*K and pET(2)*idm*J/*idm*I were mixed in a 1:1 volumetric ratio and digested with *Dpn*I for 1 hour. *Dpn*I digests the template DNA from the reaction but is also said to possess a small amount of non-specific exonuclease activity and randomly removes a few DNA bases off the ends of the insert and vector DNA, thus creating complimentary sticky ends. The *Dpn*I digested DNA was then transformed into *E. coli* XL10 Gold cells. Subsequent colony growth on 2 × TY agar plates containing ampicillin indicated the presence of the pET(2)*idm*J/*idm*K/*idm*I plasmid. Restriction digests were performed with *Sac*I-HF® and *Xho*I. Digestion of the pET(2)*idm*J/*idm*K/*idm*I plasmid with these restriction enzymes expected to yield three DNA fragments, one 1.8 kb in length corresponding to *idm*J and *idm*K as one fragment, one 1.2 kb in length corresponding to *idm*I and finally a band at 5.5 kb corresponding to pETDUET. Figure 3.33 below shows the restriction digest of the pET(2)*idm*J/*idm*K/*idm*I plasmids purified from a colony from the transformation of *Dpn*I digested DNA.

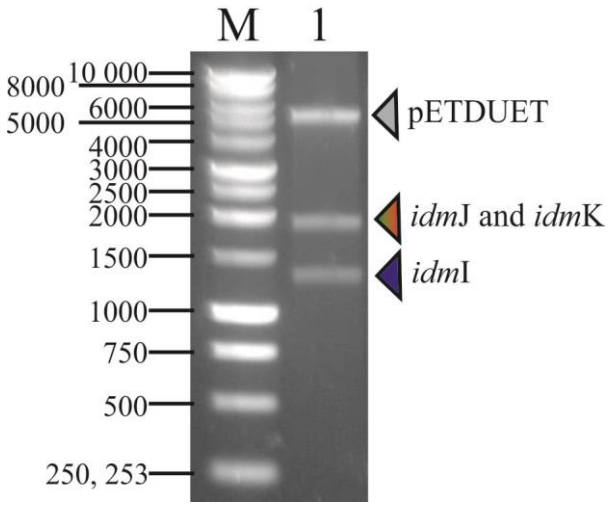


Figure 3.33-0.7 % agarose gel of the pET(2)*idm*J/*idm*K/*idm*I plasmid digested with SacI-HF® and XhoI. The pET(2)*idm*J/*idm*K/*idm*I clone yielded three DNA fragments, one at 5.5 kb, a second at 1.8 kb and a third at 1.2 kb corresponding to pETDUET, *idm*J and *idm*K in one fragment and *idm*I respectively.

Figure 3.32 illustrates the potential pET(2)*idm*J/*idm*K/*idm*I plasmid created using the FastCloning strategy. This plasmid was subsequently transformed into *E. coli* XL10 Gold cells and the plasmid sequenced with the "pETUPSTREAM" and "DUETDOWN1" sequencing primers. This confirmed the presence of *idm*K in the pET(2)*idm*J/*idm*K/*idm*I plasmid from Figure 3.32. The next stage was to attempt protein expression.

3.3.7 Recombinant tandem expression of IdmJ, IdmI and IdmK

The pET(2)*idm*J/*idm*K/*idm*I plasmid was transformed into *E. coli* BL21 (DE3) Gold cells for the purposes of protein expression. Small scale protein expression trials were carried out, testing soluble protein expression in TB, 2 × TY, LB and autoinduction medium at 37°C, 21°C and 18°C as described for IdmJ in section 3.2.4. The soluble cellular fractions for *E. coli* BL21 (DE3) Gold cells containing the pET(2)*idm*J/*idm*K/*idm*I plasmid were analysed by SDS-PAGE (Figure 3.33). The expected masses of IdmJ, IdmI and IdmK are approximately 53 kDa, 41 kDa and 11 kDa respectively, indicated by the red, blue and green arrows in Figure 3.34.

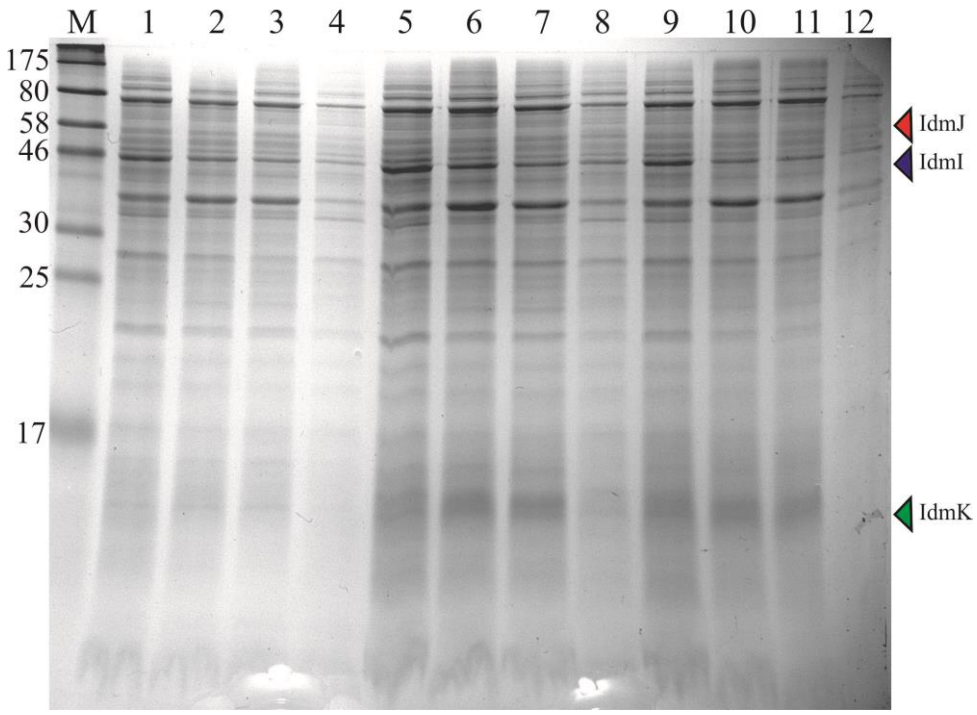


Figure 3.34- Reducing SDS-PAGE gel analysing soluble cell fractions from IdmJ, IdmI and IdmK tandem expression. Tandem expression of IdmJ, IdmI and IdmK was attempted by culturing *E. coli* BL21 (DE3) Gold cells containing the pET(2)*idm*J/*idm*K/*idm*K plasmid in TB medium at 37 °C, 21 °C, 18 °C (lanes 1, 5 and 9 respectively), 2 ×TY at 37 °C, 21 °C, 18 °C (lanes 2, 6 and 10 respectively), LB at 37 °C, 21 °C, 18 °C (lanes 3, 7 and 11 respectively) and autoinduction medium at 37 °C, 21 °C, 18 °C (lanes 4, 8 and 12 respectively). Expected molecular masses for IdmJ, IdmI and IdmK are highlighted; IdmJ has a molecular mass of approximately 53 kDa (red arrow), IdmI has a molecular mass of approximately 41 kDa (blue arrow) and IdmK has a mass of approximately 11 kDa (green arrow).

Analysis of the SDS-PAGE gel (Figure 3.34) of soluble cell fractions from the IdmJ/IdmK/IdmI tandem expression trials, in comparison to uninduced *E. coli* BL21 (DE3) Gold cells (See appendix), suggests that IdmK may have been expressed as soluble protein in all media barring autoinduction medium and at all three expression temperatures. It is unclear as to whether IdmI and IdmJ had been overexpressed in any of these conditions due to a lack of obvious protein bands at the expected masses.

In order to probe the tandem expression of IdmJ and IdmI with IdmK a western blot was performed (section 2.2.3.3). Figure 3.34 shows that IdmK was expressed as soluble protein in all conditions apart from autoinduction medium. Probing for the His_6-tag in the soluble fraction may well give a positive result solely due to the

presence of IdmK and will provide no further information as to whether IdmJ or IdmI were being expressed as soluble proteins. A western blot enabled discrimination of which proteins were being expressed as soluble proteins as they are separated by size on an SDS-PAGE gel. Figure 3.35 shows the results of the western blot performed of expression of pET(2)*idm*J/*idm*K/*idm*I plasmid in TB and LB medium at 21°C, which indicated there was soluble protein expression in a prior His_6-tag screen.

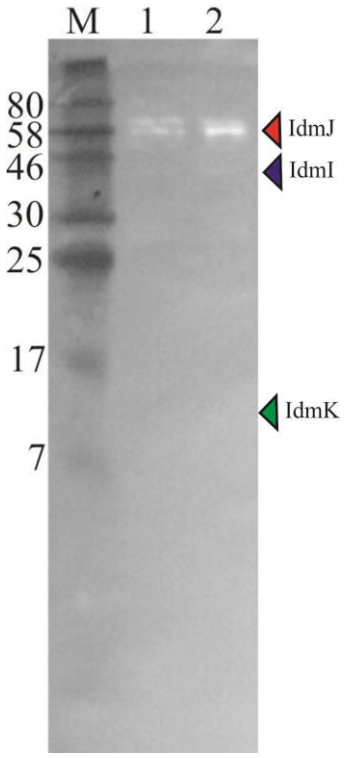


Figure 3.35- Western blot probing for soluble protein expression. A prior dot blot was performed to identify soluble protein expression. Samples that indicated the presence of soluble protein were run on an SDS-PAGE, and a western blot performed probing for the His_6-tag. Lane 1 is the pET(2)*idm*J/*idm*K/*idm*I plasmid expressed at 21°C in TB medium. Lane 2 is the pET(2)*idm*J/*idm*K/*idm*I plasmid expressed at 21°C in LB medium. The contrast of the western blot was altered for ease of viewing and superimposed on the ladder for reference.

The western blot (Figure 3.35) showed that when the pET(2)*idm*J/*idm*K/*idm*I plasmid was expressed in *E. coli* BL21 (DE3) Gold cells, soluble protein was expressed at 21°C in TB (Figure 3.35 Lane 1) and LB (Figure 3.35 Lane 2) medium. However, the approximate size of the protein expressed was comparable to that seen when expressing the pET(2)*idm*J plasmid, indicating that only soluble IdmJ was expressed (SDS-PAGE in Figure 3.13). This was also true for expression the pET(2)*idm*J/*idm*K/*idm*I plasmid at 18°C.

3.3.8 Summary and conclusions

After initial difficulties in cloning *idm*I into pET23a, the gene was successfully cloned into pKK223-3. Expression trials of the pK*idm*I plasmid however did not yield any soluble protein (Figure 3.25). In order to express soluble protein *idm*I was then cloned into a pMAL-c5X vector to create a pMAL*idm*I construct, expressing IdmI as a fusion construct with the maltose binding protein (MBP). Again, expression of a soluble fusion protein was not achieved. Probing for the His_6-tag indicated that the expression of IdmI may be somewhat temperature dependent (Figure 3.28). Tandem expression of IdmI with IdmJ and IdmK from the pET(2)*idm*J/*idm*K/*idm*I plasmid did not yield soluble IdmI either (Figure 3.35). The fact that protein is continually expressed as an insoluble aggregate may indicate that IdmI is not being folded correctly. This may be due to *E. coli*'s inability to form disulphide bonds in the cytoplasm. A cell line like Origami2 (DE3) could be used in attempts to express soluble protein (Bessette *et al.*, 1999). An additional option includes unfolding insoluble aggregates in 8 M urea followed by refolding to resolubilise IdmI. The latter of these two options will require screening for a refolding buffer and optimisation. No biochemical or structural characterisation could be carried out as no soluble protein was purified.

3.4 Discussion

Attempts at heterologous expression of the proteins responsible for starter unit construction in indanomycin biosynthesis began with cloning and expression of IdmJ and IdmI, the adenylyltransferase and dehydrogenase. IdmJ was, to some extent, successfully expressed heterologously, but contains an unidentified post-translational modification. IdmI was unable to be successfully, heterologously expressed in *E. coli* as a soluble protein.

Figure 3.20 shows a schematic to explain what may be occurring when expressing IdmJ and the IdmJ variants. It suggests that mutating C127A could be creating an active adenylyltransferase that can load phosphopantetheine, which may be present in the cell, with proline. Mechanistic determination of a PKS module by Whicher *et al.* (2014) showed that for successful chain elongation and β-keto processing, specific protein-protein interactions with the carrier protein and corresponding

enzyme within the module was required. This may indicate that C127A was able to load phosphopantetheine with proline but unable to release the subsequent product as the system lacked the cognate carrier protein. Further experimental evidence for this theory could be provided using traditional feeding as carried out with PKS, using radiolabelled proline (Birch *et al.*, 1955; Day and Mantle, 1982; Staunton and Weissman, 2001). Optimisation and purification of cysteine variants, especially C127A, may also allow biochemical characterisation, and determine if the variants were expressing as fully folded and functional enzymes.

As mentioned previously, other enzymes required for pyrrole formation in the same manner have been reconstituted in *E. coli*. The genes of pyrrole forming enzymes in clorobiocin and coumermycin biosynthesis were obtained by PCR amplification from cosmids containing genomic DNA, cloned into pET vectors and expressed from *E. coli* BL21 (DE3) cells (Garneau *et al.*, 2005). It is interesting to note that expression and MS analysis of these proteins also showed the +178 Da species seen when expressing IdmJ (Garneau *et al.*, 2005; Geoghegan *et al.*, 1999).

There are a number of strategies still available to produce soluble IdmI. One such strategy includes obtaining the gene from genomic DNA and introducing a C-terminal His_6-tag as with the dehydrogenase found in clorobiocin biosynthesis. Expression could be carried out in *E. coli* cells which aid in disulphide bond formation such as *E. coli* Origami (DE3) cells. Finally, another option is to unfold the IdmI insoluble aggregate and screen for a refolding condition. Biochemical and structural characterisation using techniques such as circular dichroism may provide evidence a fully folded functional enzyme has been obtained. Formation of a pyrrole by an NRPS in an NRPS/PKS hybrid enzyme, specifically as a starter unit, may require the rest of the enzyme, the downstream PKS modules, to dock onto and fold. This may have been why difficulties in soluble expression of individual domains were encountered. Reconstitution of the starter module with a downstream PKS module may provide evidence for this.

The next chapter in this thesis will cover cloning and expression of the third protein in this module, the prolyl carrier protein, IdmK.

4. Cloning, expression and characterisation of the peptidyl carrier protein from the starter module in indanomycin biosynthesis

As described in the introduction (section 1.4), nonribosomal peptide synthetases (NRPS) are large multimodular proteins, with a minimal module being composed of an adenylation domain, condensation domain and finally a peptidyl carrier protein. Within the NRPS modules, the carrier protein (CP), a non-catalytic protein, is essential for function. The phosphopantetheinylated CP houses the growing nonribosomal peptide (NRP) chain, receiving it from the CP within the preceding module. The CP then "passes" the extending chain to the CP in the following the module. The carrier proteins are post-translationally modified by an auxiliary enzyme, a phosphopantetheinyl transferase (PPTase), which modifies a conserved serine, adding a 4'-phosphopantetheine, providing a thiol for attachment of the extending chain (Lambalot *et al.*, 1996). Chain elongation occurs in an assembly-line like fashion with each module responsible for adding a single unit to the extending chain. The extending unit selected by the AT/A domain residing on the ACP/PCP undergoes condensation with the extending chain catalysed by the KS/C domain (Meier and Burkart, 2009; Fischbach and Walsh, 2006).

The prolyl carrier protein (IdmK) is the post-translationally modified non-catalytic essential domain in the starter module of the indanomycin biosynthetic gene cluster. After receiving the activated L-proline from the adenylyltransferase (IdmJ) (section 3.2) the IdmK bound L-proline becomes a substrate for further modification by IdmI (section 3.3) before the starter unit, pyrrole-2-carboxylate, is handed onto the next module in the NRPS-PKS multienzyme (Li *et al.*, 2009). Figure 3.1 shows the role of IdmK in starter unit biosynthesis in the indanomycin biosynthetic gene cluster.

In a number of previous studies, various NRPS and PKS carrier proteins have been recombinantly expressed in *E. coli*, not only as individual domains but also as part of reconstituted modules and extensive structural studies have been carried out (Crosby and Crump, 2012; Chan and Vogel, 2010; Meier and Burkart, 2009). Studies have been conducted attempting to phosphopantetheinylate the conserved serine with

exogenous PPTases, exploring not only the carrier protein but the PPTases tolerance for modification (Beld *et al.*, 2014).

The carrier protein, a small protein of less than 100 amino acids, is an essential part of the modular enzymes. The carrier protein must have the ability to "communicate" with other domains within its own module to receive substrates and with domains in adjacent modules in order to pass and receive the extending chain (Tran *et al.*, 2010; Weissman and Muller, 2008; Crosby and Crump, 2012). Structural and dynamic information including protein-protein interactions could provide insight into how this protein functions, in the hope of engineering it for the production of novel compounds. Figure 1.10 shows the proposed mechanism (from cryo-EM data) by which acyl carrier proteins interact with each domain involved in substrate selection, catalysis and modification meting chain elongation (Dutta *et al.*, 2014; Whicher *et al.*, 2014). These mechanistic findings show that chain elongation is very much carrier protein centric, with the carrier protein being essential for chain elongation and modification, and further emphasises the importance in understanding a carrier protein and its interactions with other domains within its own, and subsequent modules (Dutta *et al.*, 2014; Whicher *et al.*, 2014; Nguyen *et al.*, 2014).

Cloning, protein expression and purification followed by structural characterisation of IdmK will hopefully enable mechanistic insights into how chain elongation occurs during indanomycin biosynthesis as well as allowing the fundamental protein-protein interactions of the carrier protein to be determined.

4.1 Cloning the peptidyl carrier protein (*idm*K)

4.1.1 Cloning the peptidyl carrier protein (*idm*K) into pET23a

The synthetic *idm*K gene was designed with *Hind*III and *Xho*I sites at the 5' and 3' ends of the gene, respectively. As with the two previous genes (Chapter 3), the sites were selected specifically for sub-cloning idmK into pET23a using the cut-and-paste method (as illustrated in Figure 3.3). Insert DNA was prepared in an identical manner to *idm*J and *idm*I however the restriction enzymes *Hind*III-HF® and *Xho*I were used to digest the pUC*idm*K plasmid. The pET23a vector was digested with the same restriction enzymes, creating complimentary sticky ends. Vector and insert

DNA were then run on a 0.7 % agarose gel and purified as described in section 2.2.2.3. Ligation reactions between pET23a and *idm*K were performed as described in section 2.2.2.8. 4 µL of this ligation reaction was then transformed into *E.coli* XL10 Gold Ultracompetent cells and cells were subsequently plated onto 2 × TY agar plates containing ampicillin. Colony growth on the transformation plates suggested the presence of the desired plasmid, pET*idm*K. To screen for the correct plasmid, DNA from 10 colonies was purified and digested with *Hind*III-HF® and *Xho*I restriction enzymes (section 2.2.2.7). Resulting constructs were analysed by running the digested DNA on an agarose gel. Figure 4.1 shows the results of the restriction digests of plasmid DNA purified from the 10 colonies selected from the ligation between pET23a and *idm*K.

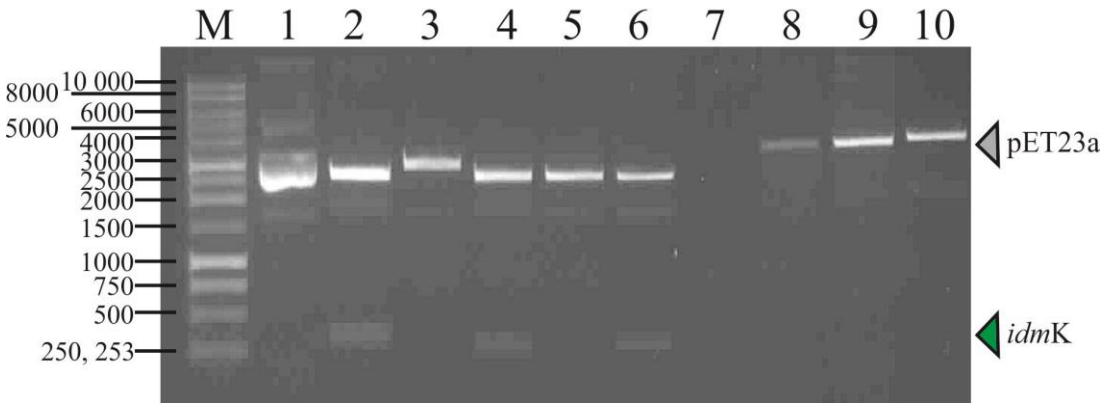


Figure 4.1-0.7 % agarose gel of isolated plasmid DNA from the 10 colonies selected from the transformation plate of the ligation reaction between pET23a and *idm*K. Plasmid DNA was digested with *Hind*III-HF® and *Xho*I to screen for the pET*idm*K plasmid. The grey arrow indicates the expected size of the pET23a vector (3666 bp) and the green arrow indicates the expected size of *idm*K (325 bp). Plasmid DNA in lane 1 appears to remain undigested plasmid DNA. In lanes 2, 4 and 6 plasmid DNA digested with *Hind*III-HF® and *Xho*I yielded two DNA fragments at approximately 2700 bp and 325 bp corresponding to the original vector pUC57 and *idm*K. Plasmid DNA in lane 3 digested with *Hind*III-HF® and *Xho*I yielded a single band at approximately 3000 bp, this corresponds to the pUC*idm*K plasmid digested at a single point. Plasmid DNA in lane 5 appears to be 2700 bp corresponding to empty pUC57. The plasmid purification in lane 7 appears to have been unsuccessful. Finally plasmid DNA in lanes 8, 9 and 10 digested with HindIII-HF® and XhoI show a single band at approximately the correct size for pET23a alone (3666 bp).

A pET*idm*K plasmid digested with *Hind*III and *Xho*I should yield two bands on a DNA gel; a band at 3666 bp corresponding to pET23a and a second band at 325 bp corresponding to *idm*K. Figure 4.1 indicates the ligation between pET23a and *idm*K

was unsuccessful. Plasmid DNA in lane 1 appears to have remained undigested by *Hind*III-HF® and *Xho*I, indicating it is not the pET*idm*K plasmid. Plasmid DNA in lanes 2, 4 and 6 show DNA fragments 2700 bp and 325 bp in length corresponding to pUC57 and *idm*K, this is indicative of pUC*idm*K contamination of the ligation reaction. Plasmid DNA in lanes 3, 5, 8, 9 and 10 digested with *Hind*III-HF® and *Xho*I show single DNA fragments of varying sizes, none of these plasmids appear to contain the expected 325 bp insert corresponding to *idm*K. Further attempts were made to clone idmK into pET23a however this approach was not successful. The next strategy was to clone IdmK into pKK223-3 plasmid as attempted with idmJ and idmI.

4.1.2 Cloning the peptidyl carrier protein (*idm*K) into pKK223-3

As with *idm*J and *idm*I the next vector selected for sub-cloning *idm*K was pKK223-3 (as illustrated in Figure 3.6), creating a plasmid designated pK*idm*K. To prepare the insert, restriction digests were carried out with the pUC*idm*K plasmid and pK*nan*A plasmid using *Eco*RI-HF® and *Pst*I-HF® creating complimentary sticky ends for ligation (section 2.2.2.7). Digested plasmids were run on a 0.7 % agarose gel and insert and vector DNA extracted from the gel (section 2.2.2.3). A ligation reaction was then carried out between pKK223-3 and *idm*K digested with *Eco*RI-HF® and *Pst*I-HF®. 4 µL of the ligation reaction was transformed into *E. coli* XL10 Gold Ultracompetent cells. Successful ligation of *idm*K and pKK223-3 was indicated by cell growth on 2 × TY agar plates containing ampicillin. 10 colonies were selected from the transformation plates, plasmid DNA purified, and the pK*idm*K plasmid screened by restriction digestion with *Eco*RI-HF® and *Pst*I-HF®. A pK*idm*K plasmid digested with *Eco*RI-HF® and *Pst*I-HF® restriction enzymes should yield two DNA fragments, one at 4600 bp and a second at 325 bp corresponding to pKK223-3 and *idm*K respectively. Figure 4.2 below shows the results from the restriction digests screening for the pK*idm*K plasmid.

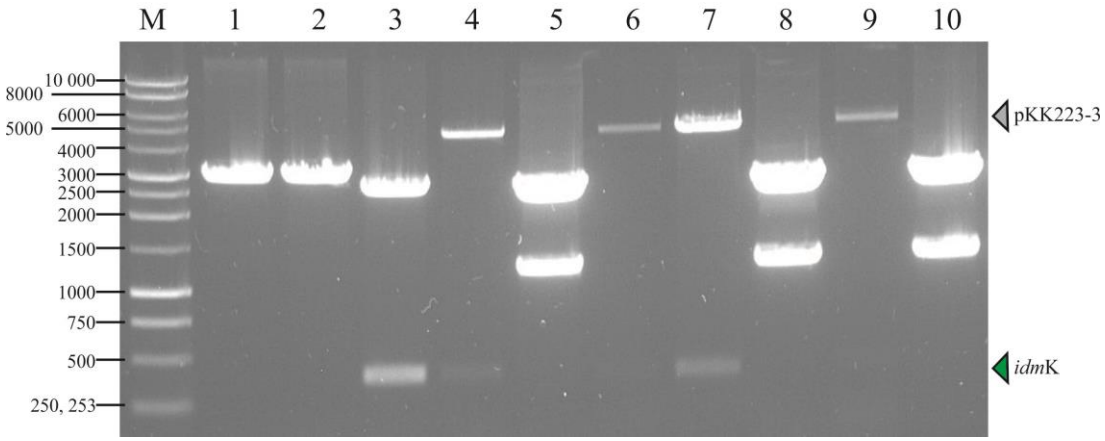


Figure 4.2-0.7 % agarose gel showing restriction digests of purified plasmid DNA from the pKK223-3 *idm*K ligation. Plasmid DNA was purified from 10 colonies and digested with EcoRI-HF® and *Pst*I-HF® to screen for the pK*idm*K plasmid. Plasmid DNA in lanes 1 and 2 digested with *Eco*RI-HF® and *Pst*I-HF® have yielded a single band approximately 3000 bp in length, corresponding to incomplete digestion of pUC*idm*K. Digestion of plasmid DNA in lane 3 yielded two DNA fragments, one 2700 bp in length corresponding to pUC57 and a second 325 bp corresponding to *idm*K. Plasmid DNA digested by *Eco*RI-HF® and *Pst*I-HF® in lanes 4 and 7 both yielded DNA fragments of approximately 4600 bp and 325 bp corresponding to pKK223-3 and *idm*K. Digestion of plasmid DNA in lane 5, 8 and 10 yielded two DNA fragments, one approximately 2700 bp (pUC57) and an DNA fragment approximately 1200 bp of unknown origin. Finally digestion of plasmid DNA in lanes 6 and 9 yielded a single band at approximately 4600 bp corresponding to pKK223-3.

Restriction digestion of plasmid DNA in lanes 4 and 9 in Figure 4.2 shows successful ligation between *idm*K and pKK223-3, creating the pK*idm*K plasmid, exhibiting DNA bands 4.6 kb and 325 bp in length corresponding to the correct sizes for the pKK223-3 vector and *idm*K insert respectively. A glycerol stock was made of *E. coli* XL10 Gold Ultracompetent cells transformed with the plasmid from lane 4. Confirmation of the pK*idm*K plasmid was obtained by DNA sequencing of plasmid 4 with the sequencing primers "PTRC-99A-FOR" and "PTRC-99A-REV" (sequences can be located in the appendix). Sequencing showed that the pKK223-3 vector contained the correct insert. The sequenced pK*idm*K plasmid was then transformed into *E. coli* BL21 Gold (DE3) cells for protein expression.

4.2 Expression of the prolyl carrier protein (IdmK)

Expression of the prolyl carrier protein was carried out using *E.coli* BL21 Gold (DE3) cells transformed with the pK*idm*K plasmid. A small scale protein expression

of the putative *S. antibioticus* prolyl carrier protein (IdmK) was tested on a 5 mL scale, in 2 × TY medium at 37°C as with IdmJ (section 3.2.4). Soluble and insoluble cellular components of lysed cells were analysed by SDS-PAGE in order to evaluate whether IdmK was expressed as a soluble protein (Figure 4.3).

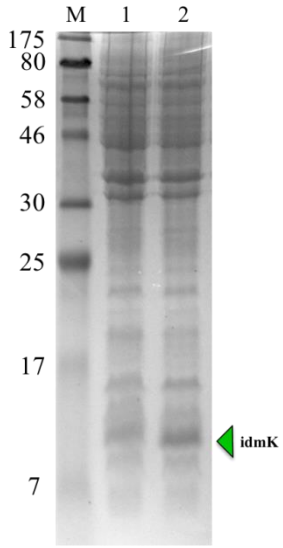


Figure 4.3- Reducing SDS-PAGE analysis of insoluble (lane 1) and soluble (lane 2) fractions suggested that IdmK was successfully expressed as soluble protein (indicated by the green arrow) in 2 × TY media at 37 °C.

The expected mass of IdmK, as calculated from the amino acid sequence of the synthetic gene, is approximately 11044 Da, as indicated by the green arrow in Figure 4.3. The SDS-PAGE gel in Figure 4.3 suggests that IdmK has been expressed as a soluble protein, indicated by the presence of a band at the expected mass in the soluble fraction of the *E. coli* BL21 (DE3) Gold cells transformed with the pK*idm*K plasmid (lane 2). This protein band is absent in the insoluble fraction providing validation that IdmK is expressed as a soluble protein.

4.3 Purification of the prolyl carrier protein (IdmK)

Protein expression was repeated on a larger (1 L) scale. IdmK was then purified by nickel affinity chromatography as described in section 2.2.3.4. Samples were taken at each stage of the purification and analysed by SDS-PAGE (Figure 4.4).

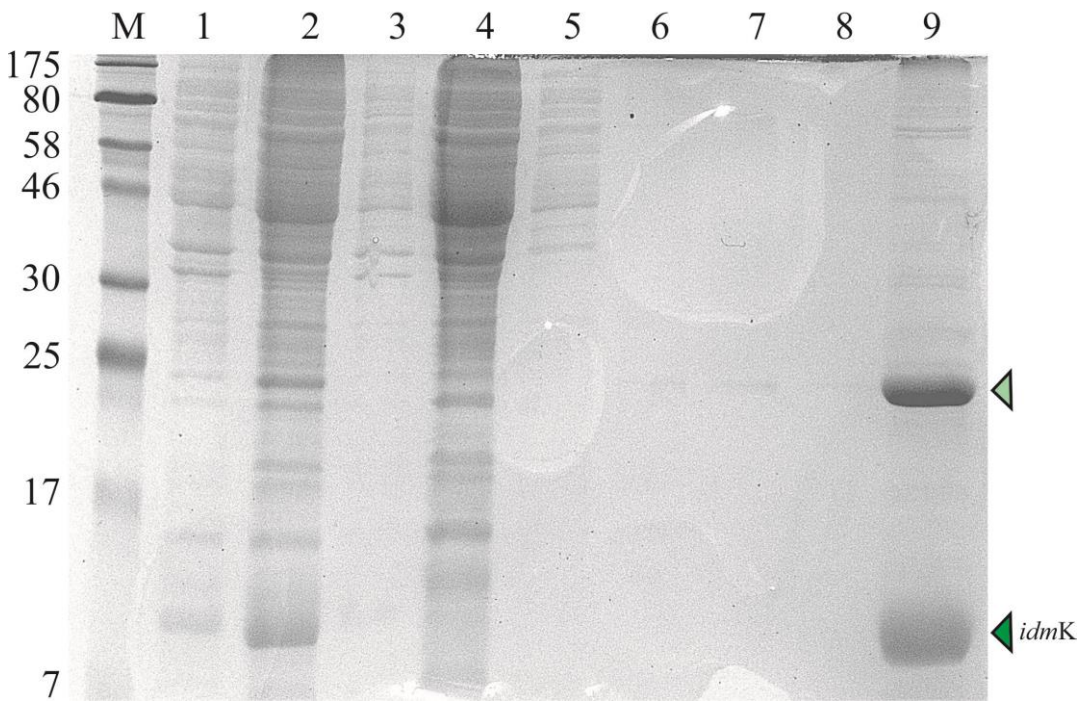


Figure 4.4- Reducing SDS-PAGE analysis of the purification of IdmK by nickel affinity chromatography. IdmK was expressed in *E. coli* BL21 (DE3) gold cells induced by IPTG (1), cells were harvested, lysed and the soluble (2) and insoluble (3) fractions separated by centrifugation. The soluble supernatant was then incubated with nickel resin for one hour, the supernatant (4) was discarded and the resin was washed with a low imidazole buffer (20 mM) (5-8) to remove non-specifically bound proteins. IdmK was then eluted from the resin using a high imidazole elution buffer (0.5 M) (9). Two dominant protein species were eluted from the resin, one approximately 11 kDa (dark green arrow) possibly corresponding to IdmK, the second (light green arrow) with an approximate mass of 22 kDa.

Figure 4.4 (lane 9) suggests that IdmK has successfully been purified using nickel affinity chromatography although two bands were seen. To provide further evidence IdmK had been successfully purified and to determine more accurate masses of the two species, a 70 µL sample of eluted protein was desalted into 50 mM ammonium acetate pH 7.4 (section 2.2.3.9) for analysis by ESI-MS (section 2.2.3.12). Figure 4.5 shows the results of ESI-MS analysis of the purification of IdmK.

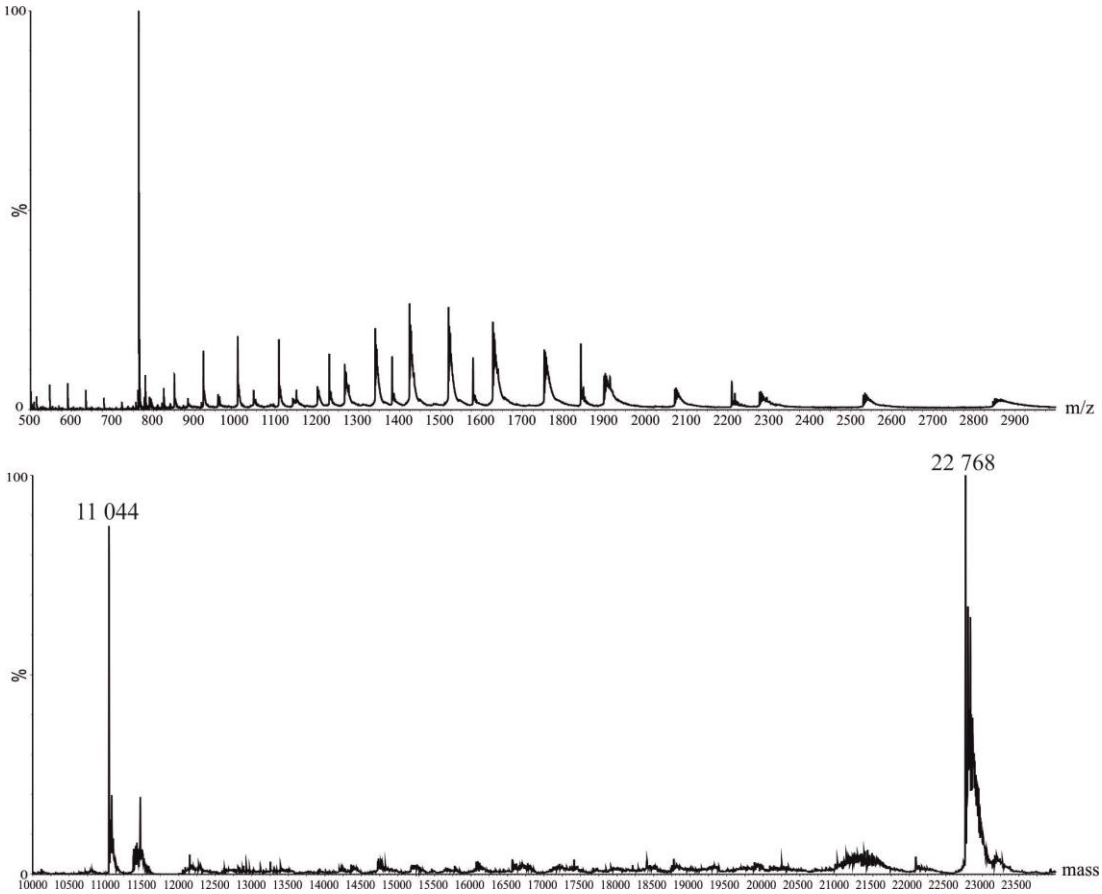


Figure 4.5- ESI-MS of the eluted protein from the purification of IdmK. A sample of eluted protein (Figure 4.4, lane 9) was desalted into 50 mM ammonium acetate pH 7.4 and analysed by ESI-MS to obtain a molecular mass of the species present. (Top) Mass spectrum raw data; (bottom) deconvoluted mass spectrum of raw data showing the masses of proteins purified. Two species were detected, one at 11 044 Da, and a second at 22 768 Da.

The expected mass of IdmK as calculated from the amino acid sequence of the synthetic gene is 11045 Da (Table 4.1). This mass of *apo*-IdmK does not include post-translational modification of the carrier protein by phosphopantetheinylation. After modification the mass expected for phosphopantetheinylated IdmK, *holo*-IdmK, should be 11 386 Da (Table 4.1). The mass spectrum (Figure 4.5) shows that *apo*-IdmK, with a mass of 11 044 Da, was purified as one of the species, while the mass of the second protein species eluted during purification (Figure 4.4 lane 9) (22 768 Da) corresponds to the mass that would be expected for dimeric *holo*-IdmK. A possibility is that dimerisation might occur through the terminal thiol of the phosphopantetheine group. This dimerisation would create a non-functional covalent dimer with an expected mass of 22 770 Da, compared to the 22 768 Da.

Sample	Nomenclature	Expected mass
IdmK	*apo*-IdmK	11 045
Phosphopantetheinylated IdmK	*holo*-IdmK	11 386

Table 4.1- Table showing the expected masses of IdmK as calculated from the amino acid sequence.

In order to test the idea that a covalent dimer was being formed through disulphide bond formation between the terminal thiol groups of the phosphopantetheine modification of *holo*-IdmK the two protein species were further purified and characterised. The protein eluted from the nickel affinity purification was subject to further purification by size exclusion chromatography, thus separating dimeric and monomeric IdmK by size. Figure 4.6 shows the size exclusion chromatography trace for IdmK. SDS-PAGE analysis was used to confirm the separation of the two species.

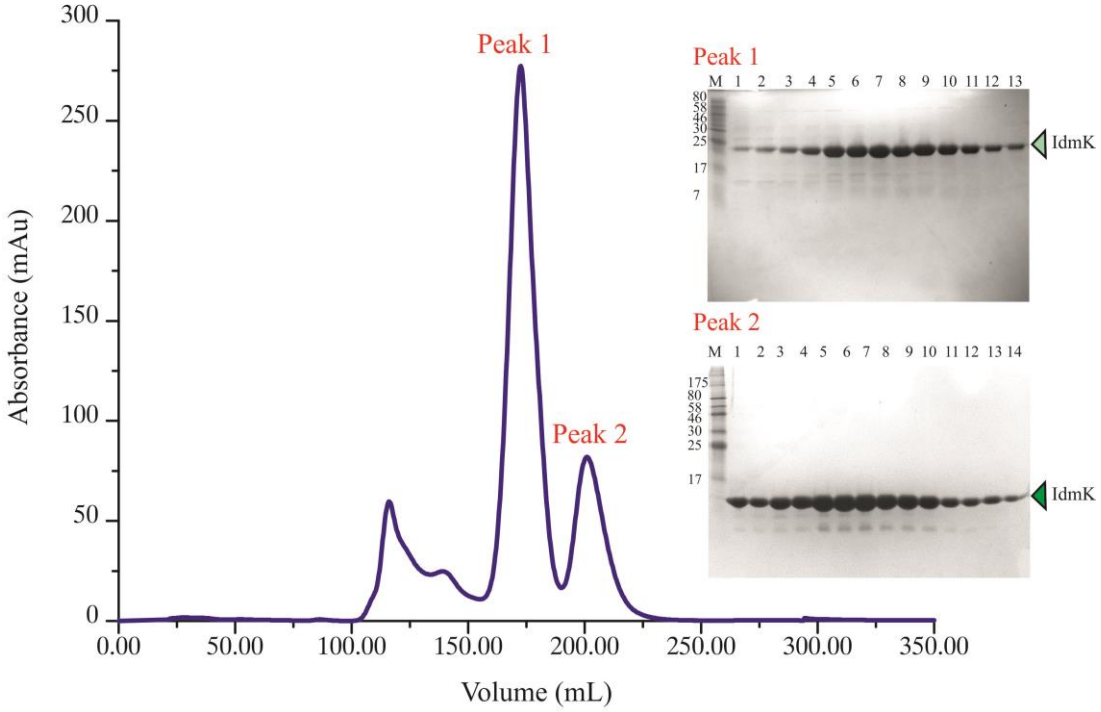


Figure 4.6- Purification of IdmK by size exclusion chromatography. Size exclusion chromatography trace, the Y axis shows the absorbance of the eluent. The X axis shows the volume. Fractions from peak 1 and peak 2 were run on SDS-PAGE gels (inset). The light green arrow indicates the potentially dimeric form of IdmK and the dark green arrow indicates the monomeric form of IdmK.

IdmK was gel filtered using a HiLoad Superdex 75 prep grade (S75) column (Figure 4.6). The dimeric species (Figure 4.6 peak 1) eluted from the S75 column between 160 and 180 mL, while monomeric *apo*-IdmK (Figure 4.6 peak 2) was eluted between 190 mL and 210 mL. Samples from the peak fractions were analysed by SDS-PAGE (Figure 4.6).

To provide evidence that IdmK was assembling as a dimer through the formation a disulphide bind between the terminal thiol of the phosphopantetheine modifications, Tris (2-carboxyethyl) phosphine (TCEP) was used as a reducing agent to break the suspected disulphide bond. A 70 µL sample was taken from peak one of the size exclusion chromatography trace shown in Figure 4.6, indicated to be dimeric IdmK, and was desalted into 50 mM ammonium acetate pH 7.4. TCEP was added to the sample to a final concentration of 5 mM and the reduced protein sample was then analysed by ESI-MS (Figure 4.7).

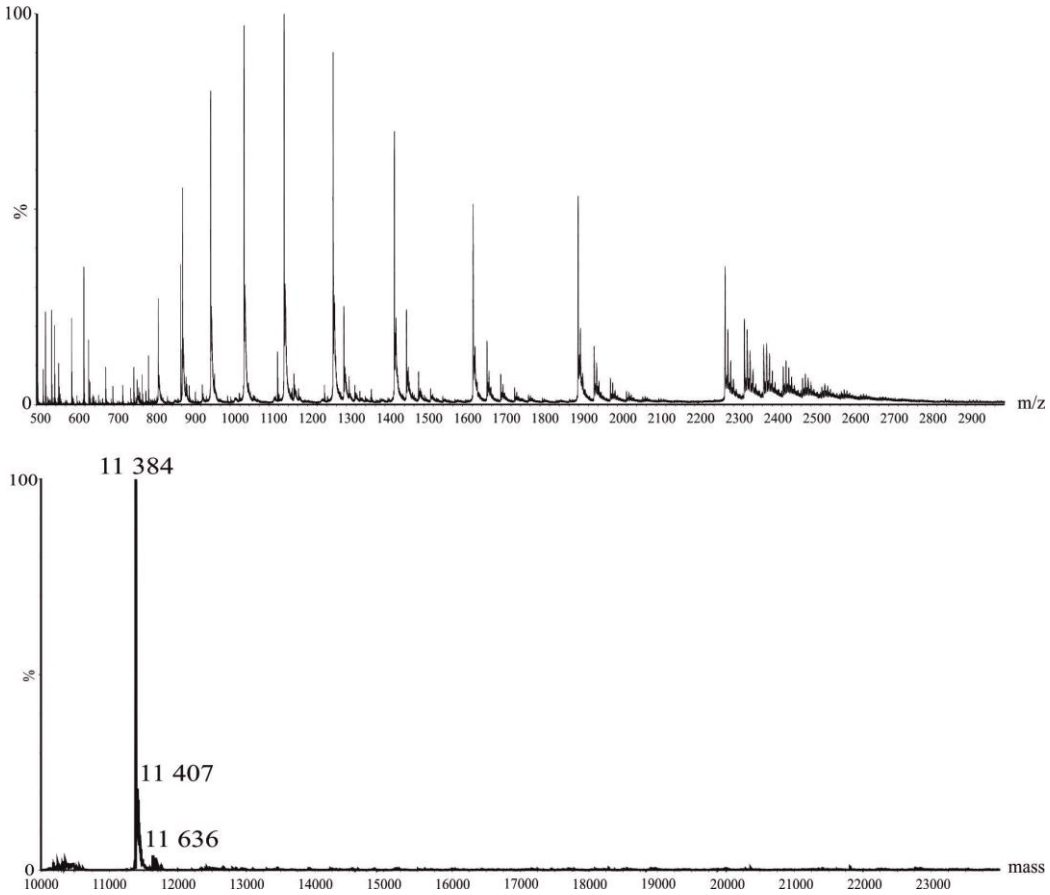


Figure 4.7- ESI-MS analysis of the reduced IdmK dimer. A sample from peak one of size exclusion chromatography was reduced by the addition of TCEP and analysed by ESI-MS. A single major protein species at 11 384 Da was observed, this corresponds to phosphopantetheinylated IdmK (*holo*-IdmK).

The mass spectrum measured after reduction (Figure 4.7) corresponds to the expected mass of the monomeric *holo*-IdmK, with a mass of 11 384 Da (Table 4.1), confirming the possibility that IdmK could be oxidised to form a covalent dimer during the growth, cell breakage or purification. The fortuitous discovery of *holo*-IdmK forming a homodimer can be utilised to purify homogeneous samples of both *apo*- and *holo*- IdmK. Incidentally, this result demonstrated that *E. coli* possesses the ability to phosphopantetheinylate the recombinantly expressed carrier protein.

4.3.1 Confirmation of the site of phosphopantetheinylation

The phosphopantetheinylation of IdmK, or any carrier protein, is required for function. The site of post-translational modification is a highly conserved serine residue (Quadri *et al.*, 1998; Walsh *et al.*, 1997; Staunton and Weissman, 2001). To confirm the presence of the correct post-translational modification, the conserved serine residue in IdmK was identified. A BLAST search followed by a multiple sequence alignment of similar amino acid sequences was performed identifying Ser44 to be conserved (appendix figure 7.3). To confirm that this was the site of phosphopantetheinylation, site-directed mutagenesis was used to mutate the conserved serine 44 to an alanine (S44A), thereby preventing the modification from occurring. To create the S44A variant the pK*idm*K plasmid DNA was purified from *E. coli* XL10 Gold Ultracompetent cells and site-directed mutagenesis performed using the Stratagene QuikChange® Site-Directed Mutagenesis kit (section 2.2.2.10) using the mutagenic primers “S44A For” and “S44A Rev” (sequences located in the appendix). DNA sequencing using the “PTRC-99A-FOR” and “PTRC-99A-REV” sequencing primers confirmed the presence of the targeted serine to alanine mutation (sequences located in the appendix). The pK*idm*K mutant plasmid was transformed into *E. coli* BL21 (DE3) Gold cells for subsequent protein expression. The IdmK variant was expressed and purified in an identical manner to wild-type IdmK and analysed by ESI-MS.

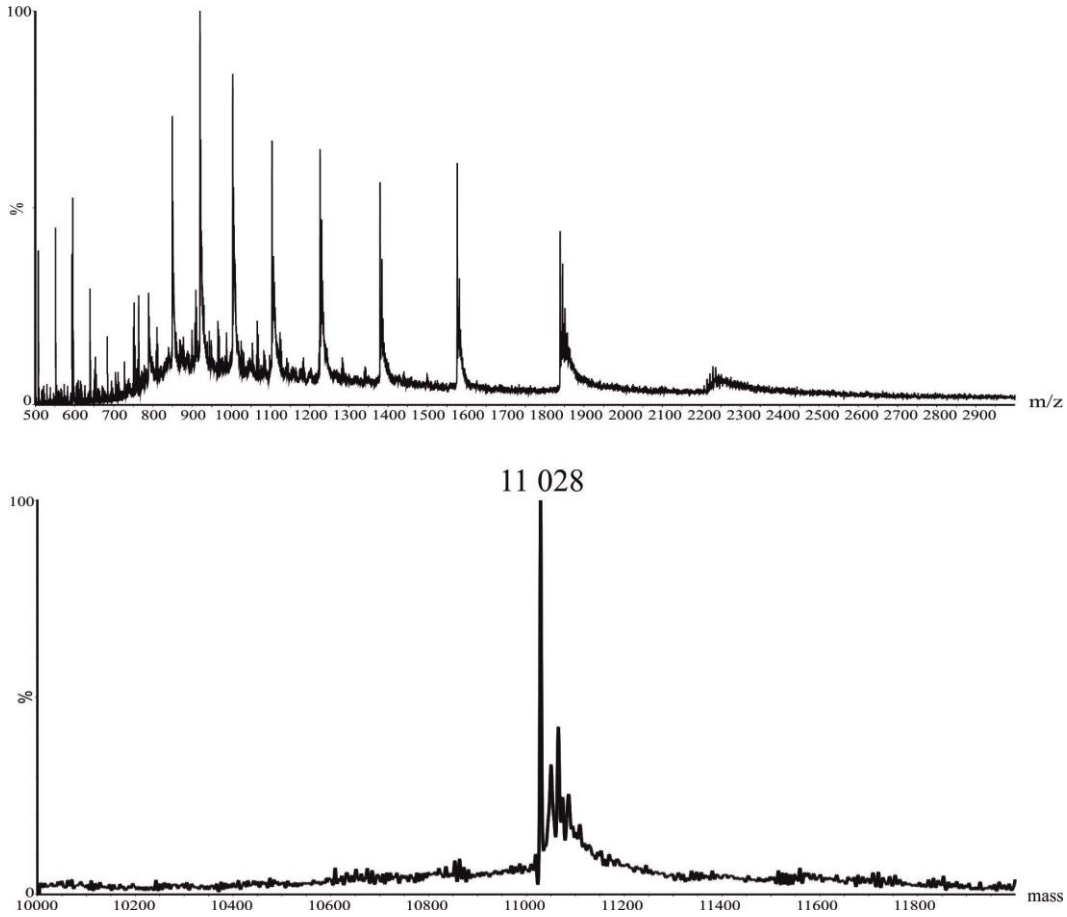


Figure 4.8- ESI-MS analysis of the purified S44A IdmK variant. The IdmK S52A variant was expressed in *E coli* BL21 (DE3) Gold cells in an identical manner to wild-type IdmK, eluted protein was desalted into 50 mM ammonium acetate pH 7.4 and analysed by ESI-MS. (Top) mass spectrum raw data (bottom) deconvoluted mass spectrum of raw data showing the mass of the purified protein.The expected mass of the IdmK variant was 11 028 Da, this was observed, in addition there is another species identified, 57 Da larger than the IdmK variant, the potential adduct being a nickel ion.

Figure 4.8 above shows the mass spectrum of the purified S44A IdmK variant. Only a single peak was observed (at 11 028 Da) corresponding to the expected mass of the *apo*-S44A variant. The lack of any covalent dimeric species along with the lack of any phosphopantetheinylated species confirmed the site of post-translational modification to be Ser44.

4.4 Structural characterisation of the prolyl carrier protein (IdmK)

Prior to this work a number of carrier protein structures from FAS, PKS and NRPSs existed in the protein data bank (PDB) the structures of which were resolved by a number of biophysical techniques. Table 4.2 shows examples of carrier proteins with structures resolved deposited in the PDB.

PDB code	Structure	Method	Reference
2JU1	*Apo*-ACP from module 2 of 6-deoxyerythronalide B	NMR spectroscopy	Alekseyev *et al.*, 2007
1T8K	*Apo*-ACP from FAS in *E. coli*	X-Ray crystallography (Multi-wavelength anomalous dispersion)	Qiu and Janson, 2004
1L0H	Butyryl-ACP from FAS in *E. coli*	X-Ray crystallography (Multi-wavelength anomalous dispersion)	Roujeinikova *et al.*, 2002
1DNY	*Apo*-PCP from module 3 of tyrocidine synthetase	NMR spectroscopy	Weber *et al.*, 2000
2LIU	*Holo*-ACP from the Curacin A PKS	NMR spectroscopy	Busche *et al.*, 2012
3GZM	*Holo*-ACP from FAS in *Plasmodium Falciparum*	X-Ray crystallography (molecular replacement)	Gallagher and Prigge, 2010
2LPK/2LL8	*Holo/apo*-ACP from FAS in *Rhodopseudomonas palustris*	NMR spectroscopy	Ramelot *et al.*, 2012
1AF8	*Apo*-ACP from the actinorhodin PKS in *S. coelicolor* A3(2)	NMR spectroscopy	Crump *et al.*, 1997
1NQ4	*Apo*-ACP from oxytetracycline PKS	NMR spectroscopy	Findlow *et al.*, 2003
2PNG	*Apo*-ACP from rat FAS	NMR spectroscopy	Ploskon *et al.*, 2008
1HY8	*Apo/Holo*-ACP from FAS in *B. subtilis*	NMR spectroscopy	Xu *et al.*, 2001

Table 4.2- Examples of carrier protein structures found in the PDB.

A comprehensive review of carrier protein structures can be found in Crosby and Crump (2012).

4.4.1 Sample preparation and screening crystallisation conditions

For structural studies of IdmK homogeneous protein samples were required. Due to the dimerisation that occurs via the terminal thiol of the phosphopantetheine post-translational modification, homogenous samples of both *apo* and *holo* IdmK

could be obtained by subjecting IdmK to further purification by size exclusion chromatography as in Figure 4.6. Fractions collected of *apo*-IdmK (Figure 4.6, peak 2) were pooled and concentrated to between 10 mg/mL for the screening of crystallisation conditions.

Due to the considerable differences between the amino acid sequence of IdmK and the previous carrier proteins available in the PDB, as well as the disparity in crystallisation conditions described, a set of commercially available crystallisation screens were set up (Hampton Research (USA) screens; Crystal screen 1 and 2, Index 1 and 2, Salt RX and from Emerald BioStructure (Germany) Wizard 3 and 4). Crystallisation was attempted by vapour diffusion of 2 µL sitting drops (1:1 protein to mother liquor ratio) in 96 well-plates at 18°C. For crystallisation trials *apo*-IdmK, in 50 mM Tris/HCl pH 7.4, was used at a concentration of 10 mg/mL. A high percentage of drops showed protein precipitation, indicating the protein concentration was close to supersaturation. Nevertheless, 11 initial crystal hits were observed after one week, crystal growth appears to be encouraged in conditions containing PEG 3350 and a higher pH (8.5). Figure 4.9 shows some of the crystal forms observed.

Crystals were cryoprotected in 25 % (v/v) glycerol, flash cooled in liquid nitrogen and screened for diffraction at 100 K on beamline I04-1 at Diamond Light Source Ltd (DLS). High-resolution diffraction data were collected from crystals grown in pH 8.2 0.056 M sodium phosphate monobasic monohydrate, 1.344 M potassium phosphate dibasic. Further optimisation of the crystallisation conditions was attempted by varying protein concentration (6-12 mg/mL), protein to mother liquor ratio (1:1, 1:2 and 2:1) and pH (7.8-8.2). Subsequent crystal trays were set up using these conditions in a hanging drop format. The best crystal growth was observed at pH 8.0 and 8.2 with protein concentrations above 9 mg/mL.

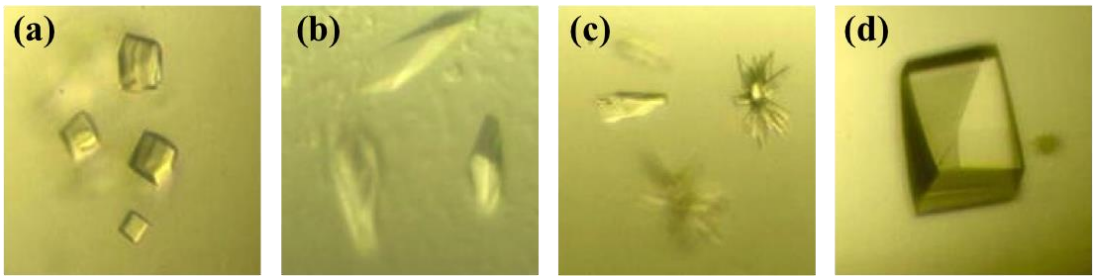


Figure 4.9- The representative crystals in this figure were obtained when crystallisation conditions were screened using *apo*-IdmK in the following conditions; (A) 0.2 M Sodium malonate pH 7.0, 20 % (w/v) PEG 3350, (B) 0.2 M ammonium sulphate, 0.1 M bis-tris pH 5.5, 25 % (w/v) PEG 3350, (C) 0.2 M sodium chloride, 0.1 M Tris/HCl pH 8.5, 25 % (w/v) PEG 3350 and (D) pH 8.2, 0.056 M sodium phosphate monobasic monohydrate, 1.344 M potassium phosphate dibasic.

X-ray data were collected at I04 beamline at DLS. Data collection and processing was carried out by Dr Chi Trinh (Astbury Centre, University of Leeds, UK). The data were reduced using the automated processing suit Xia2 at DLS. Data reduction statistics and unit cell parameters can be located in the appendix. Phasing by molecular replacement (using both Molrep (Vagin and Teplyakov, 1997) and Phaser (Mccoy *et al.*, 2007)) was unsuccessful, mainly due to the large number of monomers present in the asymmetric unit cell. An estimated 7 or 8 monomers were predicted from Matthews co-efficient (Matthews, 1968). Since molecular replacement failed for structural characterisation of IdmK there are number of alternatives that could be considered, including screening for a different crystal form, use anomalous diffraction using selenomethionine labelled protein, or use of another biophysical technique. Structural characterisation of IdmK by X-ray crystallography was not continued.

4.5 Structure determination using NMR spectroscopy

As mentioned in section 4.2 a number of carrier protein structures exist in the PDB, solved by a number of biophysical techniques. Nuclear magnetic resonance has been used to solve a number of *apo*- and *holo*- carrier protein structures including the peptidyl carrier protein from the third module of the tyrocidine biosynthetic gene cluster (Figure 1.16) from *Brevibacillus brevis* (PDB codes 1DNY, 2GDW, 2GDX, 2GDY) (Weber *et al.*, 2000; Koglin *et al.*, 2006) and of the acyl carrier protein from *B. subtilis* (PDB code 1HY8) (Xu *et al.*, 2001). These structures illustrate the

potential for using NMR to structurally characterise IdmK, and potentially make comparisons between the *apo*- and *holo*- protein species. The first step in studying IdmK by NMR was to express and purify isotopically labelled protein.

4.5.1 Initial analysis of the suitability of IdmK for structural studies by NMR spectroscopy

Previous results in this chapter have shown that IdmK had successfully been expressed and purified from *E. coli* BL21 (DE3) Gold cells transformed with the pK*idm*K construct. In order to assess the suitability of the use of NMR for structural studies, expression of IdmK in minimal medium was required. Expression of IdmK in minimal medium was carried out in M9 minimal medium (section 2.1.7) and the standard expression protocol used (section 2.2.1.5 and 2.2.3.4). The inoculation procedure of the minimal medium was augmented to prevent transfer of 2 × TY medium from the day culture: cells from the 10 mL day culture were harvested by centrifugation and resuspended in 2 mL of minimal medium, and this was used to inoculate 500 mL expression cultures. To determine if IdmK could be expressed in minimal medium small-scale expression was trialled in minimal medium in the absence of labelled medium components. To analyse the expression of IdmK, *E.coli* BL21 (DE3) Gold whole cell lysate was run on a SDS-PAGE (Figure 4.10).

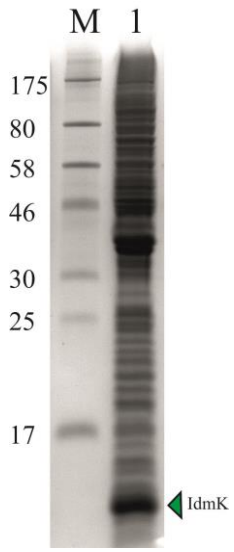


Figure 4.10- Reducing SDS-PAGE analysis of the small scale expression trial of IdmK in minimal media. Lane 1 is the whole cell lysate from *E. coli* BL21 (DE3) Gold cells containing the pK*idm*K plasmid.

As can be seen in the SDS-PAGE gel in Figure 4.10 protein over-expression was observed when expressing IdmK in minimal medium, as indicated by the protein band at approximately 11 kDa (green arrow).

Due to the observed over-expression of IdmK in minimal medium, IdmK was then expressed in minimal medium containing $^{15}NH_4Cl$ as the sole nitrogen source. Protein expression and purification were carried out by nickel affinity chromatography as previously described (section 2.2.3.4). Samples were taken at each stage of the purification and analysed by SDS-PAGE (Figure 4.11).

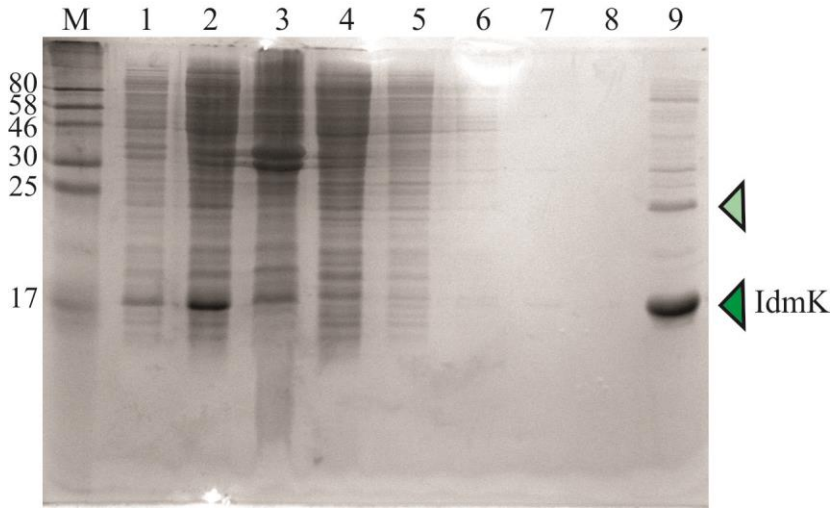


Figure 4.11- Reducing SDS-PAGE gel showing the stages of nickel affinity chromatography of IdmK grown in minimal media containing $^{15}NH_4Cl$. Broad range protein size markers are displayed (M). Cells expressing IdmK (1) were harvested by centrifugation, lysed and the soluble (2) and insoluble (3) fractions separated. Although there is a small band at approximately 11 kDa, indicating some insoluble material, the majority of protein appears to be expressed as soluble protein. After binding the soluble protein to the nickel resin (4), four washes were performed to remove any non-specifically bound protein (5-8). The bound protein was eluted by washing with 0.5 M imidazole, purified ^{15}N labelled IdmK is shown in lane (9). The dark green arrow indicates monomeric *apo*-IdmK and the light green arrow indicates the presence of dimeric *holo*-IdmK.

Figure 4.11 shows the successful purification of IdmK when expressed in minimal medium. As mentioned beforehand, both monomeric *apo*- and dimeric *holo*- species of IdmK are present upon purification by nickel affinity chromatography (Figure 4.11); therefore further purification by size exclusion chromatography was used to obtain a homogeneous protein sample of *apo*-IdmK for structural studies (Figure 4.12).

Before collecting any NMR spectra, the extent of [^{15}N] labelling of IdmK was established. A sample of purified *apo*-IdmK was desalted into 50 mM ammonium acetate pH 7.4 (section 2.2.3.9) and the mass determined by ESI-MS. The amino acid sequence of the synthetic gene and the Internet tool Protein Calculator v3.4 (http://protcalc.sourceforge.net) were used to calculate the mass of uniformly labelled monoisotopic IdmK. The expected mass for uniform [^{15}N] labelled IdmK was 11 191 Da. Figure 4.12 shows the result of attempting to express [^{15}N] labelled IdmK.

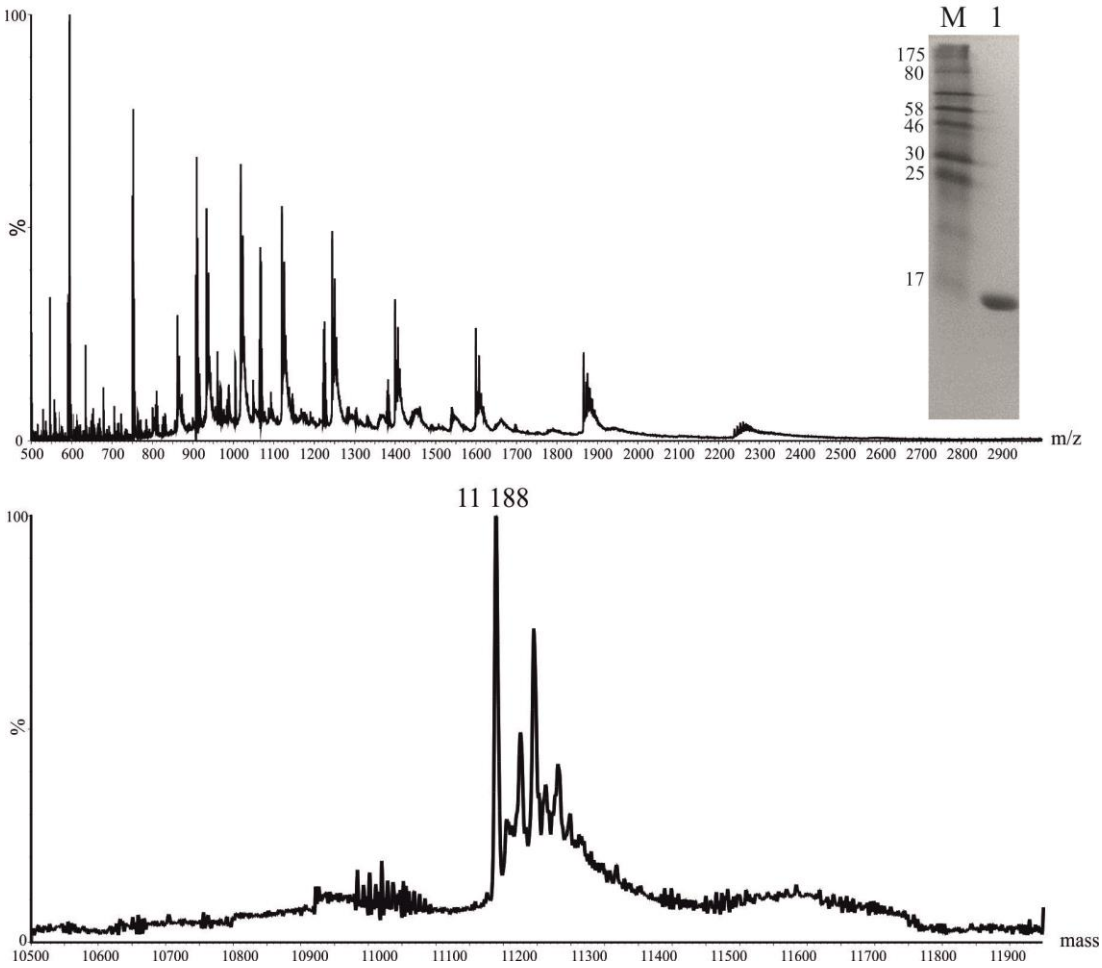


Figure 4.12- ESI-MS spectrum to analyse the extent of isotopic labelling of IdmK. *E. coli* BL21 (DE3) Gold cells containing the plasmid pK*idm*K were used for expression of IdmK in minimal media containing $^{15}NH_4Cl$ as the sole nitrogen source. *Apo*-IdmK was purified by nickel affinity chromatography followed by size exclusion chromatography. The labelled protein was then desalted into 50 mM ammonium acetate pH 7.4 and analysed by ESI-MS. A peak is observed at 11 188 Da, corresponding to fully ^{15}N- labelled IdmK. (Inset) Reducing SDS-PAGE of a sample of ^{15}N- labelled IdmK from size exclusion chromatography used in further NMR spectroscopy experiments.

Figure 4.12 shows that when IdmK is expressed in minimal medium containing $^{15}NH_4Cl$ as the sole nitrogen source it has an observed mass of 11 188 kDa, which is in excellent agreement with the calculated molecular mass of fully [^{15}N] labelled *apo*-IdmK. The next stage in determining the suitability of IdmK for structural studies using NMR spectroscopy was to collect a 2D ^{1}H-^{15}N Heteronuclear Single Quantum Coherence (^{1}H-^{15}N HSQC) spectrum.

4.5.2 Collection of a 2D heteronuclear single quantum coherence (HSQC) spectrum

Collection of a ^{1}H-^{15}N HSQC spectrum allows initial analysis as to whether the target protein is suitable for further studies by NMR spectroscopy and therefore is the first spectra collected (Cavanagh *et al.*, 2007; Rehm *et al.*, 2002). Spectra were collected of *apo*-IdmK at a concentration of 0.5 mM in 50 mM Tris.HCl pH 7.4, 0.2% (w/v) sodium azide. D_2O was added to the sample to a final concentration of 10% (v/v). The ^{1}H-^{15}N HSQC spectrum was collected at 25°C at a field strength of 500 MHz (Table 2.1 shows experimental parameters). Table 4.3 shows the coherence transfer or the path of magnetisation in the ^{1}H-^{15}N HSQC experiment.

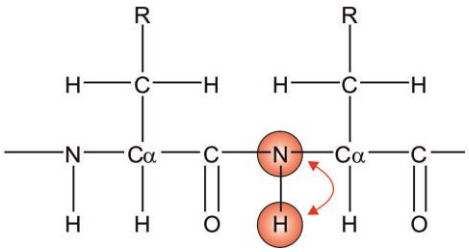
Experiment	Correlation observed	Magnetisation transfer
^{1}H-^{15}N HSQC	$^{1}H^{N}_{i}$-$^{15}N_{i}$	

Table 4.3- Coherence transfer of the 2D Heteronuclear experiments ^{1}H-^{15}N HSQC. The path of magnetisation is indicated by the double-headed red arrow, indicating the out-and-back nature of this experiment. The nuclei detected are highlighted in red circles (Cavanagh *et al.*, 2007).

Figure 4.13 shows the ^{1}H-^{15}N HSQC spectrum collected of [^{15}N] labelled *apo*-IdmK. In a ^{1}H-^{15}N HSQC each peak corresponds to a proton coupled to nitrogen, therefore in this experiment each amide in the protein should have a representative peak in the ^{1}H-^{15}N HSQC spectrum.

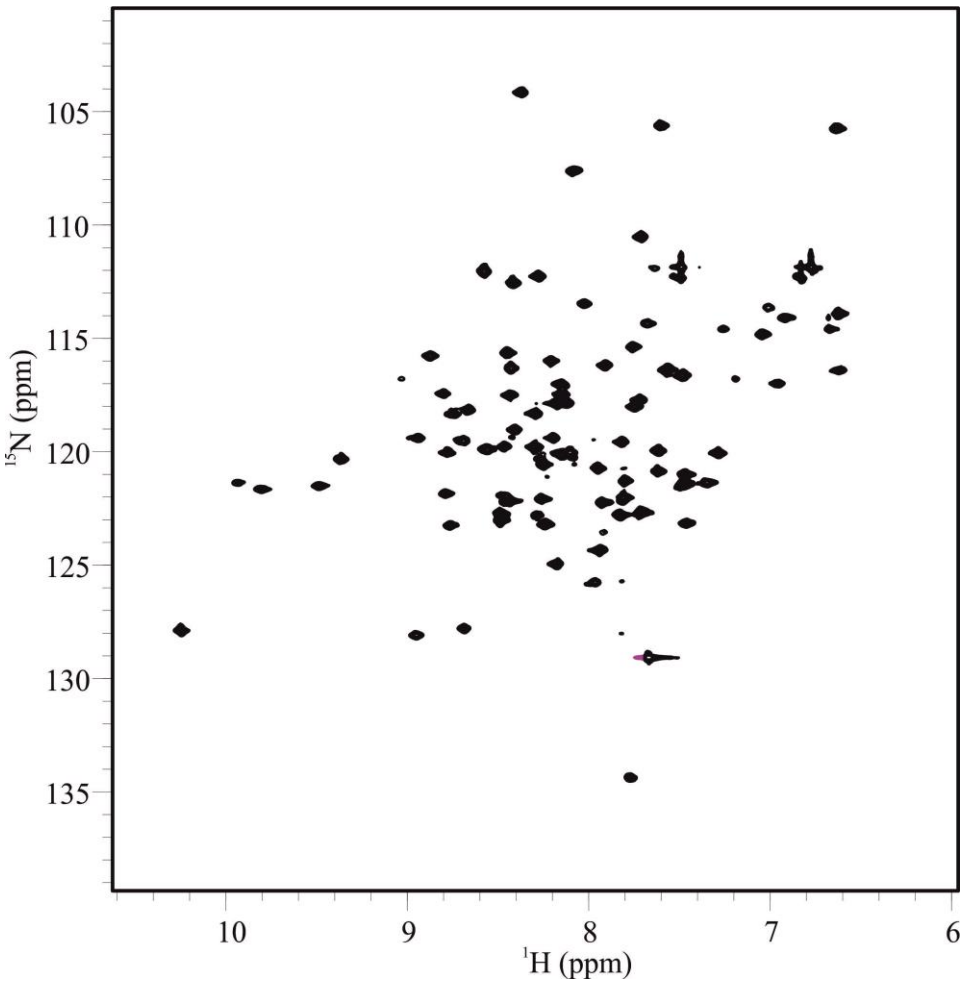


Figure 4.13- ^{1}H-^{15}N HSQC spectrum of *apo*-IdmK. The spectrum was measured in 50 mM Tris.HCl pH 7.4, 0.2% (w/v) sodium azide, at a protein concentration of 0.5 mM recorded at 500 MHz at 25 °C.

The good peak dispersion in the spectrum (Figure 4.13) and sharp line intensities indicate that the protein is folded and stable in these conditions. IdmK is a 98-residue protein including the His_6-tag; approximately 90 % of the peaks, representative of the amides, are present in the ^{1}H-^{15}N HSQC. The number and equal intensities of the peaks indicates that there is little or no protein dynamics on a mircrosecond-millisecond timescale that will hinder structure determination by NMR. This spectrum indicates that NMR may be able to be used to elucidate an overall fold for IdmK.

As each peak represents an amide in the protein the ^{1}H-^{15}N HSQC will provide the HN resonances, or 'root' resonances, to aid in the assignment of the triple resonance spectra and consequently structure determination of IdmK (described in chapter 5). The next requirement for structure determination by NMR is the expression and purification of doubly labelled [^{13}C, ^{15}N] labelled IdmK.

4.5.3 Expression and purification of [^{13}C, ^{15}N] IdmK

Initial experiments showed that IdmK could be expressed in minimal medium containing 4 g/L of glucose; however for cost efficiency when using ^{13}C-glucose as the carbon source, the concentration of glucose required for protein over expression was investigated. Expression trials with minimal medium containing 2, 3 and 4 g of glucose per litre were tested. Protein was expressed as previously described (section 4.3.1), 3 separate cultures were grown with the varying glucose concentrations. Cells were harvested by centrifugation and whole cell lysate analysed by SDS-PAGE (Figure 4.14).

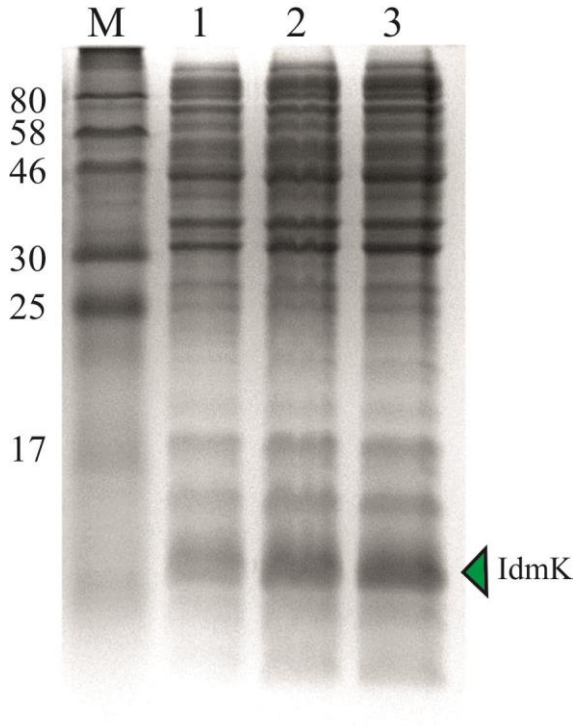


Figure 4.14- Reducing SDS-PAGE to analyse the effects of changing the concentration of glucose in minimal media on protein expression. Lane 1 is the whole cell lysate of IdmK grown in minimal media containing 2 g/L of glucose, lane 2 contains 3 g/L of glucose and lane 3 contains 4 g/L of glucose.

The expression trial showed that 2 g of glucose per litre was sufficient for protein expression. Subsequent protein expression for double labelled protein contained 10 mL 20 % (w/v) D-glucose-$^{13}C_6$ purchased from Cambridge Isotope Laboratories in 1 L. To produce [^{13}C,^{15}N] labelled protein IdmK was expressed in medium containing $^{15}NH_4Cl$ and D-glucose-$^{13}C_6$ as the sole nitrogen and carbon sources (section 2.1.7). IdmK was purified by nickel affinity chromatography. Samples from each stage of purification were run on a SDS-PAGE gel for analysis (Figure 4.15).

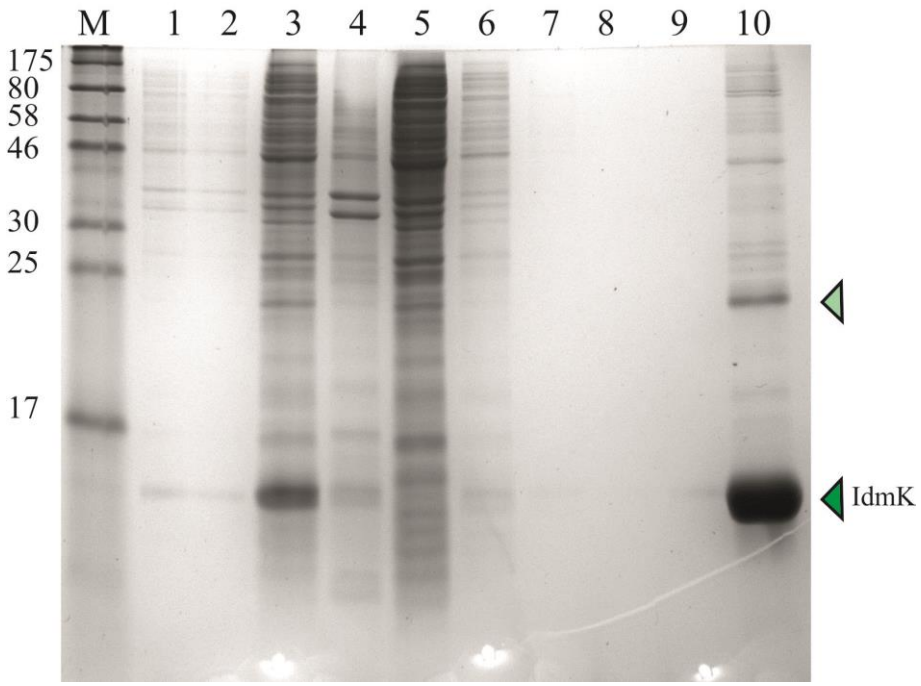


Figure 4.15- Reducing SDS-PAGE gel of [^{13}C, ^{15}N] IdmK purification. *E. coli* BL21 (DE3) Gold cells containing the pK*idm*K plasmid were expressed in minimal medium containing 2 g L^{-1} of $^{13}C_6$ glucose. Cells expressing labelled IdmK (1) were harvested and lysed (2). The soluble (3) and insoluble (4) fractions were separated by centrifugation. The soluble cell fraction was loaded onto nickel resin for one hour. The supernatant was then removed (5) and the resin washed with a low imidazole (20 mM) buffer (6-9). Finally protein was eluted from the resin using a high imidazole buffer (0.5 M). (10) Protein species eluted from the nickel resin correspond to monomer *apo*-IdmK (dark green arrow) and dimeric *holo*-IdmK (light green arrow).

As before, a homogeneous sample of *apo*-IdmK was obtained by size exclusion chromatography. The protein sample was then concentrated to 0.5 mM and the extent of labelling analysed by mass spectrometry (Figure 4.16).

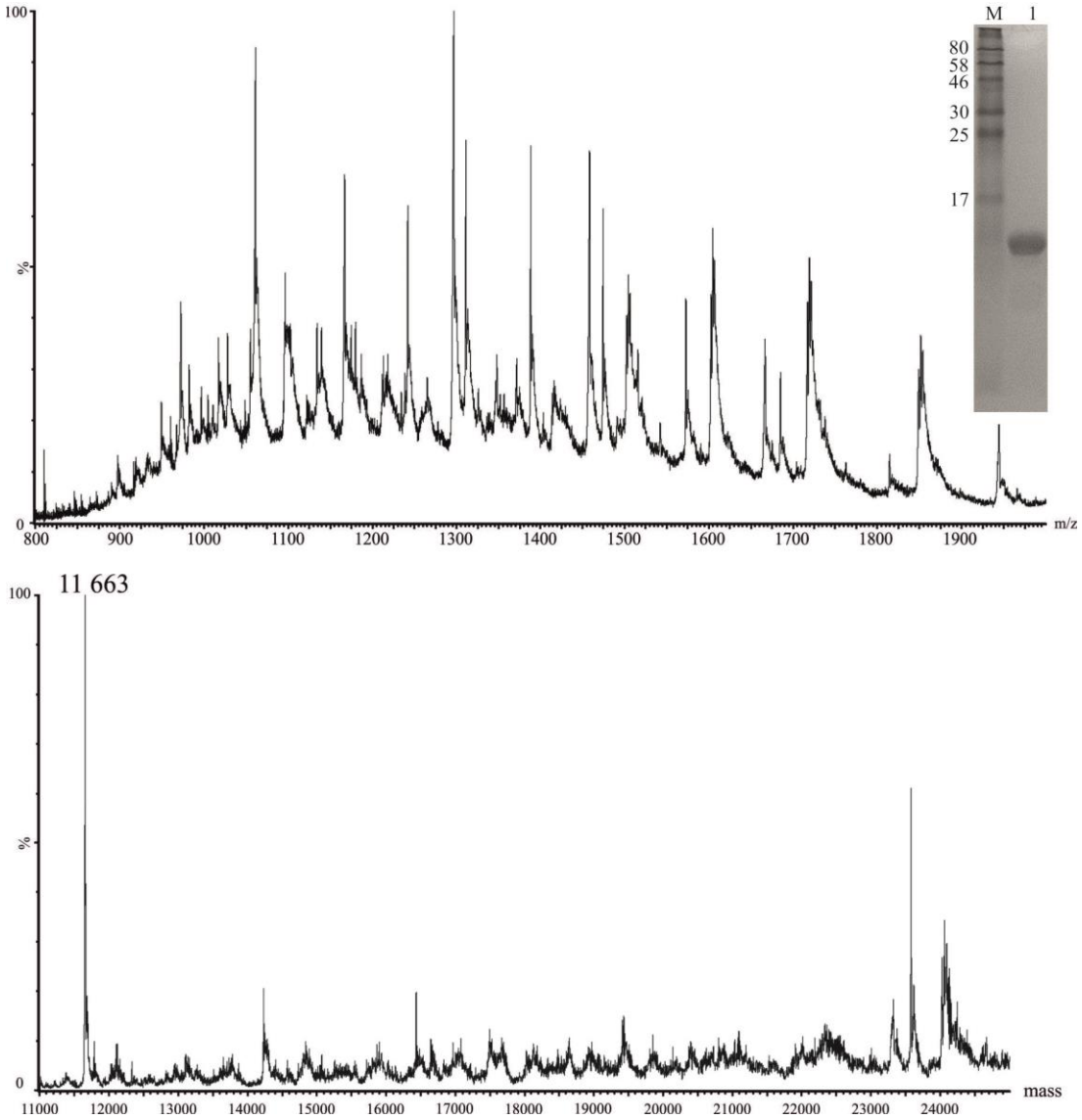


Figure 4.16- ESI-MS spectrum to analyse the extent of ^{13}C and ^{15}N isotopic labelling of IdmK. *E. coli* BL21 (DE3) Gold cells containing the plasmid pK*idm*K were used for expression of IdmK in minimal medium containing $^{15}NH_4Cl$ as the sole nitrogen source and D-Glucose-$^{13}C_6$ as the sole carbon source. *Apo*-IdmK was purified by nickel affinity chromatography followed by size exclusion chromatography. The labelled protein was then desalted into 50 mM ammonium acetate pH 7.4 and analysed by ESI-MS. A peak is observed at 11 663 Da, corresponding to fully ^{13}C, ^{15}N- labelled IdmK. (Inset) Reducing SDS-PAGE of ^{13}C, ^{15}N- labelled IdmK purified by size exclusion chromatography used in further NMR spectroscopy experiments.

The expected mass for ^{13}C, ^{15}N labelled IdmK as calculated by Protein Calculator v3.4 (http://protcalc.sourceforge.net) using the amino acid sequence of the synthetic gene was 11 668 Da. The observed mass for IdmK expressed in minimal medium containing $^{15}NH_4Cl$ and D-glucose-$^{13}C_6$ as the sole nitrogen and carbon sources was 11 663 Da indicating that 99.9 % of the protein was isotopically labelled.

Having successfully labelled *apo*-IdmK (^{15}N and ^{13}C) the next step in structure determination is assignment of the protein backbone, determining the chemical shifts of not only the ^{1}H and ^{15}N nuclei of the amide bond but also the Cα, Cβ and carbonyl nuclei. To do this 3D NMR spectra need to be collected. Chapter 5 will discuss the stages of structure determination by NMR spectroscopy.

4.6 Summary

IdmK was successfully cloned into pK223-3 and protein expressed and purified, as confirmed by SDS-PAGE and ESI-MS (Figures 4.4 and 4.5). Expression of IdmK produced both monomeric and dimeric species. Treatment of the dimeric IdmK species with TCEP, a reducing agent, showed that during expression of IdmK a PPTase endogenous to *E. coli* was able to phosphopantetheinylate IdmK. The dimeric species occurs due to the formation of a disulphide bond between the terminal thiol groups of two phosphopantetheine modifications. Contrary to expectations, both *apo*- and *holo* species were expressed in *E. coli*. BLAST searches in combination with multiple sequence alignments identified the potential serine, which becomes phosphopantetheinylated; this was confirmed by making a single point mutation of the identified serine and analysis by SDS-PAGE and ESI-MS. This showed only monomeric *apo*-IdmK was expressed when serine 44 was mutated to an alanine. The dimerisation of *holo*-IdmK allows purification of the *apo*- or *holo*-IdmK thus enabling a homogenous protein sample to be obtained. IdmK is an essential non-catalytic protein; it requires IdmJ to activate the amino acid it is to be loaded with. Initial attempts at structural characterisation of IdmK identified protein crystallisation conditions which produced crystals used for high resolution data collection. Subsequent analysis of the data collected showed that there were a large number of IdmK monomers within the asymmetric unit cell which meant molecular replacement could not be used for structure elucidation. Of the alternative options for structure determination NMR spectroscopy seemed a promising alternative, with a number of carrier protein structures having previously being solved by multidimensional NMR spectroscopy. Expression of ^{15}N labelled IdmK followed by analysis by collecting a ^{1}H-^{15}N HSQC spectrum suggested that IdmK was suitable

for structural studies by NMR and thus [^{13}C,^{15}N] labelled IdmK was expressed and purified for backbone assignments.

4.7 Discussion

Successful cloning and expression of IdmK has allowed characterisation of the site of post-translational modification. The unexpected post-translational modification, carried out by a PPTase endogenous to *E. coli* such as EntD (Lambalot *et al.*, 1996; Gehring *et al.*, 1997), enables purification of homogeneous *apo*- and *holo*- IdmK samples for biochemical and structural characterisation. Although structure determination using X-ray crystallography was problematic, screening for a new crystallisation condition, could open up this avenue for further structural characterisation. Expression of IdmK in minimal medium with $^{15}NH_4Cl$ and D-glucose-$^{13}C_6$ enabled expression of single, [^{15}N], labelled IdmK and double, [^{13}C, ^{15}N], labelled IdmK. An initial ^{1}H-^{15}N-HSQC showed that IdmK was suitable for structural studies using NMR. Chapter 5 describes the use of NMR for structure determination and how this was carried out for determining an overall fold for IdmK.

5. Structural characterisation of IdmK by nuclear magnetic resonance spectroscopy

The previous chapter presented the results of cloning and expressing the putative proline carrier protein from *S. antibioticus* NRRL 8167 in *E. coli*. Although structural characterisation by X-ray crystallography began, difficulties in solving the phase problem due to the high number of monomers in the asymmetric unit cell hindered the continued use of this method. Initial investigation into the suitability of IdmK for structural studies by NMR revealed that IdmK could be isotopically labelled, and collection and analysis of an ^{1}H-^{15}N HSQC spectrum showed that IdmK was in a folded and a single monomeric conformer, thus suitable for structure determination by NMR.

NMR as a tool for structure determination is particularly suitable for proteins of approximately 30 kDa or less, and is possible due to the intrinsic spin properties of particular nuclei (Cavanagh *et al.*, 2007). Below, a brief introduction to NMR spectroscopy will be given.

In 1946 the occurrence of nuclear magnetism was reported when Purcell *et al.* (1946) and Bloch *et al.* (1946) measured the NMR spectra of paraffin and water, respectively. This was possible due to nuclei possessing nuclear spin angular momentum. Nuclear spin angular momentum is a quantum mechanical property, characterised by the nuclear spin quantum number, *I* (Cavanagh *et al.*, 2007; Teng, 2005). Any nuclei with odd numbers of either neutrons or protons have a spin quantum number >0, some nuclei have half integer numbers if the neutrons and protons are not both odd e.g. ^{1}H has a spin quantum number of a ½; and ^{12}C has a spin quantum number of 0. Nuclei with a spin quantum number of 0 are NMR inactive (Cavanagh *et al.*, 2007; Teng, 2005). Table 5.1 shows the properties of nuclei used in biological NMR experiments.

Nucleus	Spin quantum number, *I*	Gyromagnetic ratio, γ ($T^{-1}s^{-1}$)	Natural abundance (%)
^{1}H	½	2.6752×10^{8}	99.99
^{2}H	1	4.107×10^{7}	0.012
^{13}C	½	6.728×10^{7}	1.07
^{15}N	½	-2.713×10^{7}	0.37
^{19}F	½	2.518×10^{7}	100.0
^{31}P	½	1.0839×10^{8}	100.0

Table 5.1- Properties of biologically relevant nuclei in NMR spectroscopy. *I* is the nuclear spin quantum number, γ is the gyromagnetic ratio. The natural abundance for each particular isotope is given (%) (Cavanagh *et al.*, 2007).

Nuclei with spin angular momentum, i.e. NMR active nuclei with nonzero spin quantum numbers, also have nuclear magnetic moments (μ) (Cavanagh *et al.*, 2007). The nuclear magnetic moment of each nucleus is characterised by an individual unique constant, the gyromagnetic ratio (γ), which indicates how large its magnetic field is and is related to how receptive that particular nucleus is in NMR spectroscopy (shown for each isotope in Table 5.1) (Cavanagh *et al.*, 2007; Teng, 2005)). It is the interaction of the nuclear magnetic moment with an applied, external magnetic field (B_0), which gives rise to populations of nuclei with measurable differences in energy states. The NMR signal or resonance is produced as a result of the nuclear magnetic moment dictating the nuclear rotation frequency in relation to the external magnetic field. It is the differences between the energy states which leads to different resonance frequencies. The volume of the signal from the resonance is proportional to the population difference of the states. When the spin quantum number is half an integral, for example ^{1}H, (Table 5.1) the nuclear magnetic moment can occupy one of two states due to the quantisation of angular momentum, either pointing with the main magnetic field Bo, which is typically on the Z axis (low energy, α) or against it (low energy, β). Spins in the α and β states precess around the magnetic field at their Larmor frequency (ω_0), and are randomly distributed with respect to the x-y plane (Cavanagh *et al.*, 2007; Keeler, 2002; Teng, 2005). Figure 5.1 illustrates the behaviour of nuclei in an applied magnetic field (B_0).

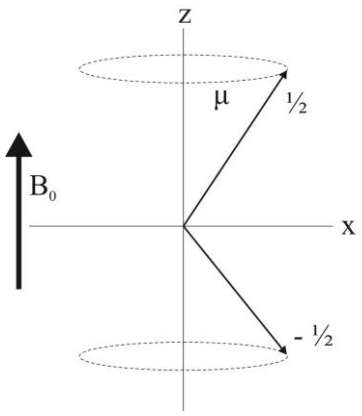


Figure 5.1- Precession of nuclear magnetic moment, μ, with a spin ½ or – ½. The vector diagram illustrates a simplification of the procession of the magnetic moment about a magnetic field (B_0) Figure adapted from Teng (2005) and Keeler (2002).

The Larmor frequency describes the frequency of the precession (Cavanagh *et al.*, 2007; Teng, 2005) and is defined by:

$$\omega_x = -\gamma_x B_0$$

Equation 5.1- Equation defining the Larmor frequency. Where γ is the gyromagnetic ratio and B_0 is the strength of the magnetic field, x is the nucleus (Cavanagh *et al.*, 2007; Teng, 2005).

As the Larmor frequency is dependent on the susceptibly of a nuclei to a magnetic field (γ) and also to the strength of the magnetic field, the frequency of precession increases in a linear manner when B_0 increases (Equation 5.1). When B_0 increases the energy difference between the two spin states that the nuclei can occupy increases, and the population difference between the two states increases due to the Boltzmann distribution. This results in a 1 in 10 000 population difference for protons at a magnetic field strength of 14.1 tesla. An NMR signal can only be observed when there is a difference in the populations of the two levels. At equilibrium, spins in the two states can be viewed as precessing as two cones pointing with and against the magnetic field (B_0) (Cavanagh *et al.*, 2007; Keeler, 2002; Teng, 2005). Summation of these spins gives a bulk magnetisation aligned with the field (Figure 5.2).

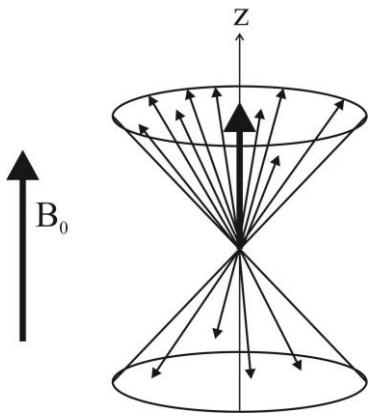


Figure 5.2- Bulk magnetisation of nuclei in an applied magnetic field (B_0) at equilibrium. At equilibrium the magnetic moments are aligned with the applied magnetic field (B_0), this is the bulk magnetisation (bold arrow) (Teng, 2005).

The field strength of an NMR magnet is measured in Tesla, however the strength of an NMR magnet is more often referred to by the proton Larmor frequency, for example a magnet with a field strength of 14.1 Tesla has a proton Larmor frequency of 600 MHz (Teng, 2005).

In the course of an NMR experiment, NMR active nuclei are excited by short radiofrequency pulse of defined power and time, which pushes the spin state away from equilibrium. Figure 5.3 shows the effects of applying a 90° pulse on the y axis ($90°_y$), which excites the bulk magnetisation from the z axis to the xy plane where it can be detected (Teng, 2005; Keeler, 2002).

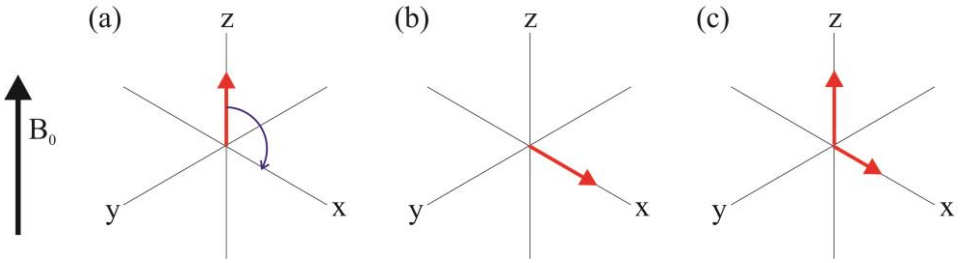


Figure 5.3- Behaviour of nuclear spins upon excitation by a radiofrequency pulse of 90° on the y axis. The red arrow illustrates the behaviour of the nuclei, as a magnetisation vector. Upon excitation by a $90°_y$ pulse (a) the magnetisation vector, nuclei, are rotated towards the xy plane (blue arrow) (b), the nuclei then begin to return to equilibrium aligning with the external magnetic field (B_0) over time (c) (Keeler, 2002).

After excitation nuclei take time to return to equilibrium, this is known as relaxation. T_1, or longitudinal relaxation describes one process for a return to equilibrium, and occurs by the nuclei exchanging energy with the environment over time leading to random spin flips from α to β and *vice versa*. The nuclei can undergo a second form of relaxation, T_2, spin-spin relaxation. This describes the loss of magnetisation on the xy plane due to local loss of magnetic field homogeneity in a molecule (Teng, 2005).

An NMR signal is observed in the form of Free Induction Decay (FID), which is produced by the oscillating magnetic field of all excited nuclei in the sample following a radio frequency pulse, The FID is measured by a receiver coil in the xy plane and is measured over time (typically < 1s). The FID signal is Fourier transformed to convert it into frequencies suitable for analysis, resulting in a NMR spectrum (Cavanagh *et al.*, 2007; Teng, 2005).

Different signals arise from nuclei within a molecule, including those from nuclei of the same isotope, due to the dependence of the magnetic field at the nucleus on the environment it is in. This dependence can be described as a chemical shift and moves the position of an NMR signal (Cavanagh *et al.*, 2007; Teng, 2005). The two main causes of the chemical shift are bonding electrons and aromatic rings, double bonds which contain pi orbitals (Cavanagh *et al.*, 2007; Teng, 2005). Bonding electrons in a molecule produces a magnetic field which opposes the induced bulk magnetisation. Therefore the greater the density of electrons, the smaller the local magnetic field and a lower frequency for the NMR transition is recorded. Aromatic rings and double bonds contain circular regions of pi electrons which can also induce large orientation dependant magnetic fields which are often described as ring current shifts and provide much of the dispersion of chemical shifts for proteins. The origin of the name chemical shift is due to its great dependence on the chemical environment, and it therefore provides a unique fingerprint of signals for a particular protein (Cavanagh *et al.*, 2007; Teng, 2005).

The normalisation of chemical shifts between different samples is achieved by referencing with compounds with chemical shifts defined as zero, such as tetramethylsilane (TMS), which is soluble in organic solvents, and the water soluble

form of TMS, DSS (4,4-demethyl-4-silapentane-1-sulfonic acid) (Teng, 2005). Chemical shift differences measured in ppm of the main magnetic field strength are not influenced by magnetic field strength though the frequencies that determine them are; therefore the same sample will produce identical spectra on the chemical shift scale when using any NMR spectrometers (Teng, 2005).

In structure determination and assignment, scalar (J) couplings provide information about the two atoms separated by one or more bonds due to interactions between bonded electrons. This is exploited for chemical shift assignments of atoms (Teng, 2005; Cavanagh *et al.*, 2007; Bax and Grzesiek, 1993), how scalar couplings are used for chemical shift assignments is described in section 5.2.

Dipolar couplings describe through space interactions of two nuclei, and are described in structure validation (section 5.3). Nuclear Overhauser effects (NOEs) arise from the transfer of energy from one spin to another spin through space by dipole-dipole interactions. NOEs can typically be observed between two protons which a distance of <6 Å and will produce NOE cross peaks in ^{1}H-^{1}H NOESY spectra which can be used in structure determination (section 5.4) (Cavanagh *et al.*, 2007; Teng, 2005).

While X-ray crystallography gives a static picture of the structure of proteins, NMR has the ability to provide dynamic information about the target protein. Figure 5.4 below illustrates the range of timescales of NMR spectroscopy can monitor *via* a range of techniques (Markwick *et al.*, 2008).

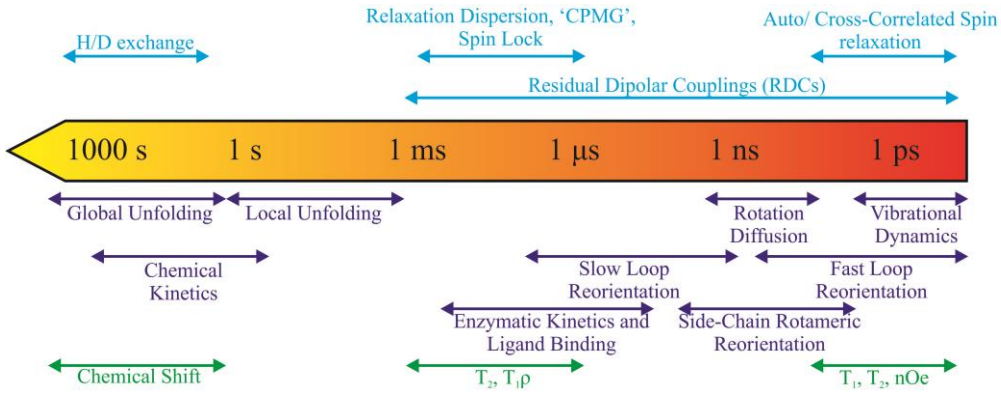


Figure 5.4- Timescales to illustrate the molecular dynamics (dark blue) that can be measured by NMR experiments (light blue) giving measurable data (green), Adapted from Markwick *et al.* (2008).

As mentioned in chapter 4, IdmK is the carrier protein in the first NRPS module of the indanomycin biosynthetic gene cluster. As with all carrier proteins, IdmK is a non-catalytic protein, phosphopantetheinylated for function. Structure determination of *apo*- carrier proteins shows a conserved canonical fold (Crosby and Crump, 2012; Li *et al.*, 2003). Figure 5.5 and 5.6 illustrate an alignment of an acyl carrier protein structures from FAS, PKS and an NRPS and a schematic of the canonical fold observed.

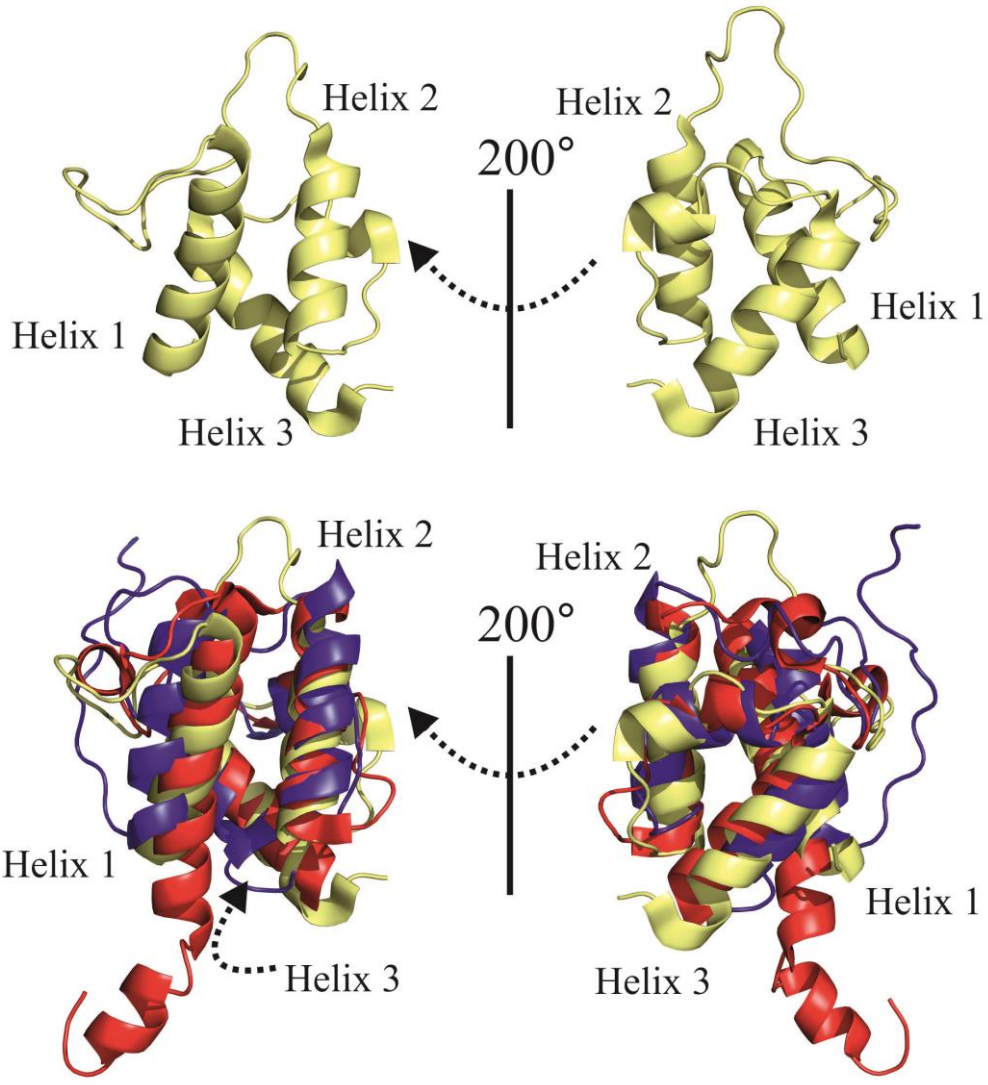


Figure 5.5- Canonical fold of carrier proteins. (Top) Structure of ACP from a fatty acid synthase (pdb code 2AVA) (Zornetzer *et al.*, 2006). (Bottom) An overlay of carrier proteins from a FAS (red, PDB code: 2AVA), a PKS (yellow, PDB code: 2JU2) and an NRPS (blue, PDB code 2GDW) (Zornetzer *et al.*, 2006; Alekseyev *et al.*, 2007; Koglin *et al.*, 2006).

Figure 5.6 shows a schematic of the secondary structural elements of the canonical fold of the homologous carrier protein structure. Helix 1 and helix 2 are antiparallel to each other, and helix 3 runs parallel to helix 2. Although slight deviations in helix lengths are seen in carrier proteins, generally they all appear to adopt the same fold (Crosby and Crump, 2012).

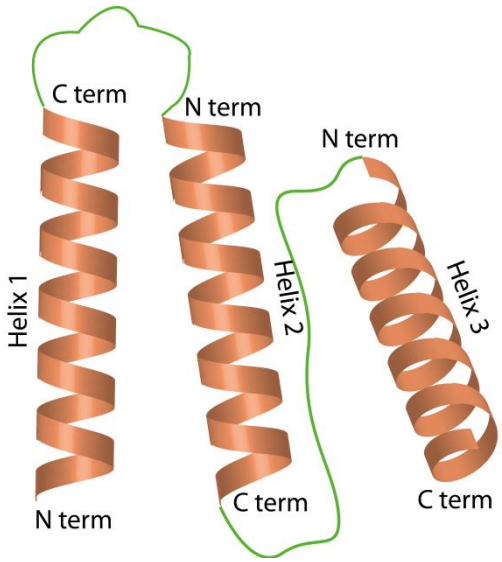


Figure 5.6- Schematic of the canonical fold of carrier proteins. Slight variations are seen in helix length. A fourth small pseudo helix is also sometimes observed between helix 2 and 3, this is more variable.

Structural studies by NMR spectroscopy of the TycC3 peptidyl carrier protein revealed conformational changes are required for access to the pantetheine co-factor, showing significant conformational changes between what is described as the A/H state, an equilibrium state adopted by both *apo*- and *holo*- conformers of the PCP, and the *holo*- and *apo*- states. These conformational changes are suggested to bring the terminal thiol, where the NRP chain is located in proximity with the adenylation and condensation domains (Weber *et al.*, 2000; Koglin *et al.*, 2006). Structure determination of IdmK by NMR spectroscopy will provide both a structure and basic dynamic information for the protein, providing a foundation for investigation into the dynamics and mechanisms of starter unit construction in indanomycin biosynthesis. This chapter describes the experiments and stages required for structure determination and validation of the initial structure calculation of the *apo*-state of IdmK.

5.1 Backbone assignments of IdmK

5.1.1 Triple resonance experiments

Prior to committing to structure determination of IdmK using NMR spectroscopy a ^{1}H-^{15}N HSQC spectrum was collected. This spectrum provided information on the conformational state of the protein and if it is folded in solution, thus indicating the suitability for further NMR experiments (Cavanagh *et al.*, 2007; Rehm *et al.*, 2002). The ^{1}H-^{15}N HSQC spectrum of IdmK (Chapter 4, Figure 4.13) indicated that the protein was suitable for structural studies by NMR as it contained the correct number of peaks and was well resolved and sharp. The next step was to collect triple resonance spectra for the sequential assignment of the protein backbone. Triple resonance experiments, originally described in 1990 for the backbone assignment of calmodulin (Ikura *et al.*, 1990; Kay *et al.*, 1990), are an alternative means of backbone assignments from the previous use of short range NOEs (Wuthrich *et al.*, 1982). Unlike the short range NOEs which use distances between protons for sequential backbone assignments, triple resonance experiments use one-bond and two-bond scalar couplings, or J couplings, which are easier to measure and interpret once the molecule has been uniformly labelled with ^{13}C and ^{15}N isotopes (Cavanagh *et al.*, 2007). Figure 5.7 illustrates the one-bond scalar couplings exploited in triple resonance experiments.

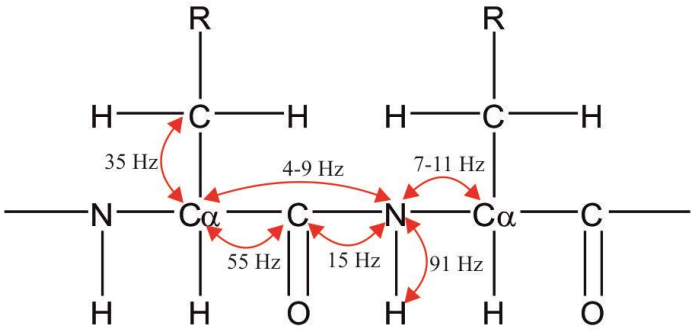
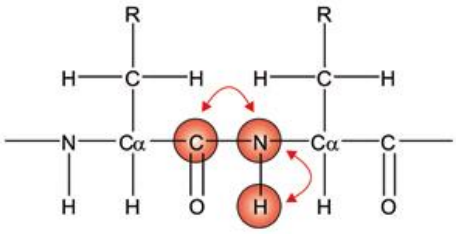
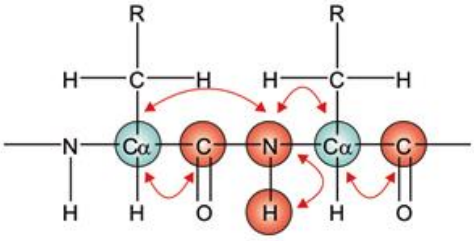
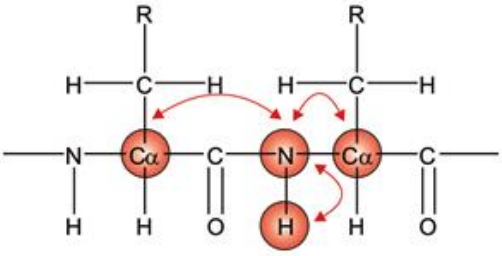


Figure 5.7- Scalar couplings observed in the protein backbone. The red arrows indicate the one and two bond scalar couplings used in triple resonance experiments (Teng, 2005).

Exploitation of these one-bond couplings (and two-bond in the case of the HN and Cα of the preceding residue) has led to the design of several sets of triple resonance experiments which correlate the spins of NMR active nuclei in the protein backbone (Bax and Grzesiek, 1993; Grzesiek and Bax, 1992; Ikura *et al.*, 1990; Cavanagh *et*

al., 2007). Running triple resonance experiments; HNCA, HNcoCA, HNCO, HNcaCO, HNCACB CBCAcoNH experiments gives chemical shift information about the $^{1}H^{N}$, ^{15}N, $^{13}C\alpha$, ^{13}CO and $^{13}C\beta$ of either the nuclei within the same residue as the H and N resonances observed (i) or the nuclei i-1 to the H and N resonances being measured and are used for sequential backbone assignments (Cavanagh *et al.*, 2007; Kay *et al.*, 1990; Grzesiek and Bax, 1992; Farmer *et al.*, 1992; Bax and Ikura, 1991; Clubb *et al.*, 1992; Bax and Grzesiek, 1993; Ikura *et al.*, 1990). Table 5.2 illustrates the transfer path used for each of the triple resonance experiments.

Experiment	**Correlation observed**	**Magnetisation transfer**
HNCO	$^{1}H^{N}_{i}$-$^{15}N_{i}$-$^{13}CO_{i-1}$	
HNcaCO	$^{1}H^{N}_{i}$-$^{15}N_{i}$-($^{13}C^{\alpha}_{i-1}$)-$^{13}CO_{i-1}$ $^{1}H^{N}_{i}$-$^{15}N_{i}$-($^{13}C^{\alpha}_{i}$)-$^{13}CO_{i}$	
HNCA	$^{1}H^{N}_{i}$-$^{15}N_{i}$-$^{13}C^{\alpha}_{i}$ $^{1}H^{N}_{i}$-$^{15}N_{i}$-$^{13}C^{\alpha}_{i-1}$	

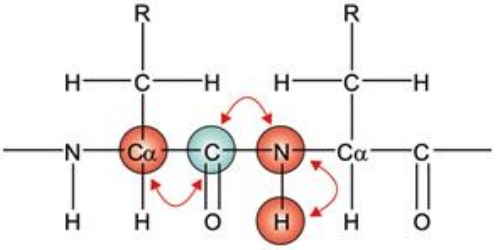
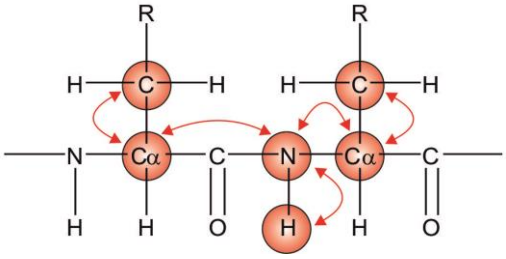
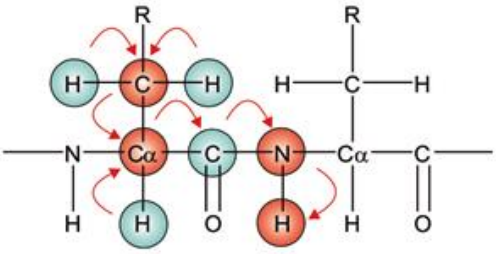
Experiment	Correlation observed	Magnetisation transfer
HNcoCA	$^{1}H^{N}_{i}$-$^{15}N_{i}$-($^{13}CO_{i-1}$)-$^{13}C^{\alpha}_{i-1}$	
HNCACB	$^{13}C^{\beta}_{i}$/$^{13}C^{\alpha}_{i}$-$^{15}N_{i}$-$^{1}H^{N}_{i}$ $^{13}C^{\beta}_{i-1}$/$^{13}C^{\alpha}_{i-1}$-$^{15}N_{i}$-$^{1}H^{N}_{i}$	
CBCAcoNH	$^{13}C^{\beta}_{i}$/$^{13}C^{\alpha}_{i}$-$^{15}N_{i+1}$-$^{1}H^{N}_{i+1}$	

Table 5.2- The coherence transfer in triple resonance experiments used for sequential backbone assignments. The arrows indicate the path of magnetisation; double headed arrows indicate transfer of an "out-and-back" nature. Red circles indicate nuclei that were detected, blue circles are those that were on the pathway but no chemical shift information was collected. Figure adapted from Cavanagh *et al.* (2007).

The triple resonance spectra HNCA, HNcoCA, HNCO, HNcaCO, HNCACB CBCAcoNH were collected for *apo*-IdmK, at a concentration of 0.5 mM, 50 mM Tris.HCl pH 7.4, 0.02% sodium azide containing 10 % (v/v) D_2O. Data were collected at 25 °C at a field strength of 600 MHz. Spectra were then processed as described in section 2.2.5.2, and imported into CCPNmr Analysis (Vranken *et al.*, 2005). CCPNmr analysis was used to peak pick the ^{1}H-^{15}N HSQC and all triple resonance spectra. Analysis was also used to assign anonymous spin systems in the triple resonance spectra to their root resonance in the ^{1}H-^{15}N HSQC. Once this was completed, semi-automated sequential backbone assignments were carried out.

5.1.2 Semi-automated backbone assignments using MARS

Sequential protein backbone assignments rely on the correlation of an NH group with the Cα, Cβ and carbonyl (CO) nuclei within the same residue and the preceding residue (i and i-1). Table 5.2 illustrates the correlation of nuclei within each experiment, and indicates whether they are used to collect chemical shift information from nuclei within the same or preceding residue. Collection of the HNCA spectra provides chemical shifts for the Cα peaks of residues (i) and (i-1) and collection of HNcoCA spectra provides chemical shift information for just the preceding residue (i-1). Superimposition of these two spectra enables the i and i-1 peaks in the HNCA spectra to be distinguished. The i-1 peak can then be matched to an i peak with an identical chemical shift in another residue, linking two residues together, sequentially. The HNCO and HNcaCO can be used to sequentially link the CO, and the HNCACB and CBCAcoNH can be used to sequentially assign residues using the Cα and Cβ nuclei in a similar manner and are important for overcoming degeneracies. Strip plots can be used to perform the sequential assignments. An example strip plot can be seen in Figure 5.8.

In order to speed up the complete backbone assignments of IdmK an automated assignment program called MARS was used (Jung and Zweckstetter, 2004). MARS works by trying to optimise local and global assignments in order to reduce the carry through of initial errors to final assignments, is conservative in the assignments it makes and is able to complete assignments with missing chemical shift data (Jung and Zweckstetter, 2004). Prior to the use of MARS both the ^{1}H-^{15}N HSQC and triple resonance spectra had been peak picked and the triple resonance peaks had been assigned to the same spin system as their root resonances in the ^{1}H-^{15}N HSQC as described above. Peaks in the same spin system are referred to as pseudoresidues (PR) prior to their assignments. Once PRs are linked into segments and then mapped onto the amino acid sequence of the protein, MARS identifies the amino acid types by characteristic chemical shifts, such as the Cβ chemical shifts of serine and threonine residues and detects sequential connectivities of the PR (Moseley and Montelione, 1999; Jung and Zweckstetter, 2004). MARS evaluates the assignments and its output indicates the reliability of assignments, classing the confidence of the assignment as High (H), Medium (M) or Low (L) (Jung and Zweckstetter, 2004). All assignments by MARS were also manually checked using strip plots (Figure 5.8).

Further analysis of side chain assignments also showed all MARS assignments used subsequently were correct. A total of 92% of backbone nuclei were assigned using this method. Chemical shifts can be found in a table in the appendix.

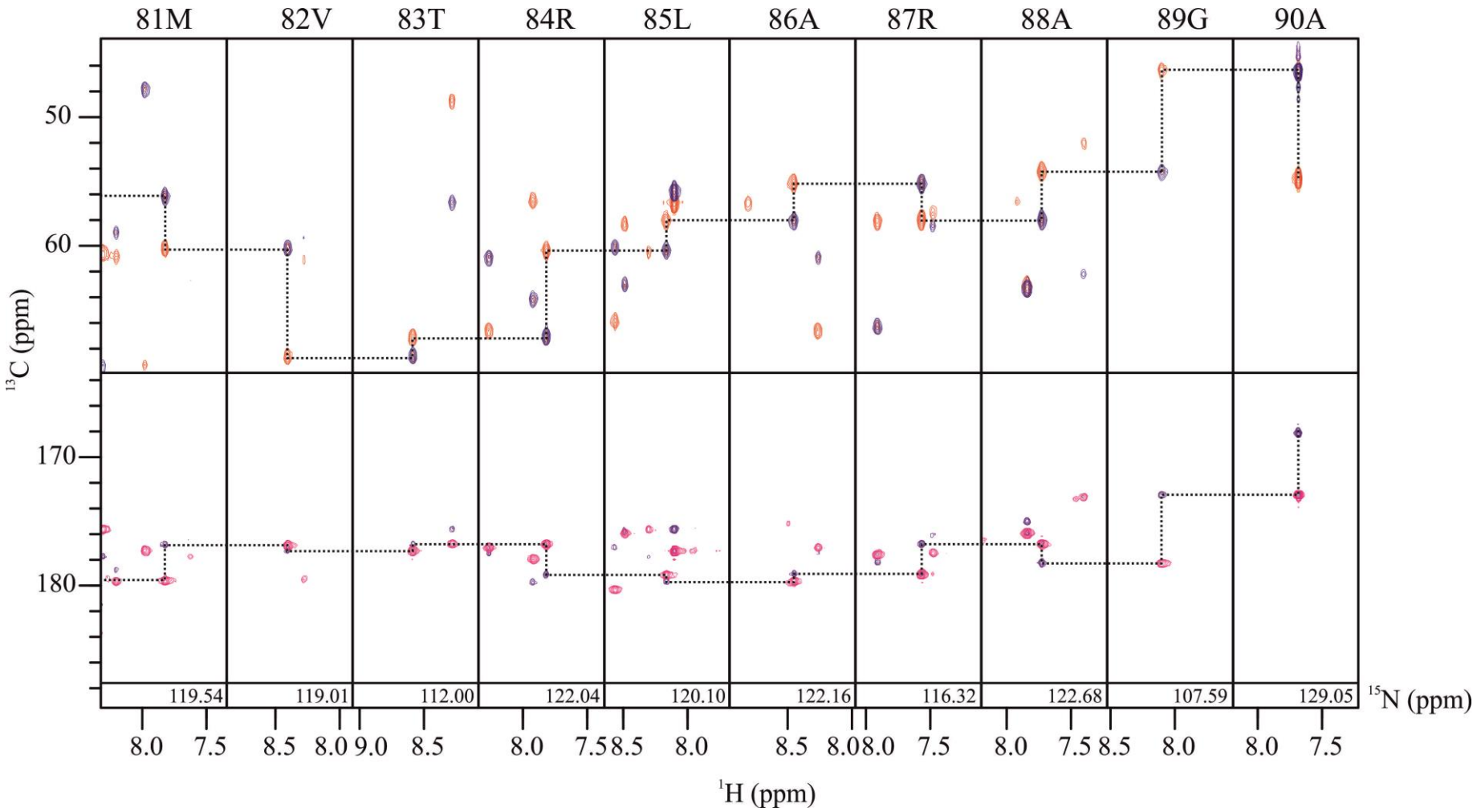


Figure 5.8- Strip plot of the last 10 residues of the backbone of IdmK. Strip plots were used to manually check the automated assignments completed by MARS. (Top) HNCA (red) and HNcoCA (blue) spectra were used to assign the Cα chemical shifts. (Bottom) HNCO (pink) and HNcaCO (purple) spectra were used to assign the carbonyl chemical shifts. Triple resonance spectra were measured in 50 mM Tris.HCl pH 7.4, 0.2% (w/v) sodium azide, at a concentration of 0.5 mM recorded at 600 MHz at 25 °C.

The successful assignment of the ^{1}H-^{15}N HSQC spectrum is shown in Figure 5.9. The triple resonance spectra provided the chemical shift information for the following nuclei of IdmK; N, HN, Cα, Cβ and CO which are given in table the chemical shift table in the appendix.

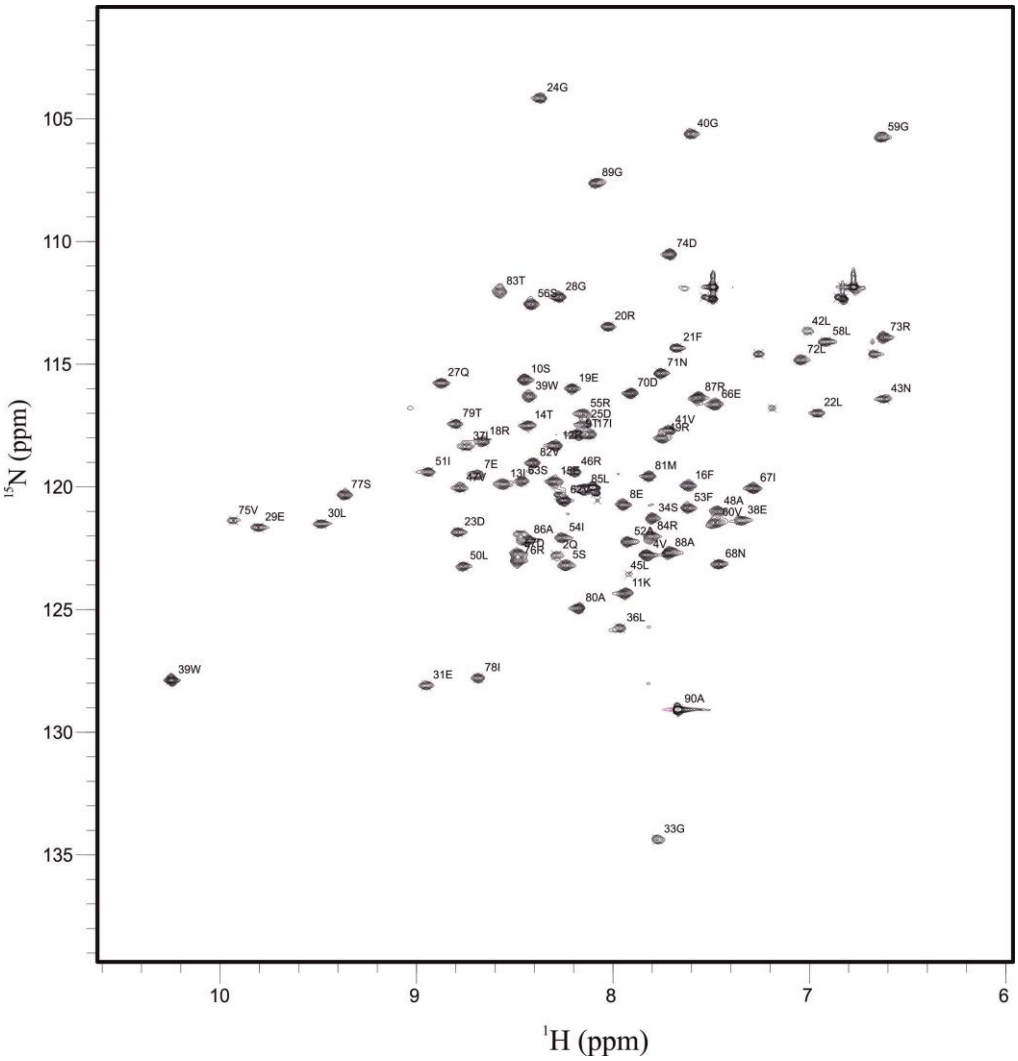


Figure 5.9- Assigned ^{1}H-^{15}N HSQC of *apo*-IdmK collected in 50 mM Tris.HCl pH 7.4, 0.2% (w/v) sodium azide, at a concentration of 0.5 mM recorded at 500 MHz at 25 °C. 92 % of the backbone residues were assigned using MARS, providing chemical shifts for the NH nuclei.

As discussed above, the good peak dispersion and sharp lines indicate a well-folded and stable protein in solution. Recent developments in NMR structure calculations now allow the use of chemical shifts alone to determine an overall protein backbone fold (Shen *et al.*, 2008; Shen *et al.*, 2009b). The chemical shifts for HN, N, Hα, Cα, Cβ and CO generated by backbone assignments were subsequently used in secondary and tertiary structure calculations.

5.2 Structure determination using chemical shifts

The unique chemical shifts for the backbone nuclei in particular secondary structure shifts and the effect of the ring current can be used to determine secondary, and more recently, tertiary chemical structures using +TALOS and CS-ROSETTA respectively (Shen *et al.*, 2009a; Shen *et al.*, 2008; Shen *et al.*, 2009b) and this approach was adopted to determine a preliminary overall fold of IdmK.

5.2.1 Determination of secondary structure

To predict the secondary structural elements of IdmK +TALOS was used (Shen *et al.*, 2009a). TALOS, Torsion Angle Likelihood Obtained from Shifts and sequence similarity, uses experimentally determined chemical shifts of HN, N, Hα, Cα, Cβ and CO nuclei to predict the phi and psi torsion angles of the protein backbone using a database of chemical shifts and the random coil shifts predicted by the primary sequence (Cornilescu *et al.*, 1999; Shen *et al.*, 2009a). TALOS is also able to predict how dynamic the residue is, quantified by an approximate value for the order parameter, S^2 (Figure 5.10 - bottom panel) (Berjanskii and Wishart, 2007).

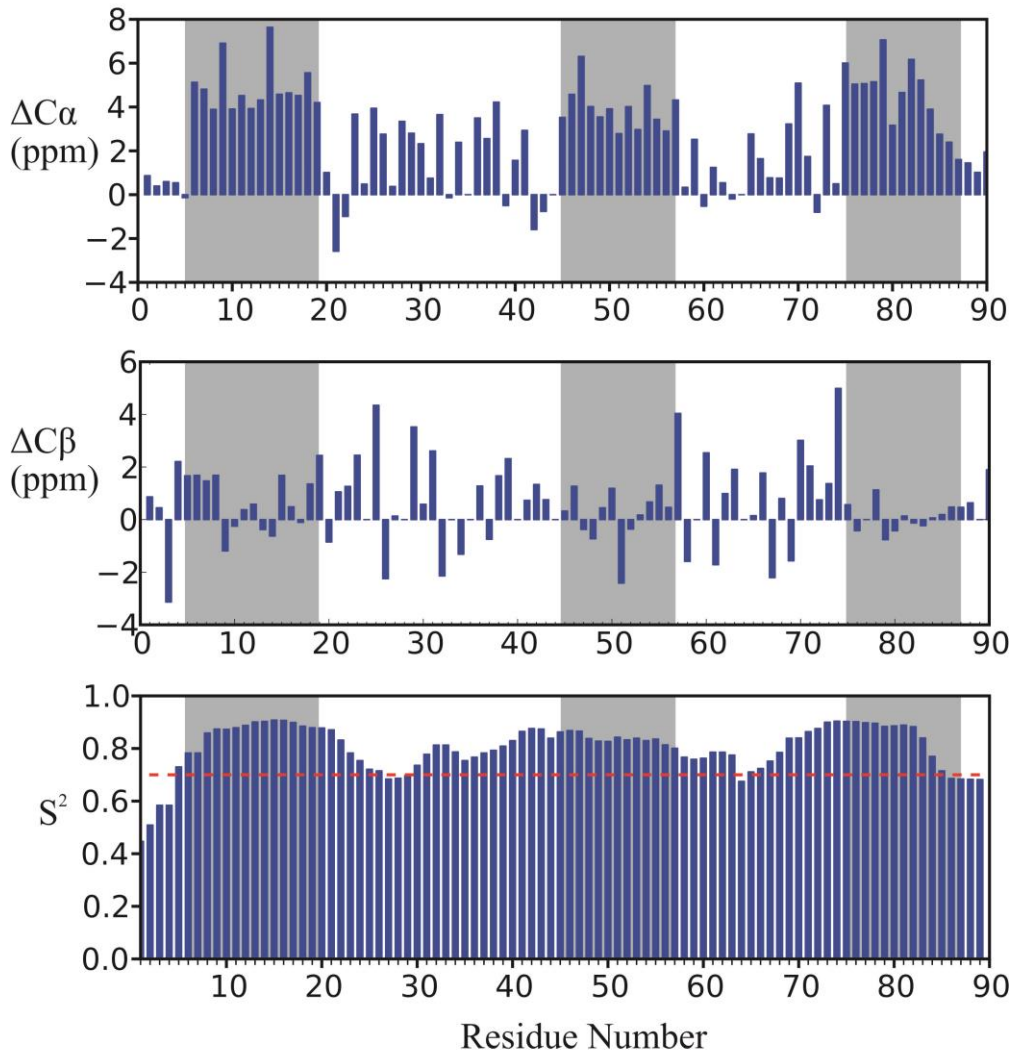


Figure 5.10- Secondary structure predictions by +TALOS (Shen *et al.* 2009a). (Top) Secondary chemical shifts of Cα atoms (ΔCα) and (middle) Cβ atoms (ΔCβ). ΔCα and ΔCβ describes the difference between measured chemical shifts and random coil chemical shifts. (Bottom) Predicted order parameter based on the secondary chemical shifts, the dashed line shows an typical cut-off of 0.7, below which residues are assumed to be dynamic and disordered (Engelke and Rüterjans, 1999). The grey boxes represent the α-helices predicted by TALOS.

The secondary chemical shifts depend on the orientation of the Cα-Hα bond in relation to each adjacent carbonyl (Tjandra and Bax, 1997b; Shen *et al.*, 2009a; Teng, 2005). Figure 5.10 shows the results of the secondary structure predictions for IdmK using +TALOS (Shen *et al.* 2009a). The top panel shows that the secondary chemical shifts of the Cα are positive, indicating that IdmK is an α-helical protein. This is reinforced by the fact that the Cβ chemical shifts show no overall pattern possibly due to slightly incorrect chemical shift referencing of the spectra. Chemical shift referencing will be completed with an internal reference compound (DSS). The positive ΔCα values indicated that IdmK has 3 helices, with helix 1 composed of

residues 6-20, helix 2 composed of residues 45-57 and helix 3 composed of residues 75-85. To compare the predicted secondary structure elements of IdmK with those of other carrier proteins the Dali server was used (Holm and Rosenstrom, 2010). Dali performs a 3D structure alignment of a target protein with those deposited in the PDB. Figure 5.11 shows the results of the alignment by Dali.

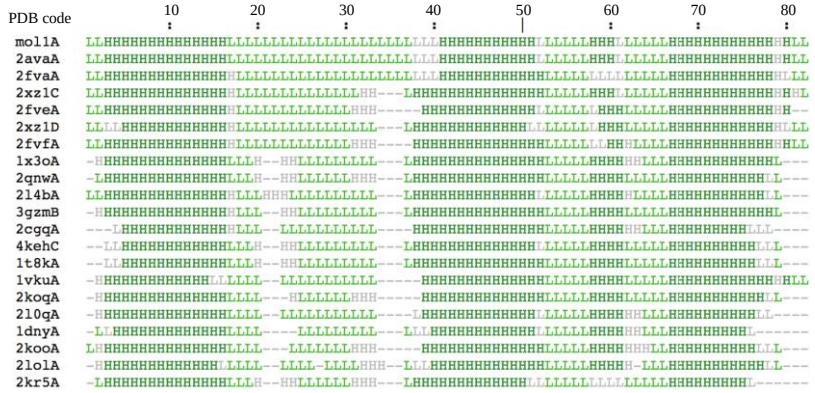


Figure 5.11- Secondary structure alignments of carrier proteins by Dali. H indicates the residue resides within a helix and L indicates the residue is in a loop. The PDB codes refer to carrier protein structures with comparable secondary structure selected by DALI deposited in the PDB.

This analysis revealed that carrier proteins consist of: helix 1, centred around residue 10, with a length of 14 residues (+/- 2), helix 2, centred around residue 46, with a length of 11 residues (+/- 4), this is followed by a small “pseudo” helix which is not always present and helix 3, centred around residue 74, which is 13 residues long (+/- 3). Exact locations of the helices vary slightly due to differences in the lengths of the carrier proteins. IdmK shows comparable helix lengths and relative locations in the sequence.

An order parameter of 0 indicates a completely disordered residue and an order parameter of 1 is a completely rigid protein (Berjanskii and Wishart, 2007; Engelke and Rüterjans, 1999). The order parameters shown in Figure 5.10 indicated that IdmK is a well-ordered protein, apart from the extreme N-and C-terminal residues. These results provide the first evidence that IdmK may have the canonical carrier protein fold. Phi and psi angles predicted by TALOS can be used as restraints in structure calculations (Section 5.2.2). The chemical shifts were then used to determine the overall fold of IdmK using CS-ROSETTA.

5.2.2 Determination of the overall fold of IdmK using CS-ROSETTA

Until recently, it was not possible to define the fold of a protein structure by use of the torsion angles determined by TALOS (Cornilescu *et al.*, 1999; Shen *et al.*, 2009a), and structure determination from assigned side chain proton requires the addition of large numbers of standard NOE distance restraints. The past seven years has seen the emergence of programs which determine a protein fold solely from the chemical shifts of backbone and Cβ nuclei; CHESHIRE, CS-ROSETTA and CS23D (Cavalli *et al.*, 2007; Shen *et al.*, 2008; Wishart *et al.*, 2008). CS-ROSETTA, a program developed by the Baker lab (Shen *et al.*, 2008; Shen *et al.*, 2009b), utilises the experimentally determined chemical shifts to determine an overall protein fold *de novo*. The ROSETTA protein structure prediction algorithm is a software package that allows the determination of a protein fold based only on the amino acid sequence and chemical shifts determined in backbone assignments. The general strategy involves the selection of fragments based on the target amino acid sequence. 9 and 3 residue fragments are slected from known structures in a database, which are then assembled using a Monte Carlo simulated annealing approach. CS-ROSETTA uses the ^{13}Cα, ^{13}Cβ, ^{13}CO, ^{1}HA, ^{1}HN and ^{15}N chemical shifts to bias the fragment selection, so as that the selected fragments have chemical shifts that agree with the experimentally measured ones and so occupy the correct part of Ramachandran space. The lack of NMR assignments in structural databases may have previously hindered searches, however development of SPARTA (which is based on TALOS (Shen and Bax, 2013)), a program that can predict the chemical shifts during ROSETTA modelling and thus allows correctly folded structures to be selected. CS-ROSETTA has been evaluated using a number of proteins up to 15 kDa in size adopting a range of conformations (Shen *et al.*, 2009b), indicating that this program may be appropriate for determination of an overall fold of IdmK. Dr Gary Thompson aided in running the CS-ROSETTA calculations.

5000 ROSETTA models were calculated using the experimentally determined ^{13}Cα, ^{13}Cβ, ^{13}CO, ^{1}HA, ^{1}HN and ^{15}N chemical shifts, and were ranked by their CS-ROSETTA scores (Shen *et al.*, 2008; Shen *et al.*, 2009b). The results showed that the five best ROSETTA structures show an overall similar fold, with helix 1 antiparallel to helix 2 and 3 which are parallel to each other, as expected for a canonical carrier protein fold (Figure 5.12 and 5.6).

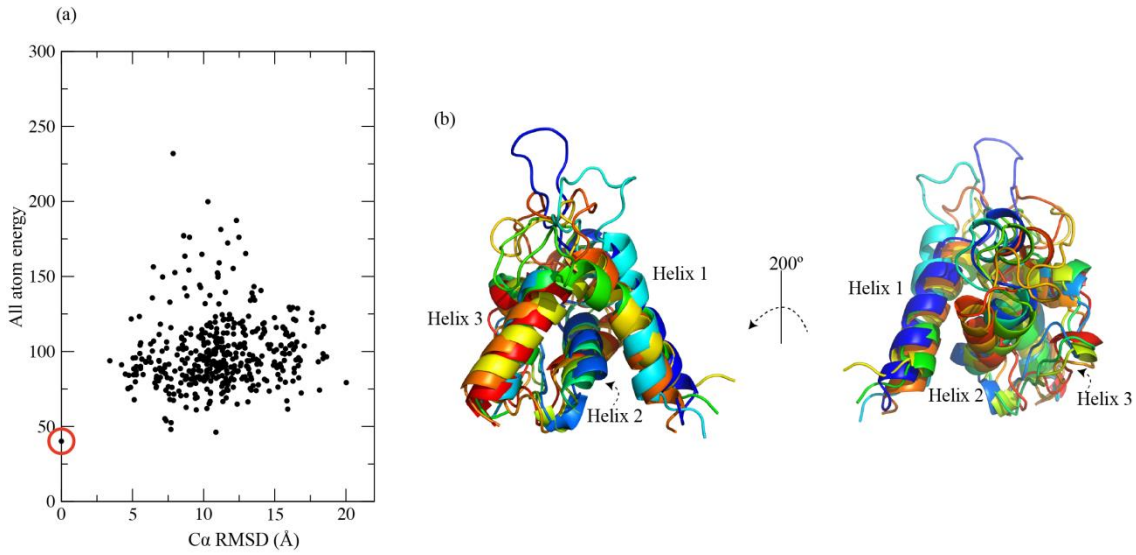


Figure 5.12- Chemical shift structure of IdmK using CS-ROSETTA. (a) Chevron plot of a subset of 500 of the best ROSETTA scored models from the 5000 models generated, the red circle indicated the model with the best ROSETTA score (b) Overlay of the 5 models with the best ROSETTA score.

Figure 5.12 shows a chevron plot with the best ROSETTA scored model highlighted. The models are analysed by plotting their all atom energy against the Cα RMSD indicates how similar the other models are to the lowest energy model. Although the CS-ROSETTA models appear to show the same overall orientation, i.e. helix 1 antiparallel to helix 2 and 3, which are parallel to each other, the exact angles between the helices deviate somewhat from the known canonical fold and between the models themselves. Though the observed structures don't show the clearest deviation in the energy vs Cα RMSD this may be attributed to the large content of loops or long meanders between helices. Figure 5.18 shows a comparison of the CS-ROSETTA model of IdmK with an acyl carrier protein from FAS.

Although CS-ROSETTA is able to predict the correct overall fold of IdmK, further experimental validation is required to assess the quality of the models. An easy experimental restraint to determine, that can be used to validate the models, are residual dipolar couplings (RDCs).

5.3 Structure validation using residual dipolar couplings

To validate the calculated CS-ROSETTA structures of IdmK, residual dipolar couplings (RDCs) were measured. RDCs depend on the distance between two NMR active nuclei (*i* and *j*) and the angle between them. As the distance between the nuclei which are directly bonded to one another is a known constant based on classical chemistry (bond lengths), RDCs (D_{ij}) provide accurate relative angular information between bonds (θ_{ij}) that is difficult to determine by other NMR methodologies (Bax, 2003; Ottiger *et al.*, 1998; Lipsitz and Tjandra, 2004; Teng, 2005).

$$D_{ij} = -\frac{\gamma_i \gamma_j \mu_0 h}{8\pi^3} \left\langle \frac{(3cos^2\theta_{ij}(t) - 1)}{2r_{ij}^3(t)} \right\rangle = 0$$

Equation 5.2- Determination of the RDC between two nuclei. Where γ_i and γ_j are the gyromagnetic ratios of the two NMR active nuclei, μ_0 is the vacuum permeability, *h* is Planks constant, r is the distance between the two nuclei and θ is the angle between them and t is time (Bax, 2003; Ottiger *et al.*, 1998).

Figure 5.12 shows the relationship between two nuclei within an external magnetic field, (B_0) defining the distance and angle between them.

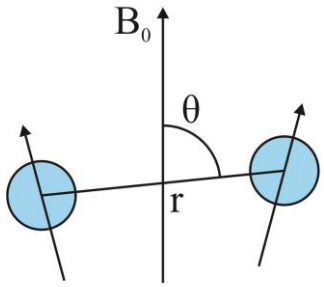


Figure 5.13- Dipole-dipole interaction between two magnetic moments. B_0 is the external magnetic field, r is the distance between the two nuclei and θ is the angle between them (Teng, 2005; Blackledge, M., personal communication).

In an isotropic solution, when a molecule is tumbling freely is solution, the RDCs are averaged to close to zero. However, in an anisotropic solution the molecular orientation is slightly restricted and this introduces measurable dipolar couplings (Equation 5.2 is no longer equal to zero).

As mentioned previously, partial alignment of the protein is required to introduce measureable dipolar couplings (Fleming and Matthews, 2004; de Alba and Tjandra, 2004). Figure 5.14 shows the partial alignment of molecules in an anisotropic solution, in order to introduce measurable dipolar couplings.

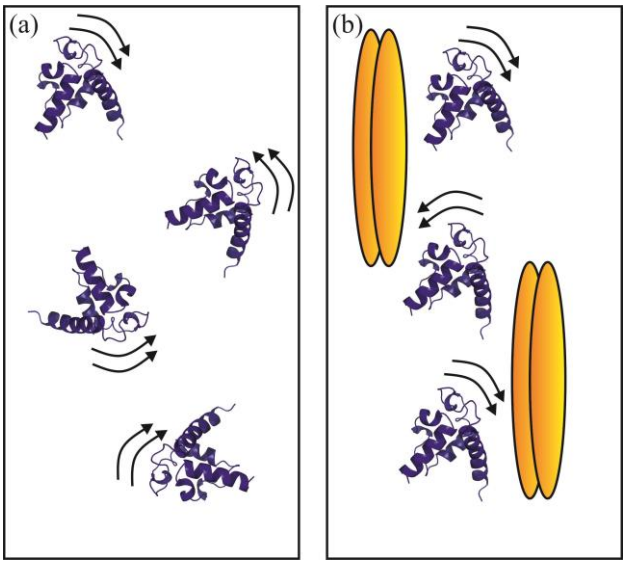


Figure 5.14- Partial alignment of a protein in an anisotropic solution. In an isotropic solution (a) the protein is tumbling freely in solution and RDCs are averaged to zero. In an anisotropic solution (b) the range of motion is partially restricted and this introduces measurable dipolar couplings (Blackledge M, personal communication).

Figure 5.15 shows a simple method by which RDCs can be directly measured by comparison of a coupled spectrum in anisotropic and isotropic solutions. It shows how the residual dipolar coupling can be measured by subtraction of the known one-bond J coupling from the coupling measured in the experiment performed in an anisotropic solution. However, for IdmK instead of measuring the difference in peak positions in an undecoupled HSQC as shown in Figure 5.14, a series of decoupled ^{1}H-^{15}N HSQC J-modulated spectra (Tjandra and Bax, 1997c) were collected with different time intervals (Δ) during which the coupling is encoded in the intensities of the peaks. This approach was used to accurately measure the RDCs in IdmK.

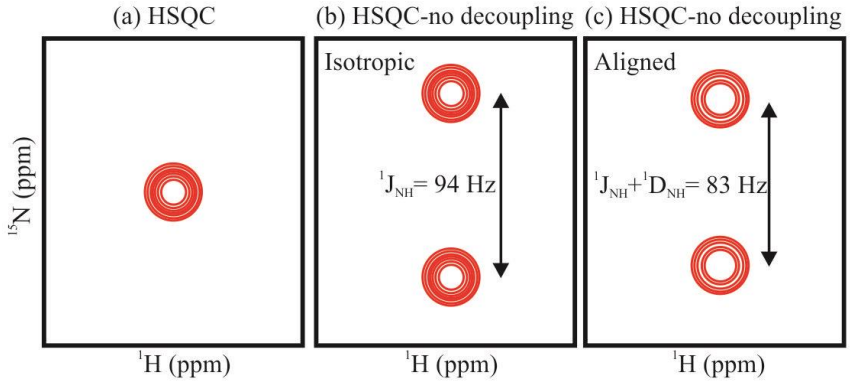


Figure 5.15- Measuring residual dipolar couplings in an isotropic and anisotropic solution. Panel (a) shows a peak in a decoupled ^{1}H-^{15}N HSQC spectrum. (b) Shows the decoupled HSQC in isotropic solution where the one one-bond j coupling is measured. (c) Shows a decoupled HSQC in an anisotropic solution where the one bond j-coupling and dipolar coupling are active (Ottiger *et al.*, 1998).

Partial alignment of the protein is often achieved using a liquid crystalline medium which can introduce measurable dipolar couplings (Tjandra and Bax, 1997a). A variety of different media can be used to partially align protein molecules such as bicelles, filamentous bacteriophage and SDS-PAGE gels (Fleming and Matthews, 2004; de Alba and Tjandra, 2004). The first liquid crystalline medium trialled for partial alignment of IdmK was Pf1 bacteriophage. Pf1 bacteriophage is a filamentous bacteriophage which aligns in parallel with the magnetic field due to being naturally susceptible to the magnetic field, thus creating a liquid crystalline medium and partially aligning the phage and the protein (Hansen *et al.*, 1998; Torbet and Maret, 1979). A ^{1}H-^{15}N HSQC spectrum was collected of IdmK at 0.5 mM in Tris.HCl pH 7.4, i.e. in an isotropic solution. Pf1 bacteriophage was then added to the isotropic solution at a final concentration of 8 mg mL^{-1} and a second ^{1}H-^{15}N HSQC spectrum was collected. Analysis of the anisotropic ^{1}H-^{15}N HSQC spectrum showed significant peak broadening also occurring, indicating an interaction of the Pf1 bacteriophage with IdmK, as indicated by by a decreased T_2. Calculation of the theoretical isoelectric point (pI) of IdmK using ExPASy ProtParam (Gasteiger *et al.*, 2003) gives a pI of 5.82. At pH 7.4 IdmK would have an overall positive charge and the Pf1 bacteriophage has an overall negative charge, explaining the interaction between the two molecules. This alignment medium was therefore unsuitable for determination of the RDCs and a new alignment medium was required.

The second alignment medium tried was a *n*-alkyl-poly(ethylene glycol)/*n*-alkyl alcohol mixture (Section 2.2.5.5) (Rückert, 2000), if these molecules are mixed in

the appropriate concentration they create a liquid crystalline medium that partially aligns the protein. This liquid crystalline medium is not charged, and is relatively insensitive to pH and applicable over a range of temperatures (Rückert, 2000). The nomenclature for the *n*-alkyl-poly(ethylene glycol)/*n*-alkyl alcohol mixture is C*m*E*n*, denoting the number of carbons in the alkyl group (*m*) and the glycol groups in the PEG (*n*) (Rückert, 2000). A C12E6, hexanol mixture (0.64 molar ratio) containing 5% PEG was selected for the alignment of IdmK. An initial ^{1}H-^{15}N HSQC spectrum was collected of IdmK in anisotropic conditions and no significant chemical shift differences were observed in the in this alignment medium. A J modulated series was then carried out. Figure 5.15 shows the how the raw data are fitted to calculate the RDCs from the change in peak intensity (Section 2.2.5.5, Equation 2.3) (Tjandra *et al.*, 1996).

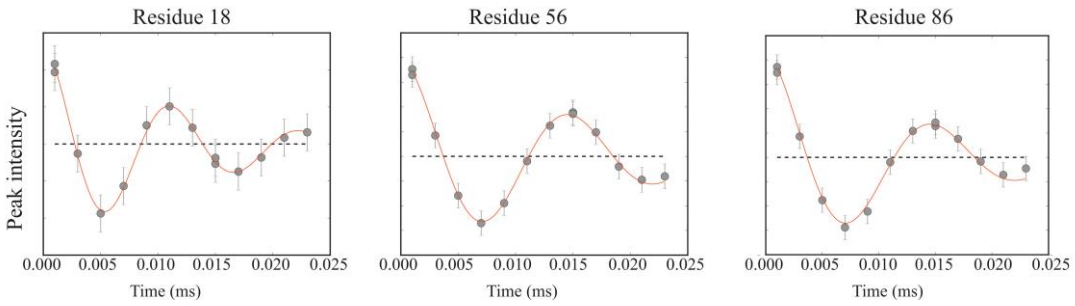


Figure 5.16- Raw RDC data of three residues from IdmK. The Raw RDC data shows a change in the peak intensities in the J modulated series as a function of time. A cosine function is fitted to extract 2 × the dipolar coupling for a residue. Error bars represent propagated errors calculated from duplicate measurements for each peak.

The RDC value extracted from the raw data by fitting a modified cosine function (see materials and methods (2.2.5.5) using the in-house python script FitR_bs (Dr Gary Thompson) which uses appropriate heuristics to determine initial values for fitting. Statistical errors were analysed using duplicate values and Monte Carlo simulations (Tjandra *et al.*, 1996). RDCS can then be plotted against residue number, for comparison with RDC values calculated from models. Figure 5.17 shows the plot of all RDCs measured for IdmK for each residue.

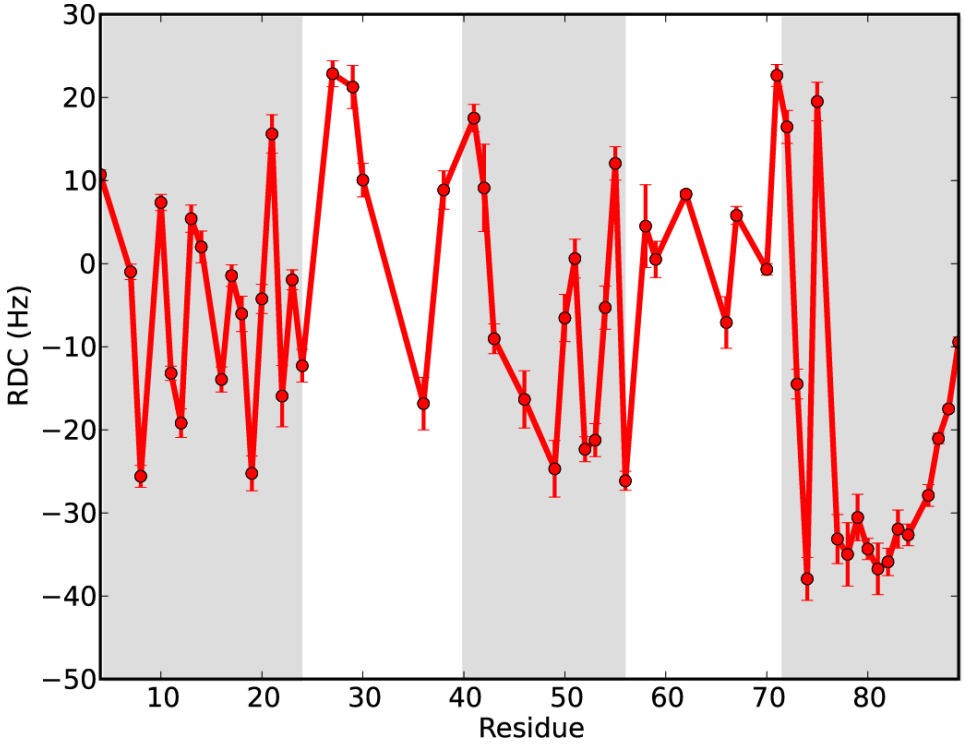


Figure 5.17- Measured RDCs for IdmK collected in 50 mM Tris.HCl pH 7.4, 0.2% (w/v) sodium azide in a C12E6, hexanol mixture (0.64 molar ratio) containing 5% PEG, at a concentration of 0.5 mM *apo*-IdmK recorded at 500 MHz at 25 °C. RDC data collected for IdmK. The grey boxes designate secondary structure elements. Error bars represent propagated errors calculated from duplicate measurements for each peak.

PALES, Prediction of alignment from structure, is a program that can fit the RDCs to a known structure (Zweckstetter, 2000), in this case the ROSETTA models. The measured RDCs define an average alignment of the molecule in three orthogonal directions, defining the alignment tensor frame (Zweckstetter, 2000). To do this the structures from ROSETTA calculation were used. To validate the CS-ROSETTA models generated, the 10 best CS-ROSETTA models were analysed using PALES in order to compare the experimentally determined RDCs with those predicted from the models. For validation of the models generated by CS-ROSETTA only RDCs for the secondary structure elements as determined by TALOS+ were used. This is because residues in loop regions tend to be less well defined and do not have well defined angles in relation to the rest of the molecule and the aligning medium. The RDC waves in figure 5.18 indicate the presence of α-helices (Mesleh *et al.*, 2002).

Figure 5.18 shows the three models that agree the best with the experimentally determined RDCs for secondary structure. All structures show the canonical fold of a carrier protein (Figure 5.5 and 5.6). Each model is ranked by their Q factor, which indicates the quality of the fit of the experimental RDCs to the model. Equation 2.3 illustrates how the Q factor is calculated to determine the overall quality of the fit of the data. A low Q factor indicates a better fit (Zweckstetter, 2000).

The structure with the lowest Q factor (Figure 5.18, top panel) was not the lowest energy ROSETTA model. Although the 2nd model selected by RDCs (middle panel) has a higher Q factor, it is clear from the RDC plots (right) that the secondary structure elements agree better with the experimentally determined data (red and blue lines overlay).

Table 5.3 shows the Q factors calculated for the 10 best CS-ROSETTA, including those in Figure 5.18.

Model	Q factor
1.1	0.479
1.2	0.527
1.3	0.528
1.4	0.544
1.5	0.561

Table 5.3- Q factors of five best CS-ROSETTA models. The best Q factor can be calculated from the experimental errors, in this case the best Q factor for the dataset is 0.1 (Tang *et al*., 2008; Karamanos *et al.*, 2014).

The best Q factor as calculated from the experimental data set for IdmK is 0.1, the Q factors states in Table 5.3 are 0.479 and above. Clearly, the Q factor calculated for illustrates that these models still require refinement. However, the pattern of the RDCs is comparable.

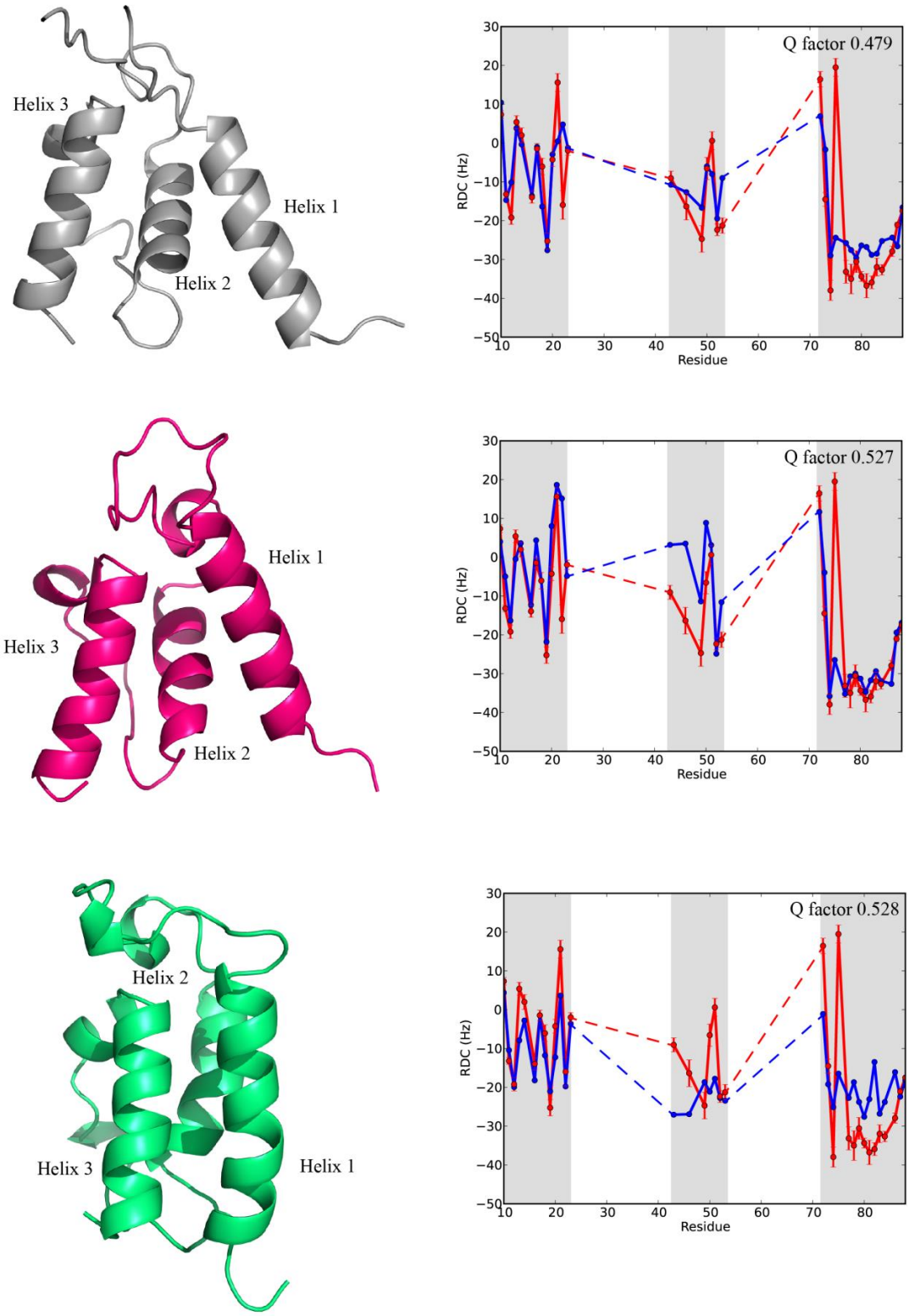


Figure 5.18- Three CS-ROSETTA models selected by RDCs. (Top panel) model 1.1, (middle panel) model 1.2 and (bottom panel) model 1.3 are the three models with calculated RDCs that agree the best with experimentally determined RDCs, as indicated by the Q factor. Dotted lines represent the residues in loops that were not used in fitting. Blue line is the expected RDCs calculated from the models. Red lines are the experimentally determined RDCs. Error bars represent propagated errors calculated from duplicate measurements for each peak.

The agreement of RDCs predicted from the models with the experimentally determined RDCs, and the overall fold predicted by ROSETTA are strong indications that IdmK probably has the canonical fold as seen for other carrier proteins. Figure 5.19 illustrates the differences in angles of the helices, comparing model 1.2 (Figure 5.17) with a carrier protein from a FAS. While helix 2 and 3 appear to be in the correct, same, orientation as the canonical fold, helix 1 appears to be rotated by 90°.

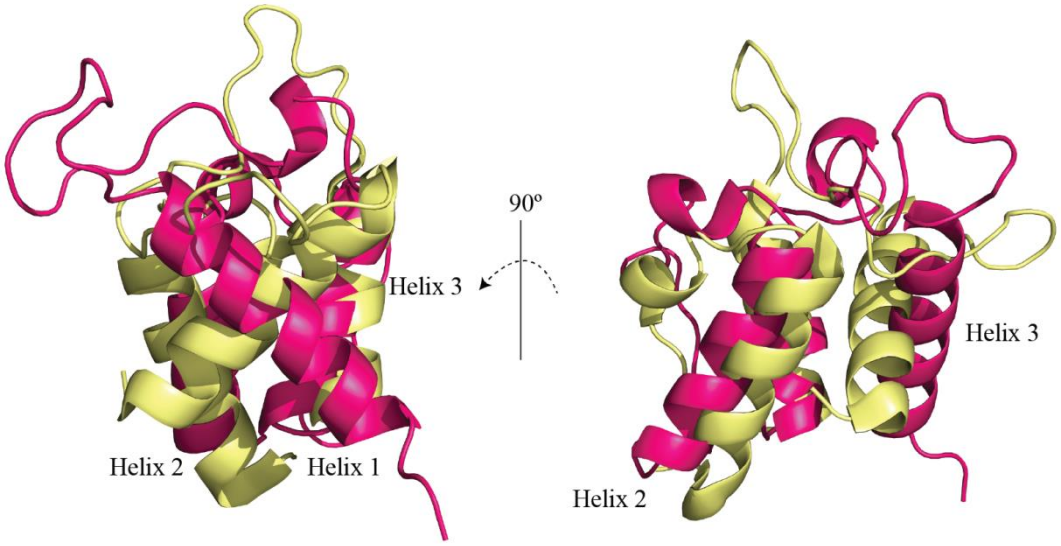


Figure 5.19- Comparison of the best CS-ROSETTA model with an ACP. Overlay of model 1.2 of IdmK (pink), selected as the best CS-ROSETTA module, of the 5 lowest energy models, this was selected due to the agreement with experimentally determined RDCs (Figure 5.12, Table 5.3) with an ACP from FAS (yellow) (PDB code: 2AVA) (Zornetzer *et al.*, 2006).

To guide the ROSETTA structure calculation, and taking advantage of further features of the CS-ROSETTA software package, instead of scoring the structures that were produced by CS-ROSETTA with chemical shifts alone the HN RDCs were also incorporated as additional restraints in the calculation. Figure 5.20 shows an overlay of the five new CS-ROSETTA structures that agreed the best with the RDC restraints for secondary structure regions. Table 5.4 shows the Q factors for these structures.

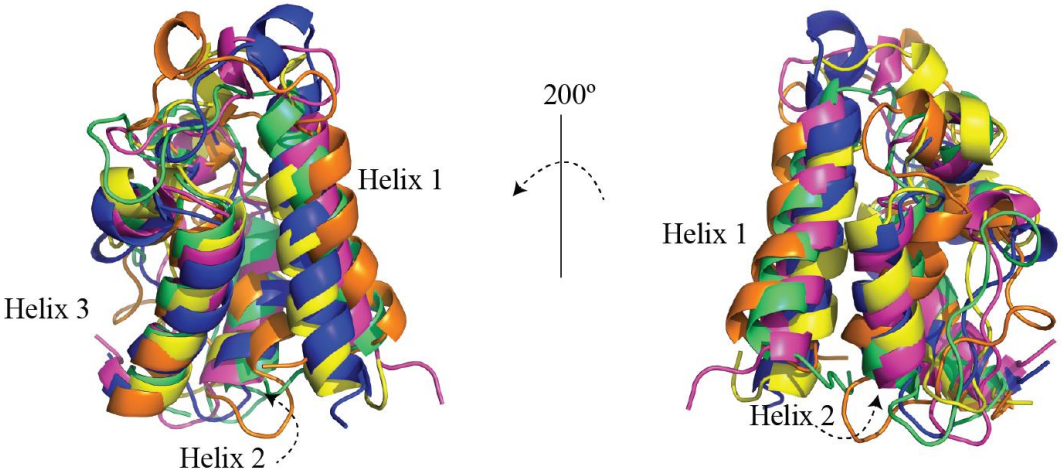


Figure 5.20-Overlay of ROSETTA structures generated with RDCs as an additional restraint. 5 models selected that best fit the RDC data as well as sharing the same secondary structure, as judged by backbone RMSD Table 5.4 shows Q factors related to RDC fits

Model	Q factor
2.1 (Green)	0.411
2.2 (Blue)	0.460
2.3 (Pink)	0.487
2.4 (Yellow)	0.503
2.5 (Orange)	0.533

Table 5.4- CS-ROSETTA structure Q factors. Q factors for secondary structure elements of CS-ROSETTA structures generated with RDCs as a restraint in the calculation. The best Q factor can be calculated from the experimental errors, in this case the best Q factor for the dataset is 0.1 (Tang *et al*., 2008; Karamanos *et al.*, 2014).

This round of structure calculations by CS-ROSETTA produced models comparable to the initial CS-ROSETTA calculation, as illustrated by the Q factors (Figure 5.17). As expected the fit to the RDC data is improved. However, within these ensembles the exact angles of the helices still appears to deviate from the canonical fold, and from each other somewhat. The 20 structures that showed the best agreement with the structure selected using RDCs were then taken forward to seed an ARIA NOE based structure calculation (section 5.4).

5.4 NOE restraints and ARIA structure calculation

In order to generate a higher resolution structure ^{1}H-^{1}H NOE distance restraints were required. In order to assign the NOE distance restraints in NOESY spectra, full amino acid side chain assignments were required. HccoNH and HBHAcbcacoNH spectra of IdmK were collected. These spectra provide a link between the ^{1}H-^{15}N HSQC and and ^{1}H shifts in the ^{1}H-^{13}C HSQC, and were used in combination with the HNCA and HNcoCA to assign cross peaks in the ^{1}H-^{13}C HSQC, this provided a starting point for side chain assignments. Structure calculations by ARIA were carried out by Dr Gary Thompson.

5.4.1 Aliphatic side chain assignments

To assign the aliphatic side chains, especially the methyl bearing side chains such as leucine, isoleucine, valine, threonine and alanine, ^{13}C-HSQC (Cavanagh *et al.*, 2007), hCCH-TOCSY and HcCH-TOCSY spectra were collected (Bax *et al.*, 1990; Olejniczak *et al.*, 1992). A ^{13}C HSQC spectrum detects all of the carbon nuclei in the protein through one bond H-C correlations and is the carbon equivalent to the ^{1}H-^{15}N spectrum. This spectrum is assigned through the use of hCCH-TOCSY and HcCH-TOCSY spectra and provides the basis for assigning NOE peaks in ^{1}H-^{13}C-NOESY-HSQC spectra during structure determination. An HcCH-TOCSY spectrum provides chemical shifts for the side chain protons by correlating them with their corresponding ^{13}C nuclei. By identification of the carbon chemical shifts in an hCCH-TOCSY spectrum, navigation to the carbon plane in the HcCH-TOCSY spectrum can match the side chain protons to the corresponding carbons. The Cα and Cβ chemical shifts assigned in the triple resonance experiments can be used as a starting place to identify the rest of the carbon nuclei in the side chain using this approach. Figure 5.21 and 5.22 below show the hCCH-TOCSY and HCCH-TOCSY spectra assignment of Leu85, respectively.

Currently, this assignment process has been carried out for assignment of approximately 77% of all amino acid side chain protons of IdmK, and 60% of heavy side chain atoms of IdmK, including all of the methyl bearing side chains. A chemical shift table can be located in the appendix.

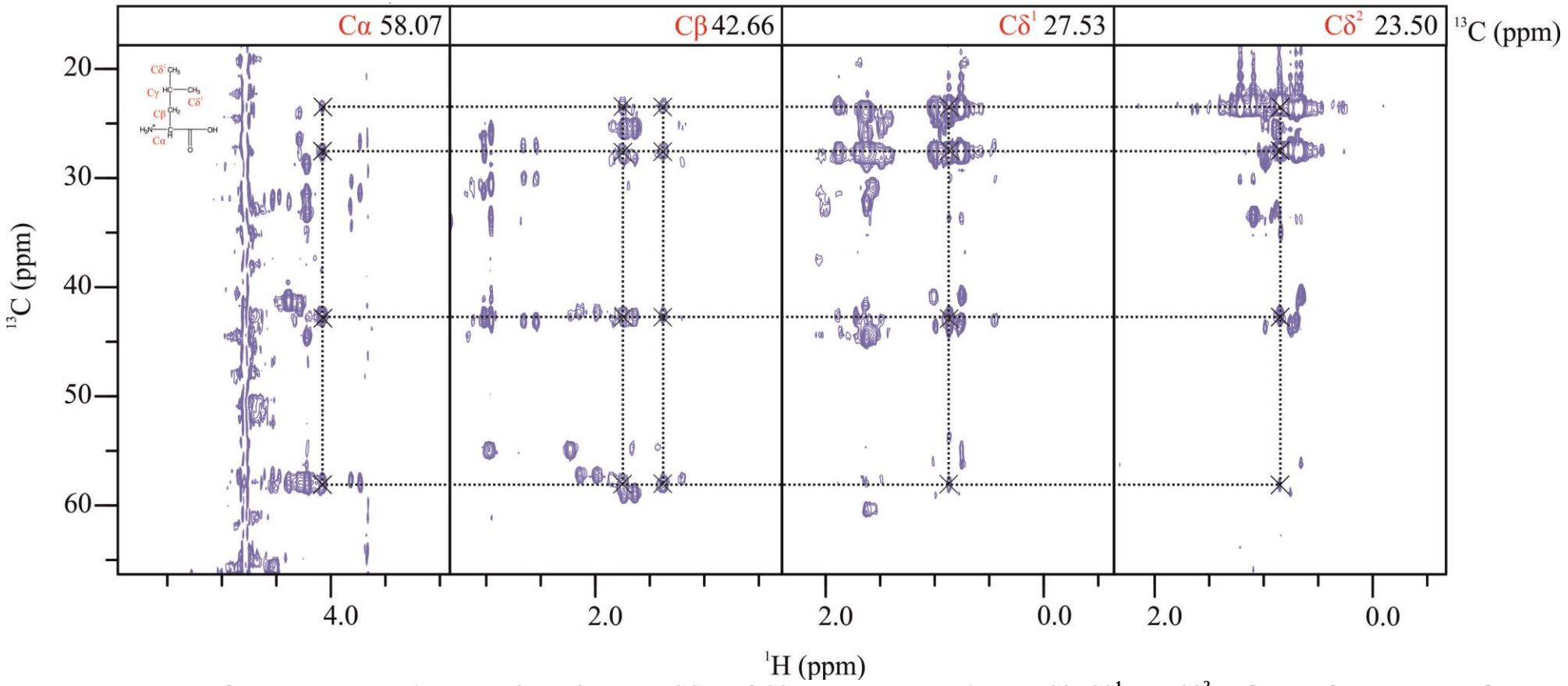


Figure 5.21- hCCH TOCSY assignment of Leu 85. The hCCH TOCSY was used to assign the Cβ, Cδ^1 and Cδ^2 carbon nuclei in Leu 85. The carbon planes for Cβ, Cδ^1 and Cδ^2 were identified in strip 1, where the peaks marked are the C*α*, Cβ, Cδ^1 and Cδ^2 (bottom to top) identified by characteristic chemical shifts. The dotted lines indicate corresponding carbon peaks in another carbon plane. Spectrum was collected in 50 mM Tris.HCl pH 7.4, 0.2% (w/v) sodium azide with *apo*-IdmK at a concentration of 0.5 mM.

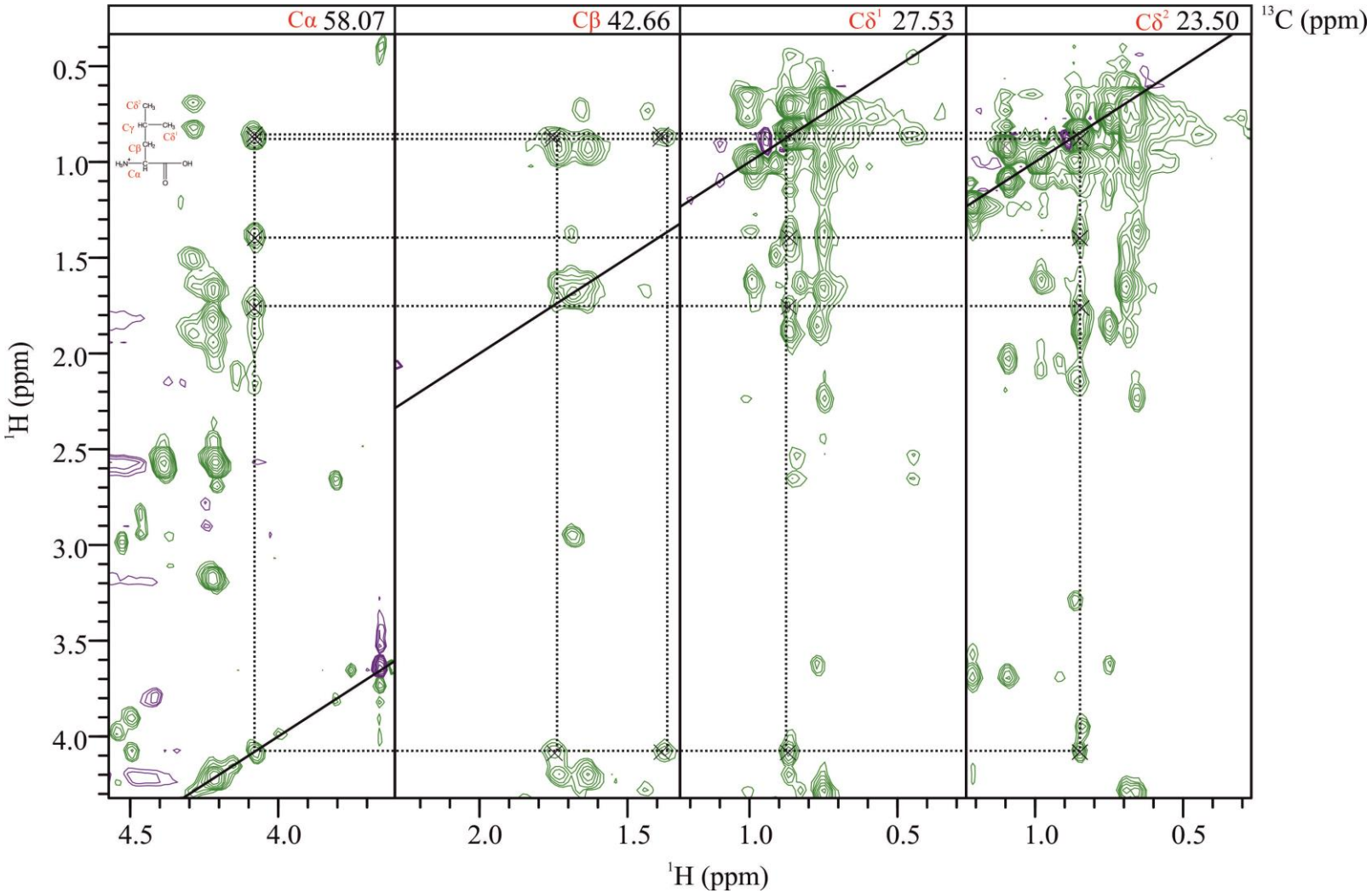


Figure 5.22- HcCH TOCSY assignment of Leu 85. The HCCH TOCSY was used to assign protons in the amino acid side chain. The hCCH-TOCSY was used to identify the carbon chemical shifts for the C*α*, Cβ, Cδ^1 and Cδ^2. Navigation to these carbon planes in the HCCH-TOCSY enabled identification of the corresponding protons. The dotted lines indicate the same protons in another carbon plane. Spectrum was collected in 50 mM Tris.HCl pH 7.4, 0.2% (w/v) sodium azide with *apo*-IdmK at a concentration of 0.5 mM.

5.4.2 Aromatic side chain assignments

The amino acid sequence of IdmK contains four aromatic residues, not including the His_6-tag, three phenylalanine residues and one tryptophan. Aromatic residues usually form part of the hydrophobic core of a protein and provide useful NOE restraints for structure determination (Prompers *et al.*, 1998). The methodology used for assignment of aromatic residue side chains differed from assignment of the other amino acid side chains. To assign the phenylalanine residues in IdmK a combination of a 2D hbCBcgcdHDceHE spectrum and a 2D aromatic ^{13}C filtered ^{1}H-^{1}H NOE spectrum were used (Yamazaki *et al.*, 1993; Marion *et al.*, 1989a; Marion *et al.*, 1989b; Cavanagh *et al.*, 2007). The hbCBcgcdHDceHE exploits the chain of one bond (J coupling) between the Cβ chemical shifts and the Hδ and Hε in the connected ring making it possible to assign the Hδ protons and Hε protons of the aromatic side chains. However, this spectrum has relatively low resolution in the ^{1}H dimension due to the use of constant time and power limits imposed by the pulse sequence and spectrometer (Yamazaki *et al.*, 1993). In addition to this the peaks in this spectrum can also be masked by those generated by the mobile His_6-tag and the overlap of Hβ chemical shifts in IdmK. To combat these problems a ^{1}H-^{1}H-aromatic ^{13}C filtered NOE spectrum was also collected, this experiment uses the ^{1}H-^{1}H NOEs observed between the Hβ and Hδ protons for assignment. The Hβ and Hδ NOE cross peaks are always, at least weakly, present due to the restraints of the covalent geometry of the ring and side chain. To assign the tryptophan residue in IdmK, short ^{1}H-^{1}H NOE distance restraints were used in addition to assignment of the characteristic chemical shifts in the ^{1}H-^{13}C aromatic optimised HSQC. This resulted in assignment of all four of the non-histidine aromatic side chains.

After assignment of of the amino acid side chains of IdmK had begun, once a good proportion had been assigned, especially those suspected to form the hydrophobic core, were used to peak pick the ^{13}C-NOESY-HSQC. In addition the ^{1}H-^{15}N HSQC was used to peak pick the ^{15}N-NOESY-HSQC (Marion *et al.*, 1989a; Marion *et al.*, 1989b; Zuiderweg and Fesik, 1989).

5.4.3 Progress towards a high resolution solution structure of IdmK

To calculate a higher resolution structure of IdmK the program ARIA, Ambiguous restraints for iterative assignment, was used (Rieping, 2007). Historically, NMR structure calculations were performed using unambiguously assigned NOE cross

peaks, converting them into distance restraints. However, the manual unambiguous assignment of cross peaks in NOESY spectra is time consuming and inaccurate (one peak may contain several distances due to resolution limitations). To deal with this issue, structure calculations have relied on iterative process where an initial 3D structure model created from the unambiguous NOEs is used to seed further NOE assignments which can be obtained by quantifying the agreement of the ambiguous NOEs with a model (Nilges, 1995; Teng, 2005). ARIA uses an automated iterative version of this strategy, however all possible assignments from an ambiguous NOE are considered for the initial structure calculation. The correct assignments of NOEs is achieved through several rounds of simulated annealing energy minimisation interspersed with automated assignment. Typically calculations start with a structure with random coordinates i.e. an unfolded polypeptide chain. During each round following an increasingly rigourous selection of assignments based on the current model the improved restraint set is used to calculate further structures, trying to satisfy as many distance restraints as possible. Over 8 rounds, only NOEs that agree with the structure generated in the previous round are selected and therefore the population of incorrect NOE assignments decreases leading to convergence (Rieping, 2007; Fossi *et al.*, 2005a).

The ARIA structure calculation of IdmK was carried out using ^{1}H-^{1}H NOEs from the filtered ^{13}C-NOESY-HSQC and ^{15}N-NOESY-HSQC spectra. The ^{1}H-^{13}C-HSQC, assigned by the HCCH- and hCCH-TOCSY experiments, was used as the root spectrum for the ^{13}C-HSQC shifts of NOE peaks used in the ARIA calculation. ^{1}H NOE chemical shifts were assigned using the ARIA protocol during the structure calculation. A total of 1123 distance restraints were obtained for IdmK, this includes long and short range NOEs as summarised in Table 5.5.

Distance constraints		IdmK
Total NOE		1123
Intraresidue		449
Interresidue	i+1	317
	i+2-4	184
Long range		173
Violations		28

Table 5.5- Unambiguous distance restraints assigned by ARIA during the structure calculation of IdmK.

To assist the convergence of the structures generated by ARIA rather than starting from a random structure, the first round of the calculation (it0) was seeded with the CS-ROSETTA structures generated when RDCs were included as a restraint (Figure 5.20). 20 structures were used to seed the first round of the calculation, and 20 structures were carried through to the next round of the calculations (it1-8). Refinement in a bath of TIP3 water was carried out for 10 structures from it8 to give the final ensemble. During the calculation, the hot and slow cooling stages were extended by a factor of four as described in (Fossi *et al.*, 2005c) in an effort to aid convergence. As mentioned previously, +TALOS phi and psi angles derived from the Cα and Cβ chemical shifts for residues shown to form the α-helices were included in the calculations.

To further improve convergence in the first round of structure calculations only the HN chemical shifts were used and the ^{1}H-^{15}N NOEs initially assigned. In this round of calculations a structure composed of three α-helices was generated, however there was no consistent overall tertiary fold. The ^{1}H-(^{15}N)-^{1}H NOEs assigned in this round of calculations for the secondary structure elements were fixed in subsequent calculations, this included 64 unambiguous and 3 ambiguous distance restraints, each of which had two ambiguous assignments (Fossi *et al.*, 2005a).

The second round of calculations used a subset of assignments, specifically those of of the methyl groups from Ile, Leu Thr and Val methyl protons and shifts Hα and Hβ protons in order to separate the spectrum into two sub-spectra for which clear and confident assignments could be made. Any other assignments within the two sub-

spectra were assigned to prevent miss-assignment of chemical shifts. The regions used were 6-3.5ppm and 1.3- -1ppm.

Finally, for the calculations initial weak NOEs restraints of those identified by the CS-ROSETTA models, between secondary structure elements, were introduced into the calculation and used with their upper bounds extended by 5 Å. Any restraints between the secondary structure elements that were assigned and violated by greater than 10 Å were removed and flagged for further investigation. RDCs were not used in the second round of calculations and so were able to be used for subsequent structure validation. Figure 5.23 shows the ensemble of IdmK structures generated by ARIA.

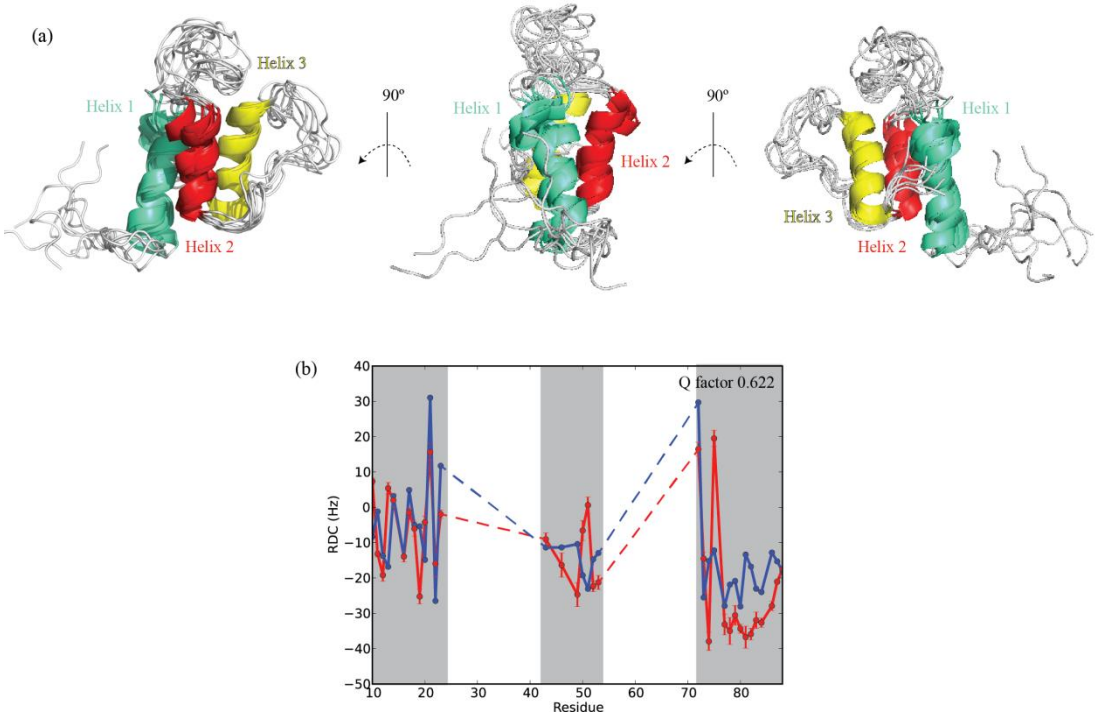


Figure 5.23- Structure ensemble of IdmK calculated by ARIA and validation. (a) Overlay of the 10 lowest energy structures generated by ARIA through unambiguous assignment of NOE distance restraints. (b) Structure ensemble was validated by comparison of the calculated RDCs against experimentally determined RDCs.

The ARIA structure ensemble shows a three α-helical bundle, with helix one (residues 6-19) running antiparallel to helix two (residues 44-54) and three (74-86). The conserved serine required for post-translational modification is located at the top of helix 2. Contrary to the CS-ROSETTA structure, the ARIA structure does not have the small-helix (or turn) seen in the long loop between helices 2 and 3. In carrier proteins this is termed as a pseudo helix, and is variable between structures,

this may be because of the limited restraint set used. This fold shows the same topology of the helices and loops as the canonical fold; however the meanders are not well ordered between the helices and relative angles of the helices to each other vary somewhat from those seen In the canonical fold of the carrier proteins (Figure 5.5). Figure 5.24 shows a comparison of the IdmK structure generated by ARIA in comparison with an acyl carrier protein from FAS.

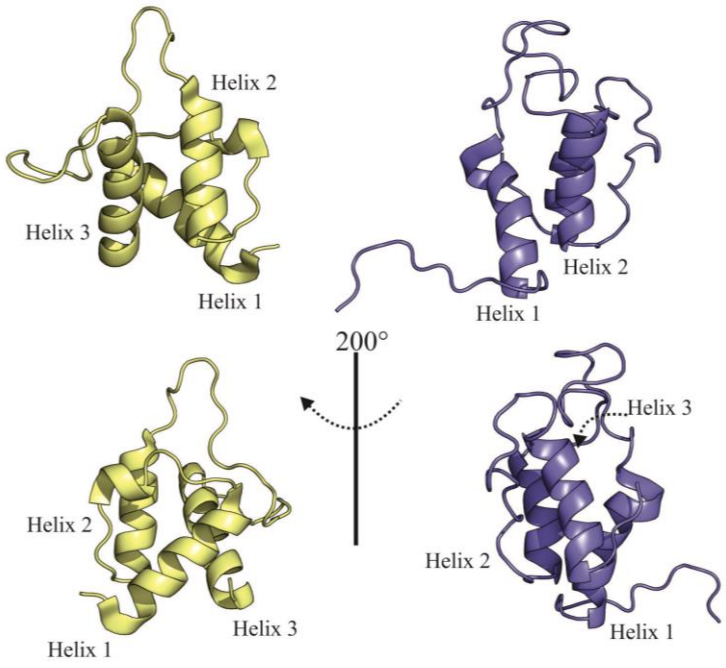


Figure 5.24- Comparison of the 3D structures of acyl carrier protein and IdmK. View of the 3D structures of 2AVA (yellow) with the best ARIA structure (blue) showing to compare orientation and positioning of helices (Zornetzer *et al.*, 2006).

Although the Q factor for the ARIA ensemble indicates that the agreement with the measured RDCs is slightly worse than that seen for the CS-ROSETTA structures, ARIA has been able to improve the convergence of the ensemble structures generated. This is shown by a lower RMSD of the secondary structure elements, an average of 0.66 Å compared with 3.43 Å of the RDC restrained CS-ROSETTA structures. The initial structure calculated with ARIA provides a solid starting point for further refinement of the structure and introduction of further NOEs to achieve a complete structural model of IdmK. Figure 5.25 shows the short range NOE assigned by ARIA. The resulting pattern of short range NOEs are in good agreement with the location and lengths of the secondary structure elements.

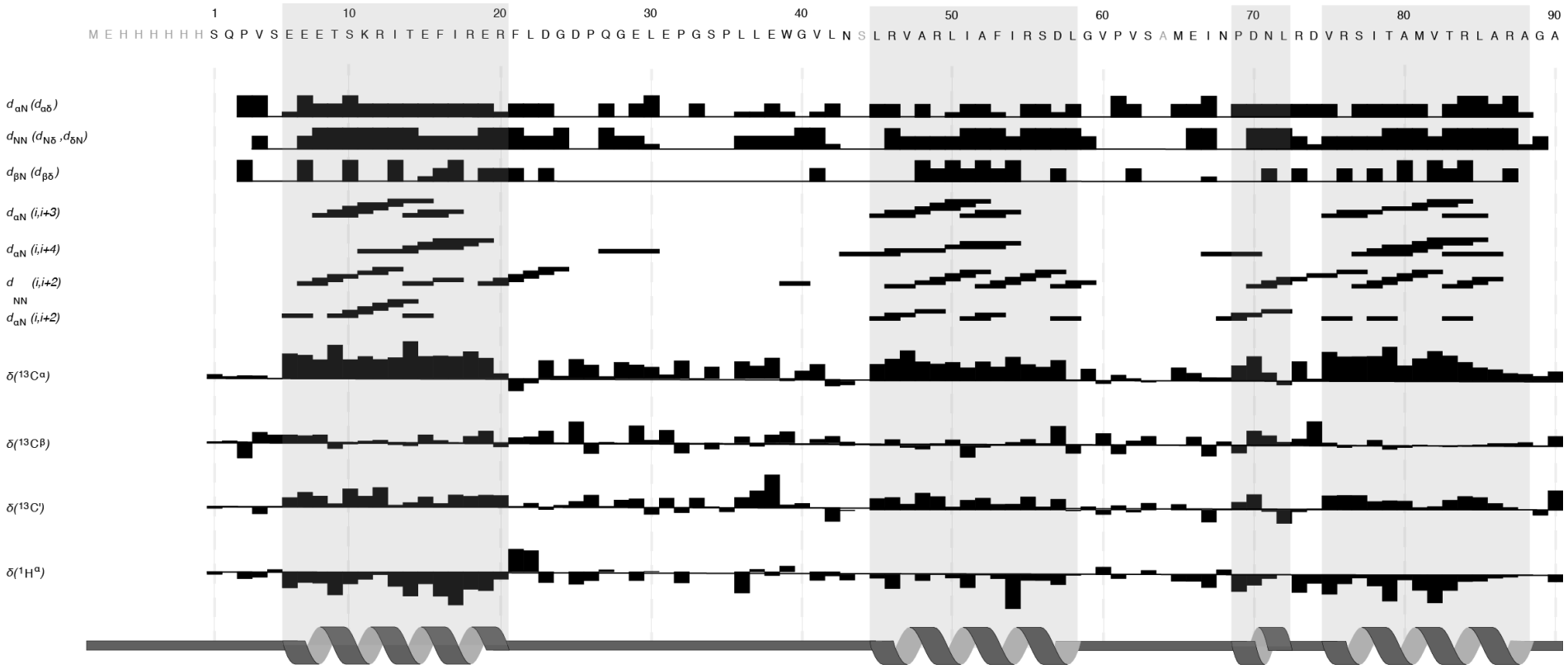


Figure 5.25- Plot showing short range NOEs assigned by ARIA. Comparison of the short range NOEs assigned by ARIA with the secondary structural elements predicted from the chemical shifts by TALOS show good agreement for the location and lengths of helices.

5.5 Initial phosphopantetheine characterisation

As with other carrier proteins, phosphopantetheinylation of IdmK is required for function. Chapter 4 showed that when IdmK is expressed in *E. coli* it is phosphopantetheinylated by an endogenous PPTase (Chapter 4, Figure 4.7). Furthermore, it was shown that it was possible to purify *holo*-IdmK (Chapter 4, Figure 4.6 and 4.7). In order to elucidate the mechanism of this module, the next step taken was to investigate structural changes observed in *holo*-IdmK, and investigate the location of the modification. Chemical shifts provide a sensitive monitor of changes in structure and dynamics (Teng, 2005).

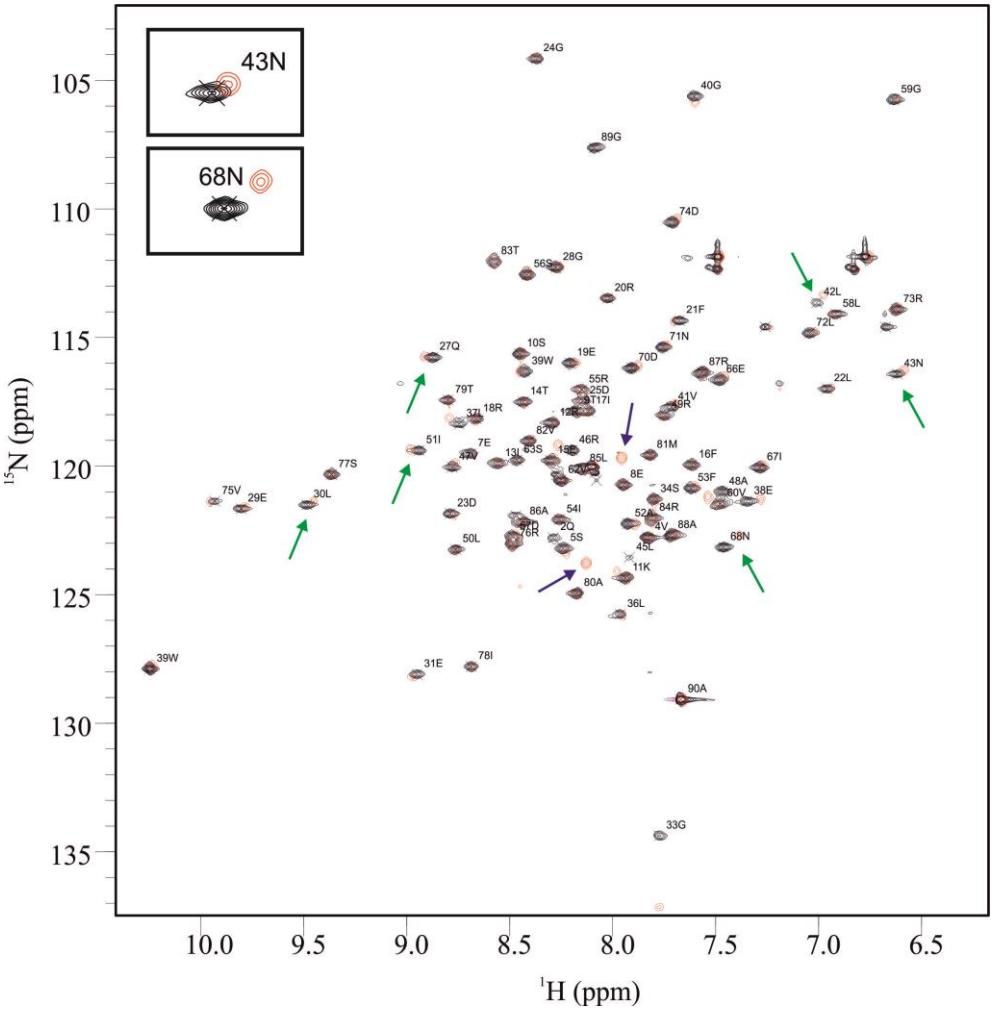


Figure 5.26- An overlay of ^{1}H-^{15}N HSQC spectra collected in 50 mM Tris.HCl pH 7.4, 0.2% (w/v) sodium azide of *apo*- and *holo*-IdmK. Chemical shift differences of the *apo*- (black) and *holo*- (red) spectra are shown by peak shifts, which indicates changes in the local environment of the nuclei. Green arrows identify peaks that have altered chemical shifts in the *holo*- spectrum. The blue arrows identify two new peaks observed in the *holo*- spectrum. Insert illustrates peak shifts observed in the overlay of the two spectra.

Superimposition of the ^{1}H-^{15}N HSQC spectra of the *apo*- and *holo*- states of IdmK identified two new peaks (blue arrows), indicating the presence of two new amides in the protein, comparable to the structure of phosphopantetheine though two amides are also missing from the apo spectrum. Chemical shift differences, characterised by peaks shifts were also observed for some other residues. Figure 5.27 shows a graph of the differences of chemical shifts between the *apo*- and *holo*- IdmK, identified using conservative shift mapping (Williamson *et al.*, 1997) (section 2.2.5.6 Equation 2.5) and those above 0.2 ppm were mapped onto the most recent, ARIA, structure for reference.

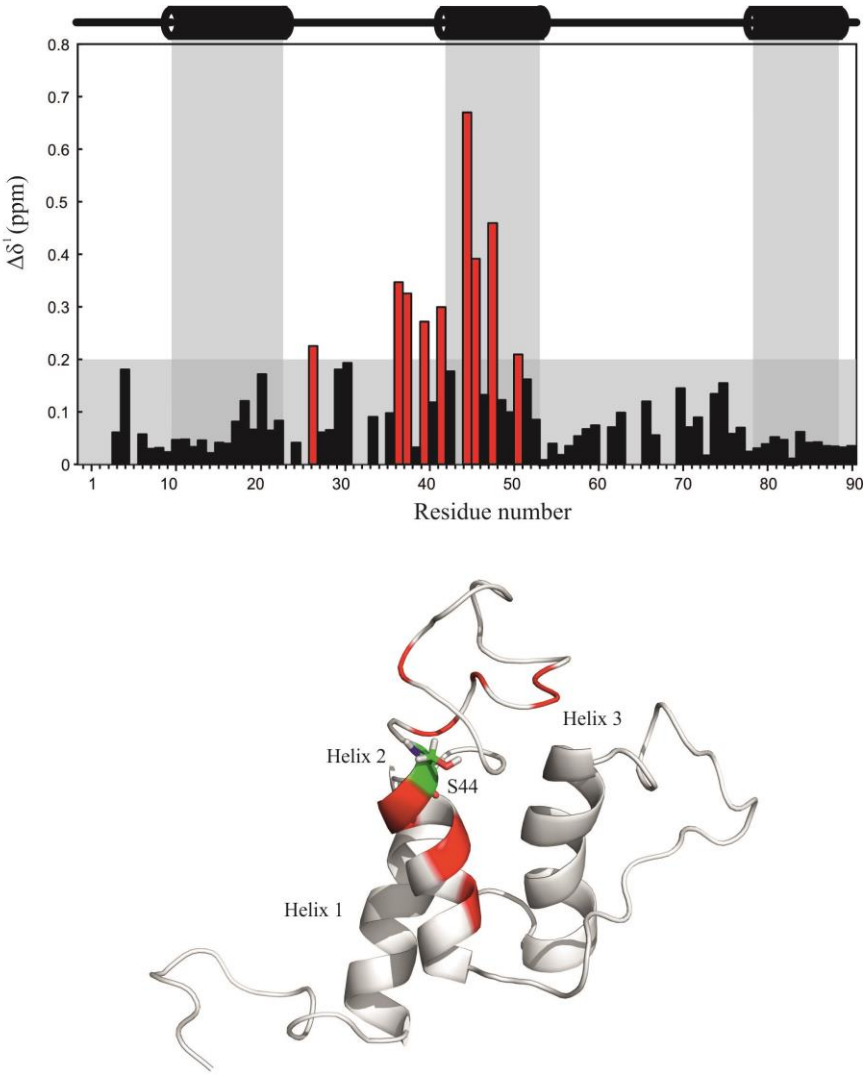


Figure 5.27- Chemical shift perturbations of *apo*- and *holo*- IdmK. The graph shows the chemical shift differences measured from the ^{1}H-^{15}N HSQSC spectra of *apo*- and *holo*- IdmK. Chemical shift differences were then mapped onto the best ARIA structure. The red patches indicate residues with chemical shift differences ≥0.2 ppm. The residue highlighted in green is the conserved serine that gets phosphopantetheinylated.

When the chemical shift differences observed in the ^{1}H-^{15}N HSQC spectra of *apo*- and *holo*-IdmK were mapped onto the *apo*- ARIA structure of IdmK, it indicated that the phosphopantetheine was lying down the side of helix 2, directed into the core of the protein. Chemical shift differences were also observed in residues within the loop between helix 1 and 2. For complete validation of the structure of IdmK with phosphopantetheine a full assignment of the phosphopantetheine moiety is required, and in addition a full set of NOE restraints need to be measured before the full *holo* state of the protein can be determined. However, this initial data suggests that the phosphopantetheine may be more occluded with a different orientation to those previously observed, unless major movements in the secondary structure elements occur (Johnson *et al.*, 2014; Nguyen *et al.*, 2014; Masoudi *et al.*, 2014).

5.6 Summary

Structure determination by NMR spectroscopy began with collection of a ^{1}H-^{15}N HSQC spectrum. This spectrum indicated that *apo*-IdmK would be suitable for structure determination by NMR spectroscopy. Backbone assignments of IdmK were completed by collection of triple resonance spectra, and using the semi-automated assignment program MARS (Jung and Zweckstetter, 2004). MARS was able to assign approximately 92% of the protein backbone; the MARS output was manually checked using strip plots (Figure 5.8). Assignment of the protein backbone provided chemical shifts for $^{1}H^{N}$, ^{15}N, $^{13}C\alpha$, ^{13}CO and $^{13}C\beta$ nuclei. These chemical shifts were then used in a structure calculation by CS-ROSETTA to generate an overall fold for IdmK (Shen *et al.*, 2008; Shen *et al.*, 2009b). The models generated were validated by RDCs measured by collection of a J modulated series. CS-ROSETTA models were also generated using RDCs as a restraint in the calculation. Models generated by CS-ROSETTA showed the observed topology expected for a carrier protein (Figure 5.6), however the angles of the helices and relative positioning differed from the observed expected canonical fold of other carrier proteins (Figure 5.19). RDC data (Figure 5.18) indicated that the secondary structure elements in the models agreed with experimental data collected. In an effort to elucidate a higher resolution structure short and long range NOEs were used. Side chain assignments were completed for 60-77% of side chains atoms of residues, assisting in assignment of

NOE distance restraints by ARIA (Rieping *et al.*, 2007). Structure calculations carried out by ARIA generated a converged ensemble. The structures showed a three-helix bundle with the expected topology for carrier proteins. The ARIA structures satisfy NOEs expected from the canonical fold, the orientations of the helices are comparable to the CS-ROSETTA structures. Although the fit of the RDC data is slightly worse than the CS-ROSETTA structures this data provides a strong foundation for determination of a high-resolution structure, using ARIA.

5.7 Discussion

Initial analysis of the solution structure ensemble of IdmK generated by ARIA with other carrier proteins from fatty acid synthases and polyketide synthases has inferred that IdmK adopts an almost canonical carrier protein fold differing in the orientations of the helices, although more work needs to be done to complete the models generated. Investigations into the protein dynamics of peptidyl carrier proteins has been carried out on the TycC3 PCP. Figure 5.27 shows the multiple conformers this carrier protein was seen to adopt in the *apo*- and *holo*- forms, the A/H state is seen in both *apo*- and *holo*- PCP states. The first solution structure (PDB code: 1DNY) of the TycC3 PCP is comparable to the A/H state elucidated some years later (Weber *et al.*, 2000; Koglin *et al.*, 2006). The PCP from the tyrocidine biosynthetic gene cluster appears to adopt different conformations, with the *holo*-state being the most expanded, and the A/H state being the most compact. Variations in helix angles with respect to each other differ between states, as does the length of helices. The small helix/turn between in the A/H state is also missing in the A and H states (Koglin *et al.*, 2006). The dynamic nature and more open conformation of *holo*-TycC3-PCP may be to assist in protein-protein interactions and access of the phosphopantetheine modification to other domains and downstream modules to elongate the growing peptidyl chain.

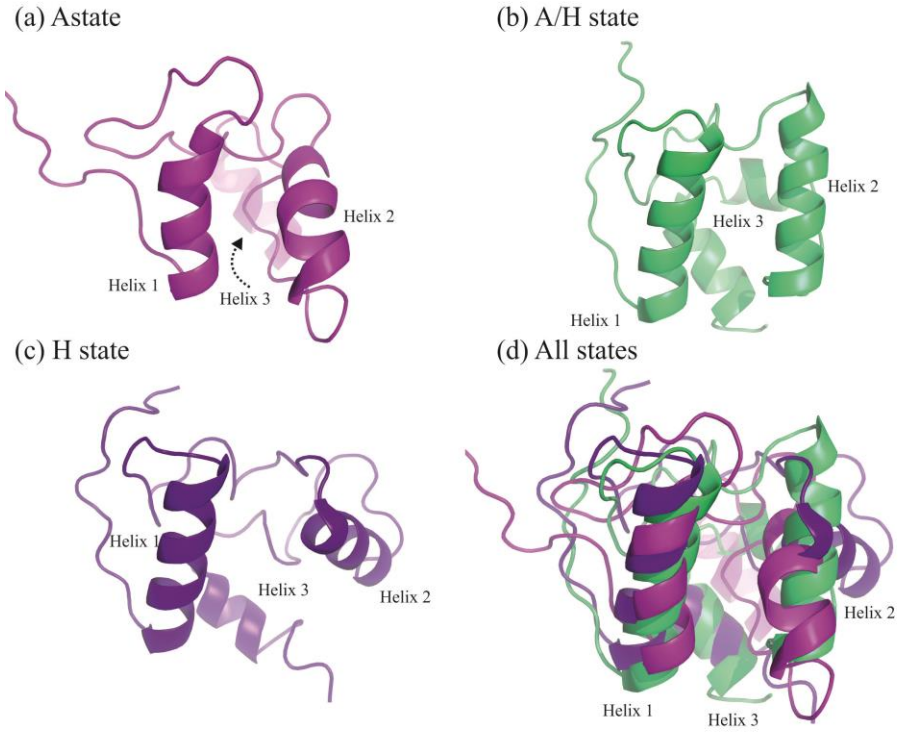


Figure 5.28- Conformational states of the TycC3-PCP. (a) and (c) are the conformations TycC3-PCP adopts when unmodified and phosphopantetheinylated, respectively (PDB codes: 2GDY and 2GDX) (Koglin *et al.*, 2008). The A/H state, equivalent to the TycC3-PCP structure 1DNY, is an equilibrium state seen in both *apo*- and *holo*- states of TycC3-PCP (Weber *et al.*, 2000; Koglin *et al.*, 2006). (d) Overlay of three structures illustrating deviations in helix angles and lengths.

So far, there has been no obvious evidence of IdmK being as dynamic in nature as the TycC3-PCP. However, collection of perhaps a [^{15}N, ^{1}H]-TROSY spectrum and full characterisation of the phosphopantetheine co-factor and *holo*-state of IdmK may shed some light on whether IdmK is as dynamic, as the TycC3-PCP (Koglin *et al.*, 2006).

Characterisation of the phosphopantetheine modification and *holo*-state of the carrier protein is in its preliminary stages. Results so far indicate that the phosphopantetheine may be positioned down the side of helix 2, directed into the hydrophobic core of the protein. Further validation of the current ARIA models is required before assuming any definitive conformations of the phosphopantetheine modification. Assignment of the phosphopantetheine can be carried out using a standard triple resonance HNCA experiment that has already been collected, in addition to collection of ^{1}H-^{15}N-NOESY data. Recently, vibrational spectroscopy and crystallography were able to probe the movements of the phosphopantetheine arm in a *holo*-ACP, showing that prior to being loaded with its substrate the

phosphopantetheine arm is solvent exposed, directed away from the core of the protein, upon acylation large conformational changes were observed in helix 2 and 3, causing dissociation from the loading enzyme, potentially in preparation for subsequent protein-protein interactions downstream (Johnson *et al.*, 2014; Nguyen *et al.*, 2014; Masoudi *et al.*, 2014). It would be interesting to probe the conformational changes that may occur when loading *holo*-IdmK with L-proline, either enzymatically or chemically.

NMR structure calculations are difficult as they require the assignment of the entire protein backbone, amino acid side chains and NOE distance restraints (Teng, 2005). This chapter described the use of MARS to assign the backbone of IdmK. This program provided fast and reliable assignments. MARS is a powerful program for small proteins such as IdmK (Jung and Zweckstetter, 2004). The increasing number of assignments required for structure calculations can be difficult and laborious, especially as confidence in the assignment is important as mistakes can severely affect the outcome (Markwick *et al.*, 2008; Nilges, 1995).

ROSETTA has been used successfully for *ab initio* protein structure prediction for proteins up to 20 kDa (Shen *et al.*, 2008; Sgourakis *et al.*, 2014; Bouvignies *et al.*, 2011). A combination of unambiguous NOE assignments in the ROSETTA structure calculation could and has been used to help higher resolution structure prediction, even for larger proteins (Loquet *et al.*, 2012; Raman *et al.*, 2010; Sgourakis *et al.*, 2011). In the structure determination of IdmK using CS-ROSETTA, RDCs were used as a restraint to aid the structure calculations, producing a carrier protein with the same topology as the canonical fold. The structures produced were equivalent to the CS-ROSETTA structure calculation without the RDCs, showing the high predictive power of the technique.

Assignment of NOESY spectra can be very tedious even for small proteins. ARIA performs an iterative assignment of ambiguous NOEs to determine long range distance restraints for structure determination (Rieping *et al.*, 2007). Although this program has obviously sped up structure determination, it still requires the input of good quality NOESY spectra. Overlap and miss assignment of peaks will hinder the ARIA structure calculation and severely affect the outcome of the structure calculation. More recently, CS-ROSETTA structure calculations have been

completed using RDCs and sparse Ile, Leu and Val NOE restraints, making it possible to calculate a structure with an RMSD of 0.5 Å with respect to the crystal structure (Sgourakis *et al.*, 2014).

Structure determination of IdmK using CS-ROSETTA provided a solid foundation for high resolution structure determination using ARIA. Seeding ARIA calculations with the CS-ROSETTA models, in addition to including some weak initial NOE restraints was able to produce a structure of IdmK in some agreement with other carrier proteins. Further structure refinement is required of IdmK. Once the high-resolution structure is complete, investigation into protein-protein interactions and dynamics of both *apo*- and *holo*- IdmK will provide some insight into peptidyl chain elongation and starter unit construction in indanomycin biosynthesis.

6. Summary, future work and perspectives

6.1 Summary

This thesis has described the successes and difficulties in heterologous expression of the indanomycin nonribosomal peptide synthetase (NRPS) proteins in *E. coli*. Attempts to clone and express the three proteins, IdmJ, IdmI and IdmK, from indanomycin starter unit biosynthesis are presented.

Chapter 1, the introduction to the focus of this thesis, outlined a huge foundation of research into NRPS and PKSs, from their biosynthetic gene clusters to production of novel compounds. It was decided that a heterologous expression of the target module would be the appropriate first step to take in engineering substrate specificity in the domains required for starter unit biosynthesis of indanomycin. This research began by designing and purchasing synthetic genes, codon optimised for expression in *E. coli*. These genes encoded an adenylyltransferase (IdmJ), responsible for selection of L-proline and activating it, a carrier protein (IdmK), responsible for transfer of the substrate to the downstream module, and the dehydrogenase (IdmI), which oxidises the L-proline to a pyrrole. Chapter three presented the issues encountered when attempting to express IdmJ and IdmI in *E. coli* outlining the approaches taken in order to produce soluble protein. Complications encountered with expression of IdmJ and investigation into a potential post-translational modification is outlined in this work. Unfortunately, the measures taken to express IdmI as a soluble protein did not solve the problem. Chapter four discussed the successes of cloning and expressing IdmK followed by the fortuitous discovery that IdmK was also post-translationally modified in *E. coli*. This chapter also described the first attempts at structural characterisation by crystallography, and why NMR was chosen as an alternative for structure determination. Chapter five describes the steps taken for structure determination by NMR and has presented a protein fold elucidated solely from chemical shifts in addition to a higher resolution structure of IdmK. The first steps in characterisation of *holo*-IdmK and the phosphopantetheine are also described. This data so far suggests that IdmK has canonical carrier protein fold, and the phosphopantetheine modification may be pointing into the core of the protein.

Structure determination of IdmK has provided a solid foundation for investigations into dynamics of the NRPS module, and potentially the carrier proteins interactions

with an adjacent PKS module. Including refinement of the carrier protein structure and assignment of the phosphopantetheine modification, there are a number of directions for future work.

6.2 Extension of current work

6.2.1 Expression and purification of the dehydrogenase (IdmI)

Soluble protein expression of IdmI was not achieved; however there are still a number of strategies which could be trialled, including expression in Origami™ 2 (DE3) cells, to aid in disulphide bond formation in *E. coli*. Although genes were synthesised to be codon optimised for expression in *E. coli*, previous dehydrogenases from aminocoumarin biosynthetic gene clusters have been cloned from genomic DNA with a C-terminal His_6-tag (Garneau *et al.*, 2005), therefore this could be attempted. Finally, insoluble aggregates could be unfolded, and refolded for purification.

6.2.2 Purification and characterisation of the adenylyltransferase

Expression and purification of IdmJ resulted in an unknown post-translational modification at C127. Site directed mutagenesis of the cysteine residues resulted in purification of IdmJ variants without the post-translational modification. Further purification by size exclusion chromatography of IdmJ alanine variants (C104A, C127A, C169A and C308A), characterisation of secondary structure by CD and assaying activity using an ATP/PPi exchange assay could provide insight into to whether the alanine variants are folded correctly and functional.

6.2.3 Structural characterisation of the carrier protein (IdmK)

The current solution structure of IdmK requires further refinement and validation; once this has been achieved it provides the basis for investigation into structural changes that may occur when the carrier protein is phosphopantetheinylated or loaded with its substrate. Exploration into protein-protein interactions with a variety of PPTases could be carried out, investigating the binding surface of the protein. In

lieu of the adenylyltransferase, investigation into the use of chemical modification using thiols as an alternative method to chemically load the carrier protein with the substrate, using the method described for conversion of cysteines into dehydroalanine followed by a reaction with a thiol, could be carried out (Chalker *et al.*, 2011; Bernardes *et al.*, 2008; Timms *et al.*, 2013). NMR could be used to investigate structural changes that have occurred as a consequence of *holo*-IdmK being loaded with the appropriate and alternate substrates.

6.3 Future perspectives

Following any successes highlighted above structural and biochemical characterisation could be carried out. Alternatively, investigation into substrate specificity could be carried out by substituting IdmJ with the adenylation domains found in clorobiocin and coumermycin biosynthesis, known to be successfully cloned and expressed in *E. coli* (Garneau *et al.*, 2005). Identification of residues which contribute to substrate specificity and mutagenesis could aid in engineering substrate specificity in the adenylation domains (Stachelhaus *et al.*, 1999). In addition to biochemical assays such as ATP/PPi exchange, NMR could be used to follow the progress of loading the carrier protein with alternative substrates.

IdmK is a peptidyl carrier protein, located in an NRPS module, that is required for the transfer of the starter unit, pyrrole-2-carboxylate, to a downstream PKS module. Investigating the interaction of IdmK with PKS domains could provide insight into communication between NRPS and PKS modules in hybrid enzymes. However, this would require the simultaneous expression of PKS modules, such as the downstream PKS module in indanomycin biosynthesis IdmL (Li *et al.*, 2009).

Combining chemical loading of the carrier proteins via a dehydroalanine or enzymatically through the use of an adenylation domain able to accept alternative substrates, with investigation of the interaction of IdmK with the downstream module IdmL could generate novel products.

6.4 Concluding remarks

This research has provided a basis for investigation into NRPS module dynamics based around the carrier protein using NMR spectroscopy as a tool. Although difficulties were experienced when attempting to clone and express the adenylyltransferase and dehydrogenase, characterisation of the carrier protein has been successful. Structure determination of a carrier protein involved in communication between an NRPS and PKS module could provide insight into this interaction, potentially highlighting differences between the interactions of carrier proteins in NRPS and PKS only multienzymes. This could aid in the efforts to create an NRPS/PKS toolbox for combinatorial synthesis of novel compounds.

7. Appendix

7.1 Sequences

7.1.1 Adenylyltransferase (*idm*J) DNA sequence

GAGCTCAAGGAGATATACCATGGAACATCATCATCATCACCACAACCTGCATCAACTGCTGGTCGA
TACGGCGGCTAAAGAACCGGATCGTCTGGCTGTCGCTGGCACCGCAGCACGTCTGACGTACGCAGA
ACTGGATTCTACCGCAAACGCACTGGCACATCGTCTGCGCGCCCTGGGTGTTGGTCCGGGTGACCG
CGTGGTTCTGTGGAGTGATAAATCCCCGGCAGTCGTGGCAGCTATGCAGGCAGTGCTGCGTCTGGG
TGCAGCATATGTTCCGGCTGATGGCGCGCTGCCGATTGCCCGCGTGGCAGCTATGGCAGATGACTG
CAGCGCAGCAGCACTGCTGGCTCCGGCAGACCGTCTGGCACCGGTTGCAGATGTCCTGGGTCCGCG
TTGTCCGGCTGCAGACCTGGCACAGCGTCCGGACCCGGCAGCAGAACCGCTGAATGCTCTGGTTGC
GCCGGATGACCTGGCCTATATTCTGTACACCAGCGGTTCTACGGGCGCACCGAAGGGTGTGTGCAT
CAGCCATCGTAACGCCCGCGCATTTGTGGATTGGGCTGTGGAAGAACTGGCACCGGGTCCGCAAGA
TCGTTTCAGTAATCACGCCCCGTTTACCTTCGACCTGTCCGTTCTGGATCTGTATGCTGCGTTTAGCG
CTGGTGCGTCTGTCCATCTGATTCCGTCTGAACTGGCCTACGCACCGGAACAGCTGGTTGAATTTCT
GCACGATCGCCAAATCACCGTTTGGTATTCAGTCCCGTCGGCGCTGACGCTGATGATGCGTGACGG
TGGTCTGCTGGATCGTCCGGCACCGCGTGCACTGCGCACCGTCCTGTTTGCAGGTGAACCGTTCCCG
CTGCCGGGTGTGCGTGCTCTGGCAGGTTGGACGGATGCACGCCTGCTGAACCTGTATGGTCCGACC
GAAACGAATGTTTGTACCCGTCATGAAGTCCGTCCGACGGATCTGGACGGTGATCGTCCGCTGCCG
ATCGGTACCGCAGTGAGTGGTGACCGTGCATGGGCAGAAGGTCCGGATGGCCGCCTGGCAGCACC
GGGTGAAGAAGGCGAACTGCTGGTCGATGGTCCGACGGTGATGCTGGGCTATTGGGGTGGTGAACC
GCACACCGGTCCGTACCGTACGGGTGACCTGGTGCGTCCGCTGCCGGATGGTTCCTTTGCGTACCTG
GGCCGTCGCGACCATATGGTGAAAGTTCGTGGTCACCGCGTTGAACTGGGCGAAGTGGAATCAGTT
CTGGCTCTGCATCCGGATGTGGCCGAAGCTGCGGCCGTTGTCGTGGGTTCGGGCATGGACGGTCGC
CTGGTTGCCTTCGTTGTCCCGGAACCGGATCGTCGCCCGGGTGTGCTGAGTCTGGTTCGTCACGCAG
CTCAACGTCTGCCGCGCTATATGGTGGCAGATGAAGTCCGTGTGCTGCCGGGTCTGCCGCGTACCC
GCAACGGCAAAGTTGATCGTCTGGCACTGCGTGATACGGCTGAAGCACCGGCACCGGGCGCTGCGG
CACAGTAATAAGTCGAC

7.1.2 Adenylyltransferase (IdmJ) amino acid sequence

MEHHHHHHNLHQLLVDTAAKEPDRLAVAGTAARLTYAELDSTANALAHRLRALGVGPGDRVVLWSD
KSPAVVAAMQAVLRLGAAYVPADGALPIARVAAMADDCSAAALLAPADRLAPVADVLGPRCPAADL
AQRPDPAAEPLNALVAPDDLAYILYTSGSTGAPKGVCISHRNARAFVDWAVEELAPGPQDRFSNHAPFT
FDLSVLDLYAAFSAGASVHLIPSELAYAPEQLVEFLHDRQITVWYSVPSALTLMMRDGGLLDRPAPRAL
RTVLFAGEPFPLPGVRALAGWTDARLLNLYGPTETNVCTRHEVRPTDLDGDRPLPIGTAVSGDRAWAE
GPDGRLAAPGEEGELLVDGPTVMLGYWGGEPHTGPYRTGDLVRPLPDGSFAYLGRRDHMVKVRGHRV
ELGEVESVLALHPDVAEAAAVVVGSGMDGRLVAFVVPEPDRRPGVLSLVRHAAQRLPRYMVADEVRV
LPGLPRTRNGKVDRLALRDTAEAPAPGAAAQ

7.1.3 Prolyl dehydrogenase (*idm*I) DNA sequence

GGATCCAAGGAGATATACCATGGAACACCACCACCACCACCACGACTTTGACCTGACCGAAAAAC
AACGCCAACGCTACGACGATGTGCTGTCCGCCGTCCGCGAACGTCTGGGTGAACCGCCGAGCGGTG
AACCGTTTACCCCGGCACGTTGGCAGGATGCAGCACGTATTGGTCTGACCGGCCTGTGCCTGCCGA

CGGAATTTGGCGGTGGCGGTCTGGGTGCACTGGATACCGCTCTGTGCCTGGAAGCCTTTGGTCGTG
GTTGTCCGGACACGGGTCTGGTCTTCGCAGTGAGCGCACATCTGCTGGCATGTGCTGTGCCGATCCG
CGATTTTGCGGACGTCTCTGTGCGTGGTGAACTGCTGTCAGGTCTGGCCTCGGGCGAACTGGTTGCA
GCTAACGCAATGACCGAAGATGACGCTGGCAGCGATCTGTCTCGCCTGGCAGTCACGGCTGAACGT
AAAGATGACGGTTATCTGCTGAACGGCGAAAAAAGTTTCGCGTCCAATGCCCCGGCGGCCGATGTT
CTGGTCACCTACGGTACGAGCGACCCGGCAGCTGGTTTTCTGGGTGTTACCGCATTCGTCCTGCGTG
CAGATCTGCCGGGTGTGCAGGTTGGCGAACCGTTTCGTAAAATGGGTCTGAGTGGTTGCCCGGCAG
GTCGTGTCGCATTCCGTGATTGTCATGTGCCGACCTCCCACCGTCTGGGTACGGAGGGTCAGGGTTC
AATTATCTTTCAACACTCGATGGGCTGGGAACGCGCAATTCTGTTCGCTGGTTATCTGGGCCTGATG
GAACGTCAGCTGGAACAATGCGTGCGTCATGCGCGCGAACGTCGCCAGTTTGGCCACGGTATCGGC
GAATTTCAAGCCGTTAGCCATCGCGTGGTTGGTATGAAACAGCGTCTGGAAGCAGCACGTCTGCTG
CTGTATCGTGCGTGTTGGCTGATGGATCAAGGTCGTGACCACTCAACCGCGGTGGCCCTGAGTAAA
ACGGCCGTGTCCGAAGGCGCAGTTGCTAATTCGCTGGATGCAATTCAGATCTTTGGCGGTAGCGGT
TACCTGTCTCCGGCAGGCATTGAACAGCAACTGCGTGACGCGGTTCCGTCTACCATCTTCAGCGGCA
CGACGGACATCCAACGCGAAATTGTGGCACGCGAAATCGGTCTGTAATAAGAATTC

7.1.4 Prolyl dehydrogenase (IdmI) amino acid sequence

MEHHHHHHDFDLTEKQRQRYDDVLSAVRERLGEPPSGEPFTPARWQDAARIGLTGLCLPTEFGGGGLG
ALDTALCLEAFGRGCPDTGLVFAVSAHLLACAVPIRDFADVSVRGELLSGLASGELVAANAMTEDDAG
SDLSRLAVTAERKDDGYLLNGEKSFASNAPAADVLVTYGTSDPAAGFLGVTAFVLRADLPGVQVGEPF
RKMGLSGCPAGRVAFRDCHVPTSHRLGTEGQGSIIFQHSMGWERAILFAGYLGLMERQLEQCVRHARE
RRQFGHGIGEFQAVSHRVVGMKQRLEAARLLLYRACWLMDQGRDHSTAVALSKTAVSEGAVANSLD
AIQIFGGSGYLSPAGIEQQLRDAVPSTIFSGTTDIQREIVAREIGL

7.1.5 Prolyl carrier protein (*idm*K) DNA sequence

AAGCTTAAGGAGATATACCATGGAACATCATCATCATCATCATAGCCAGCCGGTTAGCGAAGAAGA
AACGAGCAAACGCATCACCGAATTTATCCGCGAACGCTTTCTGGATGGCGACCCGCAGGGTGAACT
GGAACCGGGCAGTCCGCTGCTGGAATGGGGTGTCCTGAACAGCCTGCGTGTTGCGCGCCTGATTGC
CTTTATCCGTAGCGACCTGGGCGTGCCGGTTAGTGCGATGGAAATTAATCCGGACAACCTGCGTGA
TGTTCGCTCTATTACCGCTATGGTGACCCGTCTGGCCCGTGCTGGTGCCTAATAACTCGAG

7.1.6 Prolyl carrier protein (IdmK) amino acid sequence

MEHHHHHHSQPVSEEETSKRITEFIRERFLDGDPQGELEPGSPLLEWGVLNSLRVARLIAFIRSDLGVPVS
AMEINPDNLRDVRSITAMVTRLARAGA

7.2 Vector maps

7.2.1 pET23a

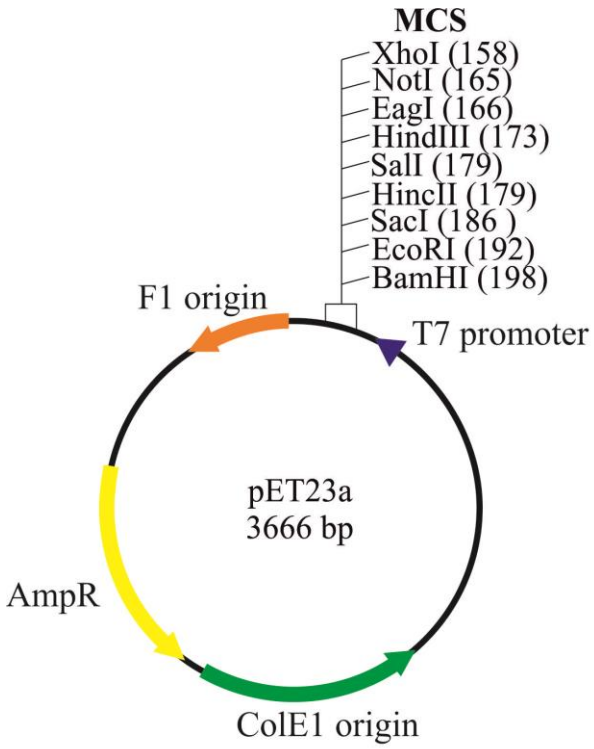


Figure 7.1- Vector map of the pET23a vector purchased from Merck4Biosciences (Nottingham, UK) highlighting key features such as the multiple cloning site (MCS), origin or replication, antibiotic resistance and promoter.

7.2.2 pK*nan*A plasmid map

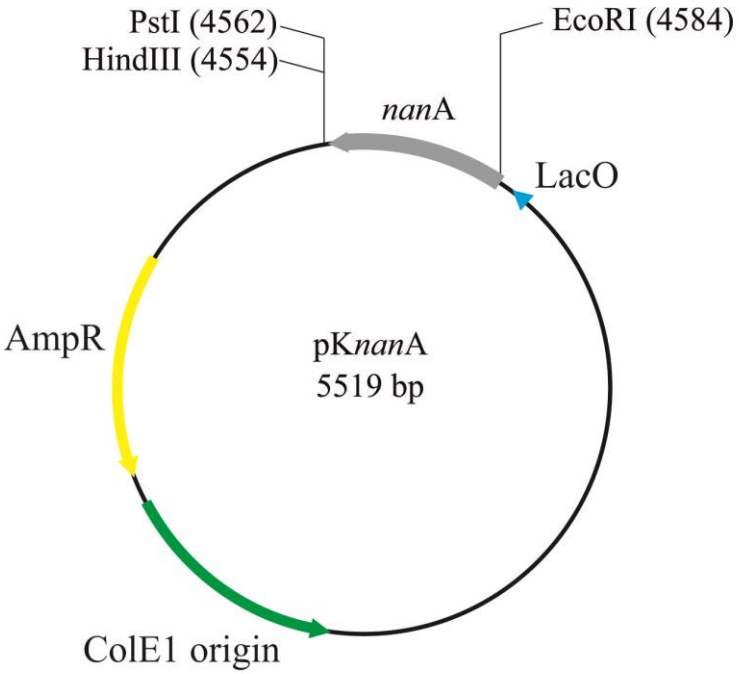


Figure 7.2-Plasmid map of the pK*nan*A plasmid (Timms *et al*., 2013) The *nan*A gene (933 bp) resides between the *Eco*RI and *Pst*I restriction sites. The origin of replication, antibiotic resistance and promoter are also highlighted.

7.2.3 pETDUET

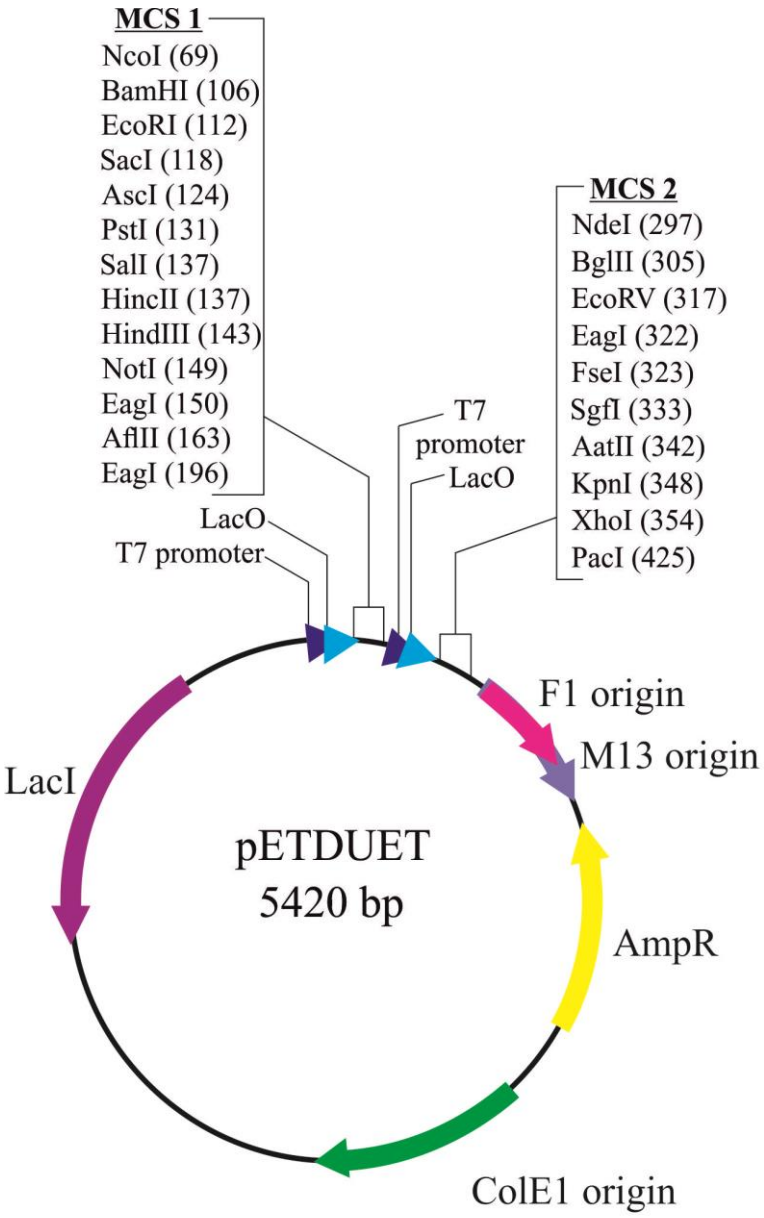


Figure 7.3- Vector map of the pETDUET vector purchased from Merck4Biosciences (Nottingham, UK) highlighting key features such as the multiple cloning sites 1 and 2 (MCS1 and 2), origin or replication, antibiotic resistance and promoter for each multiple cloning site.

7.2.4 pMAL c5X

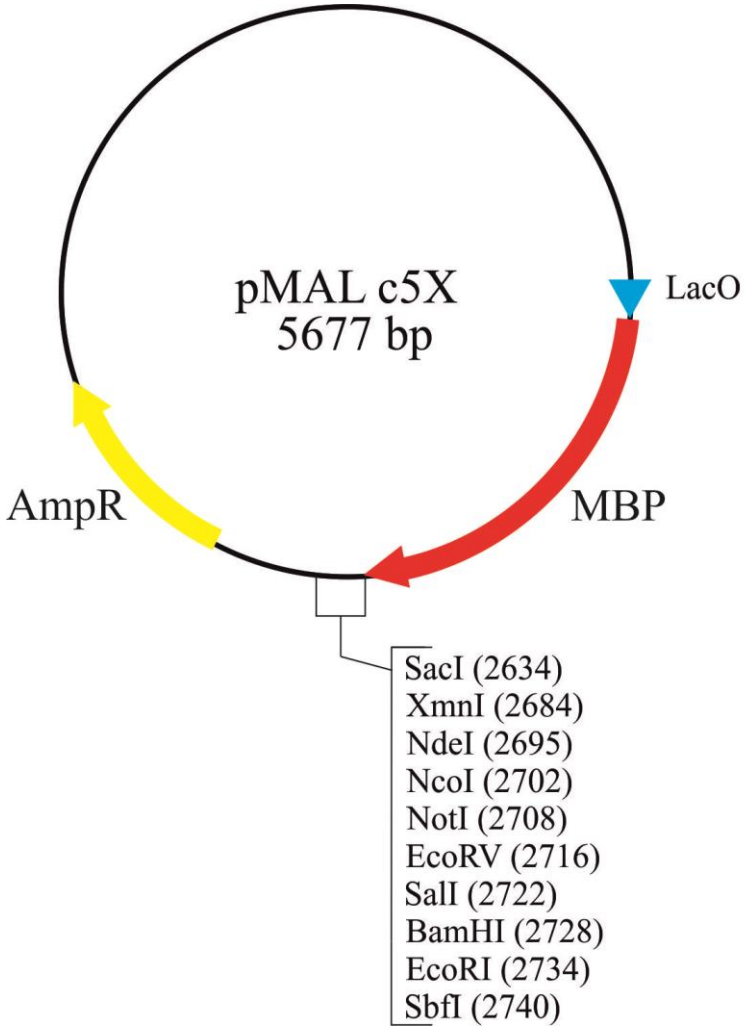


Figure 7.4- Vector map of pMAL c5X purchased from New England Biolabs (Ipswich, MA, USA) highlighting key features such as the multiple cloning site, promoter, maltose binding protein (MBP) and antibiotic resistance marker.

7.3 Oligonucleotide sequences

Oligonucleotide primers	Sequence
Sequencing primers	
pTRC-99A-FOR	5'-GAC ATC ATA ACG GTT CTG-3'
pTRC-99A-REV	5'-CTG AGT TCG GCA TGG GG-3'
pET-UPSTREAM	5'-ATG CGT CCG GCG TAG A-3'
T7 Term	5'-CTA GTT ATT GCT CAG CGG-3'
DUETDOWN1	5'-GAT TAT GCG GCC GTG TAC AA-3'
MAL-E	5'-GGT CGT CAG ACT GTC GAT GAA GCC-3'
M13-40FOR	5'-GTT TTC CCA GTC ACG AC-3'
PCR amplification primers	
Upstream-*idm*I-*Eco*RI	5'-GGG CCC GGA ATT CAA GGA G-3'
Downstream-*idm*I-*Pst*I	5'-AAC GAC GGC CAG TCT GCA GTT ATT ACA GAC CG-3'
*idm*I-For-TripleClone	5'-CGT ATT GTA CAC GGC C-3'
*idm*J-Rev-TripleClone	5'-CTG CAG CGC AAG CTT GTC GAC TTA TTA-3'
*idm*K-For-TripleClone	5'-AAG CTT GCG CTG CAG AAG GAG ATA TAC CAT GGA A-3'
*idm*K-Rev-TripleClone	5'-GGC CGT GTA CAA TAC GCT CGA GTT ATT AGG CAC C-3'
*idm*I-pMAL-Upstream-*Nde*I	5'-TCA GAT CAT ATG GAA CAC CAC CAC CAC-3'
*idm*I-pMAL-Downstream-*Eco*RI	5'-CTA GCT GAA TTC ATT ATT CAG ACC GAT TTC-3'
Site-directed mutagenesis primers	
*idm*J-C104A-For	5'-GCT ATG GCA GAT GAC GCC AGC GCA GCA GCA CTG-3'
*idm*J-C104A-Rev	5'-CAG TGC TGC TGC GCT GGC GTC ATC TGC CAT AGC-3'
*idm*J-C127A-For	5'-CTG GGT CCG CGT GCT CCG GCT GCA GAC-3'
*idm*J-C127A-Rev	5'-GTC TGC AGC CGG AGC ACG CGG ACC CAG-3'
*idm*J-C169A-For	5'-GCA CCG AAG GGT GTG GCC ATC AGC CAT CGT AAC-3'
*idm*J-C169A-Rev	5'-GTT ACG ATG GCT GAT GGC CAC ACC CTT CGG TGC-3'
*idm*J-C308A-For	5'-CGA CCG AAA CGA ATG TTG CTA CCC GTC ATG AAG TCC G-3'
*idm*J-C308A-Rev	5'-CGG ACT TCA TGA CGG GTA GCA ACA TTC GTT TCG GTC G-3'
*idm*K-S44A-For	5'-TGT CCT GAA CGC CCT GCG TGT TG-3'
*idm*K-S44A-Rev	5'-CAA CAC GCA GGG CGT TCA GGA CA-3'

Table 7.1- Table of sequences of oligonucleotide used.

7.4 PSI-BLAST search

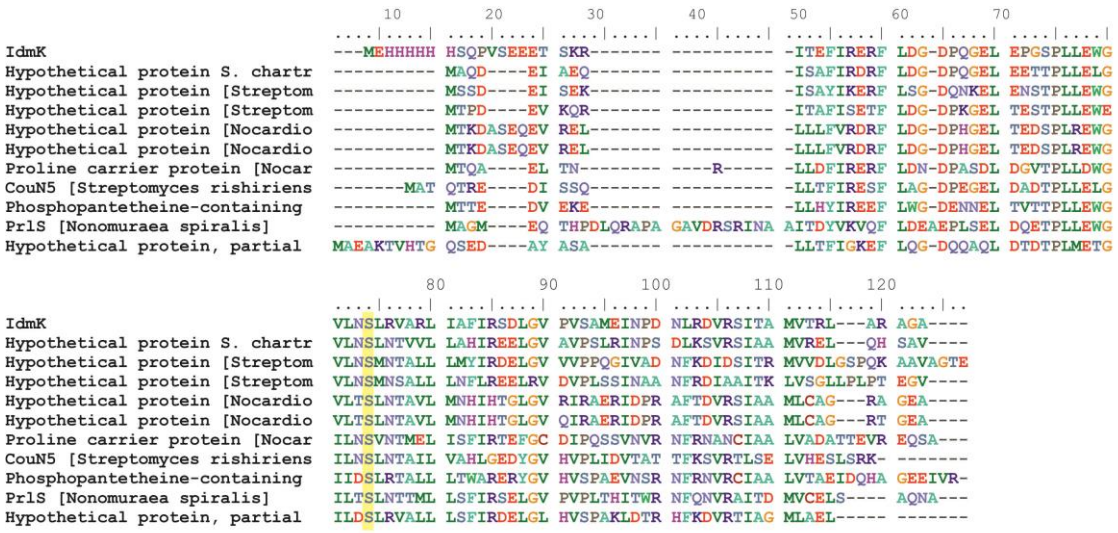


Figure 7.5-Top 10 sequences from the PSI-BLAST search with IdmK as the target sequence. The conserved serine is highlighted in the sequence.

7.5 SDS-PAGE analysis of uninduced *E. coli* BL21 (DE3) Gold

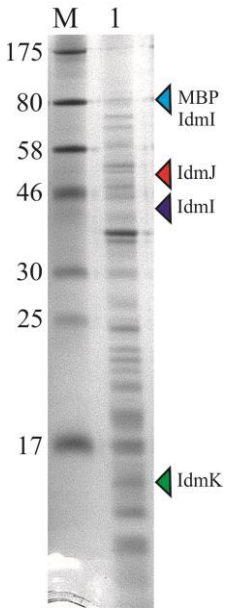


Figure 7.6- Reducing SDS-PAGE of uninduced *E. coli* BL21 (DE3) Gold cells. Expected masses for IdmJ, IdmI, IdmI-MBP and IdmK are highlighted for comparison.

Table 7.2-Table of chemical shift assignments (ppm) of IdmK. The protons have not been stereospecifically assigned.

seq	type	H	N	Hα	Hβ	CO	Cα	Cβ							
1	Ser	-	-	4.42	-	173.90	59.30	64.90							
2	Gln	8.30	122.80	-	-	173.68	54.70	30.00							
3	Pro	-	-	4.34	-	175.90	63.40	31.70	3.79-Hδ2	3.62-Hδ3	51.9-Cδ				
4	Val	7.84	122.80	3.96	1.80	175.00	63.10	35.00	0.86-Hγ1*	0.85-Hγ2*	22.9-Cγ1	23.0-Cγ2			
5	Ser	8.24	123.20	4.51	4.09,3.91	174.80	58.50	65.70							
6	Glu	-	-	3.87	2.04,2.04	178.10	61.30	30.60	2.30-Hγ1	2.30-Hγ2	37.9-Cγ				
7	Glu	8.69	119.50	4.03	1.96,2.08	179.10	60.90	30.30	2.31-Hγ2	2.31-Hγ3	37.5-Cγ				
8	Glu	7.96	120.70	4.09	2.06,2.06	178.60	60.00	30.60	2.27-Hγ2	2.27-Hγ3	37.1-Cγ				
9	Thr	8.18	117.90	3.69	4.26	175.60	68.90	68.80	1.22-Hγ2*	23.6-Cγ2					
10	Ser	8.46	115.60	4.09	3.92,4.02	178.50	62.60	63.80							
11	Lys	7.94	124.30	4.06	2.00,2.00	178.80	61.20	33.60	1.69-Hδ2	1.69-Hδ3	2.96-Hε2	2.96-Hε3	30.7-Cδ	43.1-Cε	
12	Arg	8.29	118.30	4.24	2.09,2.09	180.30	60.20	31.50	1.86-Hγ2	1.86-Hγ3	3.15-Hδ1	3.27-Hδ2	28.2-Cγ	44.4-Cδ	
13	Ile	8.56	119.90	3.77	2.04	177.10	65.90	38.50	1.84-Hγ12	1.26-Hγ13	0.97-Hγ2*	0.81-Hδ1*	30.2-Cγ1	19.2-Cγ2	14.1-Cδ1
14	Thr	8.44	117.50	3.57	4.35	175.60	69.50	69.40	1.22-Hγ2*	22.7-Cγ2					
15	Glu	8.30	119.70	3.92	1.95,2.08	177.80	60.50	30.60	2.30-Hγ1	2.44-Hγ2	37.5-Cγ				
16	Phe	7.61	120.00	3.82	2.59,2.39	176.50	62.60	40.20	6.38-Hδ*	6.86-Hε*	6.83-Hζ	131.7-Cδ*	130.1-Cε*	128.1-Cζ	
17	Ile	8.12	117.80	3.06	1.85	178.80	66.10	38.80	1.08-Hγ12	1.08-Hγ13	0.95-Hγ2*	0.86-Hδ1*	30.8-Cγ1	18.9-Cγ2	15.2-Cδ1
18	Arg	8.67	118.20	3.73	-	178.60	62.00	32.30							
19	Glu	8.20	116.00	3.72	1.92,1.80	178.30	60.20	31.30	2.07-Hγ1	2.29-Hγ2	37.3-Cγ				
20	Arg	8.04	113.50	4.00	0.75,0.30	178.40	57.30	30.10	0.45-Hγ1	0.86-Hγ2	2.55-Hδ1	2.66-Hδ2	27.1-Cγ	43.2-Cδ	

seq	type	H	N	Hα	Hβ	CO	Cα	Cβ								
21	Phe	7.69	114.40	5.28	-	176.00	55.40	40.80	6.90-Hζ	130.1-Cζ	7.09-Hδ*	7.22-Hε*	130.8-Cδ*	133.5-Cε*		
22	Leu	6.95	116.90	4.95	2.11,1.97	178.20	54.60	43.70	1.01-Hδ1*	0.99-Hδ2*	28.6-Cδ1	24.0-Cδ2				
23	Asp	8.80	121.90	4.37	2.99,2.62	175.10	56.70	40.80								
24	Gly	8.38	104.10	4.26,3.25	-	174.10	45.90	-								
25	Asp	8.15	117.40	4.52	2.24,2.99	174.21	54.90	42.70								
26	Pro	-	-	4.28	2.31,2.31	178.60	66.00	32.60	3.92-Hδ2	3.50-Hδ3	28.6-Cɣ	51.8-Cδ				
27	Gln	8.91	115.80	4.38	2.30,2.00	176.80	56.60	29.70								
28	Gly	8.28	112.30	3.78,3.67	-	175.60	48.70	-								
29	Glu	9.81	121.60	4.09	2.16,1.99	178.00	58.80	32.40	2.18-Hɣ1	38.4-Cɣ						
30	Leu	9.48	121.40	4.34	1.86,1.21	175.60	57.50	43.00	1.50-Hɣ	0.92-Hδ1*	1.02-Hδ2*	27.9-Cɣ	28.6-Cδ1	25.4-Cδ2		
31	Glu	8.98	128.20	-	-	175.72	54.90	31.50								
32	Pro	-	-	4.21	1.79,2.35	175.30	66.70	32.70	3.87-Hδ2	51.8-Cδ						
33	Gly	7.79	98.80	4.51,3.49	-	175.80	45.00	-								
34	Ser	7.80	121.30	-	-	172.55	59.10	62.70								
35	Pro	-	-	-	-	175.00	-	-								
36	Leu	7.95	125.80	3.63	1.85,1.13	179.50	58.90	43.70	0.78-Hδ1*	0.76-Hδ2*	27.4-Cδ1	23.7-Cδ2				
37	Leu	8.78	118.30	4.30	1.51,1.89	180.70	58.00	41.70	1.65-Hɣ	0.84-Hδ1*	0.69-Hδ2*	27.8-Cɣ	26.4-Cδ1	24.1-Cδ2		
38	Glu	7.34	121.30	4.14	2.22,2.22	182.50	60.20	30.50	2.36-Hɣ3	2.36-Hɣ2	37.4-Cɣ					
39	Trp	8.46	116.30	4.83	3.14,3.42	177.20	57.10	32.10	127.8-Nε1 121.2-Cζ3	7.13-Hδ1 124.9-Ch2	10.25-Hε1 7.33-Hε3	7.58-Hζ2	6.99-Hζ3	7.33-Hh2	123.9-Cδ1	114.7-Cζ2
40	Gly	7.62	105.80	4.17,3.85	-	174.70	46.80	-								
41	Val	7.70	117.60	3.69	2.04	176.20	65.30	33.60	1.10-Hɣ1*	0.93-Hɣ2*	23.6-Cɣ1	23.0-Cɣ2				
42	Leu	7.01	113.60	4.21	1.16,1.42	174.30	53.80	43.80	1.43-Hɣ	0.88-Hδ1*	0.71-Hδ2*	28.1-Cɣ	27.7-Cδ1	24.6-Cδ2		

seq	type	H	N	Hα	Hβ	CO	Cα	Cβ							
43	Asn	6.61	116.40	4.53	3.18,2.96	175.26	52.50	39.80							
44	-	-	-	-	-	-	-								
45	Leu	7.92	123.60	4.21	1.64,1.74	179.70	59.00	42.80	0.91-Hδ1*	0.90-Hδ2*	28.5-Cδ1	25.3-Cδ2			
46	Arg	8.20	119.40	3.86	-	178.90	60.90	32.20							
47	Val	8.79	120.00	-	-	177.20	68.80	32.40							
48	Ala	7.49	120.90	4.08	1.59	181.20	56.80	18.50							
49	Arg	7.74	118.10	4.30	2.29,2.07	179.10	59.90	31.40							
50	Leu	8.78	123.30	4.21	2.42,1.75	177.90	59.20	43.60	0.90-Hδ1*	1.00-Hδ2*	25.2-Cδ1	28.1-Cδ2			
51	Ile	8.94	119.30	3.66	2.17	178.00	64.20	36.50	1.41-Hγ12	1.41-Hγ13	0.86-Hγ2*	0.68-Hδ1*	28.5-Cγ1	18.3-Cγ2	11.1-Cδ1
52	Ala	7.92	122.20	4.18	1.60	179.80	56.50	18.90							
53	Phe	7.62	120.80	4.44	2.73,2.73	177.10	60.90	39.90	6.84-Hδ*	6.99-Hε*	132.4-Cδ*	130.0-Cε*	7.00-Hζ	129.0-Cζ	
54	Ile	8.26	122.00	3.04	1.88	177.50	66.60	39.60	0.73-Hγ12	0.73-Hγ13	0.92-Hγ2*	0.64-Hδ1*	30.4-Cγ1	19.8-Cγ2	14.7-Cδ1
55	Arg	8.15	117.00	4.14	1.84	179.10	59.80	32.20	1.67-Hγ2	1.67-Hγ3	3.05-Hδ2	3.05-Hδ3	29.2-Cγ	44.7-Cδ	
56	Ser	8.43	112.50	4.25	3.86,3.91	175.20	61.60	64.50							
57	Asp	8.48	122.80	4.47	1.99,2.14	177.50	57.40	42.40							
58	Leu	6.92	114.10	4.30	1.38,2.25	177.60	56.00	40.90	0.67-Hδ1*	1.02-Hδ2*	23.7-Cδ1	27.7-Cδ2			
59	Gly	6.63	105.80	3.71,3.77	-	174.40	47.50	-							
60	Val	7.49	121.50	4.47	1.65	173.03	60.00	35.40	0.73-Hγ1*	0.73-Hγ2*	22.0-Cγ1	22.0-Cγ2			
61	Pro	-	-	4.43	2.20,1.85	176.80	64.00	33.10	3.78-Hδ2	3.62-Hδ3	28.5-Cγ	51.6-Cδ			
62	Val	8.27	120.60	4.11	1.98	175.90	63.10	33.80	0.89-Hγ1*	0.89-Hγ2*	22.4-Cγ1	22.4-Cγ2			
63	Ser	8.48	119.80	4.56	4.24,3.99	175.78	58.30	66.00							
64		-	-	-	-	-	-	-							
65	Met	-	-	4.22	2.58,2.58	177.50	58.40	33.10	2.05-Hγ2	2.05-Hγ3	32.8-Cγ				
66	Glu	7.48	116.60	4.17	2.35,1.90	176.10	57.40	30.70	2.20-Hγ1	2.31-Hγ2	36.3-Cγ				

seq	type	H	N	Hα	Hβ	CO	Cα	Cβ							
67	Ile	7.29	120.00	3.76	2.05	173.30	62.20	36.70	1.71-Hγ12	1.71-Hγ13	0.78-Hγ2*	0.78-Hδ1*	27.9-Cγ1	19.4-Cγ2	14.4-Cδ1
68	Asn	7.46	123.10	5.01	2.95,2.69	173.24	52.10	39.90							
69	Pro	-	-	4.02	2.27,1.88	177.60	66.40	33.20							
70	Asp	7.91	116.10	4.40	2.59,2.59	178.20	58.10	41.30							
71	Asn	7.76	115.30	4.59	3.15,2.69	175.10	55.30	41.10							
72	Leu	7.05	114.80	4.30	1.43,1.67	174.70	54.80	43.20	0.71-Hδ1*	0.76-Hδ2*	23.8-Cδ1	27.7-Cδ2			
73	Arg	6.62	113.90	3.73	1.73,2.03	175.80	60.50	32.30	1.63-Hγ2	1.63-Hγ3	3.17-Hδ1	3.22-Hδ2	27.6-Cγ	44.7-Cδ	
74	Asp	7.71	110.40	4.48	3.12,3.35	175.40	53.30	43.30							
75	Val	9.95	121.30	3.50	2.20	179.20	68.80	33.40	0.93-Hγ1*	1.12-Hγ2*	22.1-Cγ1	24.7-Cγ2			
76	Arg	8.49	123.00	4.05	1.78,1.59	179.30	61.30	30.50	1.58-Hγ2	1.58-Hγ3	3.13-Hδ2	3.13-Hδ3	28.3-Cγ	44.4-Cδ	
77	Ser	9.38	120.30	4.34	-	177.60	63.50	-							
78	Ile	8.70	127.80	3.88	1.83	178.40	66.90	40.00	2.34-Hγ12	2.34-Hγ13	0.90-Hγ2*	0.66-Hδ1*	31.4-Cγ1	19.8-Cγ2	15.2-Cδ1
79	Thr	8.80	117.40	3.53	3.52	176.50	68.90	69.20							
80	Ala	8.18	124.90	4.15	1.53	179.60	56.20	18.80							
81	Met	7.82	119.50	3.97	2.15,2.15	176.90	60.30	33.10							
82	Val	8.40	119.00	3.30	2.14	177.30	68.70	32.70	0.88-Hγ1*	0.96-Hγ2*	23.2-Cγ1	25.7-Cγ2			
83	Thr	8.58	112.00	3.86	4.19	176.80	67.10	69.80							
84	Arg	7.81	122.00	4.10	2.05,1.88	179.20	60.40	31.00							
85	Leu	8.16	120.10	4.09	1.39,1.76	179.70	58.00	42.70	0.86-Hδ1*	0.88-Hδ2*	23.5-Cδ1	27.5-Cδ2			
86	Ala	8.45	122.20	4.14	1.47	179.10	55.20	19.70							
87	Arg	7.56	116.30	4.23	1.84,1.95	176.80	58.00	31.40	1.68-Hγ1	1.82-Hγ2	3.18-Hδ2	3.18-Hδ3	28.7-Cγ	44.6-Cδ	
88	Ala	7.72	122.70	4.28	1.46	178.20	54.20	19.90							
89	Gly	8.09	107.60	3.90,3.94	-	173.00	46.30	-							
90	Ala	7.68	129.00	4.12	1.34	182.51	54.80	21.20							

7.7 Crystallography statistics

Parameter	
Space group	$C222_1$
a, b, c (Å)	48.8, 237.5, 128.1
A, β, γ	90, 90, 90
High resolution limit (Å)	118.75-2.70
R_{merge}	0.062(0.498)
R_{pim} (I)	0.030(0.249)
Observed reflections	117 058
Unique reflections	20 716
Completeness (%)	98.6
Multiplicity	5.7
I/σ	11.5(2.2)

Table 7.3- Table of crystallography statistics of IdmK. Values in parentheses correspond to the lowest resolution range. R_{pim} indicates precision (multiplicity-weighted) R_{merge}.

References

References

Abe, I. & Morita, H. (2010). Structure and function of the chalcone synthase superfamily of plant type III polyketide synthases. *Nat.Prod. Rep. ,* **27,** 809-838.

Akey, D. L., Kittendorf, J. D., Giraldes, J. W., Fecik, R. A., Sherman, D. H. & Smith, J. L. (2006). Structural basis for macrolactonization by the pikromycin thioesterase. *Nat Chem Biol,* **2,** 537-42.

Alberts, A. W., Chen, J., Kuron, G., Hunt, V., Huff, J., Hoffman, C., Rothrock, J., Lopez, M., Joshua, H., Harris, E., Patchett, A., Monaghan, R., Currie, S., Stapley, E., Albersschonberg, G., Hensens, O., Hirshfield, J., Hoogsteen, K., Liesch, J. & Springer, J. (1980). Mevinolin - a highly potent competitive inhibitor of hydroxymethylglutaryl-coenzyme-A reductase and a cholesterol-lowering agent. *Proc. Natl. Acad. Sci. U.S.A. ,* **77,** 3957-3961.

Alekseyev, V. Y., Liu, C. W., Cane, D. E., Puglisi, J. D. & Khosla, C. (2007). Solution structure and proposed domain domain recognition interface of an acyl carrier protein domain from a modular polyketide synthase. *Protein sci,* **16,** 2093-107.

Alekshun, M. N. & Levy, S. B. (2007). Molecular mechanisms of antibacterial multidrug resistance. *Cell,* **128,** 1037-1050.

Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. (2006). The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics,* **22,** 195-201.

Asturias, F. J., Chadick, J. Z., Cheung, I. K., Stark, H., Witkowski, A., Joshi, A. K. & Smith, S. (2005). Structure and molecular organization of mammalian fatty acid synthase. *Nat Struct Mol Biol,* **12,** 225-32.

Baltz, R. H. (2008). Renaissance in antibacterial discovery from actinomycetes. *Curr Opin Pharmacol,* **8,** 557-63.

Barnwal, R. P., Van Voorhis, W. C. & Varani, G. (2011). Nmr structure of an acyl-carrier protein from B*orrelia burgdorferi*. *Acta Crystallogr F,* **67,** 1137-1140.

Bax, A. (2003). Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci,* **12,** 1-16.

Bax, A., Clore, G. M. & Gronenborn, A. M. (1990). ^{1}H-^{1}H correlation via isotropic mixing of ^{13}C magnetization, a new 3-Dimensional approach for assigning ^{1}H and ^{13}C spectra of ^{13}C-enriched proteins. *Journal of Magnetic Resonance,* **88,** 425-431.

Bax, A. & Grzesiek, S. (1993). Methodological advances in protein NMR. *Accounts Chem Res,* **26,** 131-138.

Bax, A. & Ikura, M. (1991). An efficient 3D NMR technique for correlating the proton and ^{15}N backbone amide resonances with the alpha-carbon of the preceding residue in uniformly ^{15}N/^{13}C enriched proteins. *Journal of Biomolecular NMR,* **1,** 99-104.

Beld, I., Sonnenschein, E. C., Vickery, C. R., Noel, J. P. & Burkart, M. D. (2014). The phosphopantetheinyl transferases: Catalysis of a post-translational modification crucial for life. *Natural Product Reports,* **31,** 61-108.

Belshaw, P. J., Walsh, C. T. & Stachelhaus, T. (1999). Aminoacyl-coas as probes of condensation domain selectivity in nonribosomal peptide synthesis. *Science,* **284,** 486-9.

References

Bender, C. L., Alarcon-Chaidez, F. & Gross, D. C. (1999). *Pseudomonas syringae* phytotoxins: Mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. *Microbiol Mol Biol Rev,* **63,** 266-292.

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C. H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'neil, S., Rabbinowitsch, E., Rajandream, M. A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B. G., Parkhill, J. & Hopwood, D. A. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature,* **417,** 141-147.

Berjanskii, M. V. & Wishart, D. S. (2007). The RCI server: Rapid and accurate calculation of protein flexibility using chemical shifts. *Nucleic Acids Res,* **35,** W531-7.

Bernardes, G. J. L., Chalker, J. M., Errey, J. C. & Davis, B. G. (2008). Facile conversion of cysteine and alkyl cysteines to dehydroalanine on protein surfaces: Versatile and switchable access to functionalized proteins. *Journal of the American Chemical Society,* **130,** 5052-5053.

Bessette, P. H., Aslund, F., Beckwith, J. & Georgiou, G. (1999). Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. *Proc. Natl. Acad. Sci. U.S.A.,* **96,** 13703-8.

Bevitt, D. J., Cortes, J., Haydock, S. F. & Leadlay, P. F. (1992). 6-deoxyerythronolide-B synthase-2 from *Saccharopolyspora erythraea* - cloning of the structural gene, sequence-analysis and inferred domain-structure of the multifunctional enzyme. *Eur J Biochem,* **204,** 39-49.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L. & Schwede, T. (2014). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res,* **42,** W252-W258.

Birch, A. J., Massywestropp, R. A. & Moye, C. J. (1955). Studies in relation to biosynthesis .7. 2-hydroxy-6-methylbenzoic acid in *Penicillium griseofulvum* dierckx. *Aust J Chem,* **8,** 539-544.

Bloch, F., Hansen, W. W. & Packard, M. (1946). The nuclear induction experiment. *Physical Review,* **70,** 474-485.

Borchert, S., Patil, S. S. & Marahiel, M. A. (1992). Identification of putative multifunctional peptide synthetase genes using highly conserved oligonucleotide sequences derived from known synthetases. *FEMS Microbiol Lett,* **92,** 175-180.

Bouvignies, G., Vallurupalli, P., Hansen, D. F., Correia, B. E., Lange, O., Bah, A., Vernon, R. M., Dahlquist, F. W., Baker, D. & Kay, L. E. (2011). Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature,* **477,** 111-U134.

Broadhurst, R. W., Nietlispach, D., Wheatcroft, M. P., Leadlay, P. F. & Weissman, K. J. (2003). The structure of docking domains in modular polyketide synthases. *Chem Biol,* **10,** 723-731.

Bruner, S. D., Weber, T., Kohli, R. M., Schwarzer, D., Marahiel, M. A., Walsh, C. T. & Stubbs, M. T. (2002). Structural basis for the cyclization of the

lipopeptide antibiotic surfactin by the thioesterase domain SrfTE. *Structure,* **10,** 301-310.

Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. & Jones, D. T. (2013). Scalable web services for the PSIPRED protein analysis workbench. *Nucleic Acids Res,* **41,** W349-W357.

Buchholz, T. J., Geders, T. W., Bartley, F. L., Reynolds, K. A., Smith, J. L. & Sherman, D. H. (2009). Structural basis for binding specificity between subclasses of modular polyketide synthase docking domains. *ACS Chem Biol,* **4,** 41-52.

Busche, A., Gottstein, D., Hein, C., Ripin, N., Pader, I., Tufar, P., Eisman, E. B., Gu, L. C., Walsh, C. T., Sherman, D. H., Lohr, F., Guntert, P. & Dotsch, V. (2012). Characterization of molecular interactions between ACP and halogenase domains in the curacin a polyketide synthase. *ACS Chem Biol,* **7,** 377-385.

Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W. H., Yang, X. C., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. & Emili, A. (2005). Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature,* **433,** 531-537.

Butler, M. S., Blaskovich, M. A. & Cooper, M. A. (2013). Antibiotics in the clinical pipeline in 2013. *J Antibiot (Tokyo),* **66,** 571-91.

Caffrey, P., Bevitt, D. J., Staunton, J. & Leadlay, P. F. (1992). Identification of DEBS1, DEBS2 and DEBS3, the multienzyme polypeptides of the erythromycin-producing polyketide synthase from *Saccharopolyspora erythraea*. *FEBS lett,* **304,** 225-8.

Campbell, C. D. & Vederas, J. C. (2010). Biosynthesis of lovastatin and related metabolites formed by fungal iterative PKS enzymes. *Biopolymers,* **93,** 755-763.

Cane, D. E., Celmer, W. D. & Westley, J. W. (1983). Unified stereochemical model of polyether antibiotic-structure and biogenesis. *J Am Chem Soc,* **105,** 3594-3600.

Cane, D. E., Walsh, C. T. & Khosla, C. (1998). Harnessing the biosynthetic code: Combinations, permutations, and mutations. *Science,* **282,** 63-8.

Cavalli, A., Salvatella, X., Dobson, C. M. & Vendruscolo, M. (2007). Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. U.S.A.,* **104,** 9615-9620.

Cavanagh, J., Fairbrother, W. J., Palmer, A. G., Rance, M. & Skelton, N. J. (2007). Protein NMR spectroscopy principles and practice second edition preface. *Protein NMR Spectroscopy: Principles and Practice, 2nd Edition,* V-Vi.

Chain, E., Florey, H. W., Gardner, A. D., Heatley, N. G., Jennings, M. A., Ewing, J. O. & Sanders, A. G. (1940). Penicillin as a chemotherapeutic agent. *Lancet,* **2,** 226-228.

Chalker, J. M., Gunnoo, S. B., Boutureira, O., Gerstberger, S. C., Fernandez-Gonzalez, M., Bernardes, G. J. L., Griffin, L., Hailu, H., Schofield, C. J. & Davis, B. G. (2011). Methods for converting cysteine to dehydroalanine on peptides and proteins. *Chemical Science,* **2,** 1666-1676.

Challis, G. L. (2006). Engineering *Escherichia coli* to produce nonribosomal peptide antibiotics. *Nat Chem Biol,* **2,** 398-400.

Chan, D. I., Chu, B. C. H., Lau, C. K. Y., Hunter, H. N., Byers, D. M. & Vogel, H. J. (2010). NMR solution structure and biophysical characterization of *Vibrio*

harveyi acyl carrier protein A75H effects of divalent metal ions. *J Biol Chem,* **285,** 30558-30566.

Chan, D. I. & Vogel, H. J. (2010). Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem J,* **430,** 1-19.

Chan, Y. A., Podevels, A. M., Kevany, B. M. & Thomas, M. G. (2009). Biosynthesis of polyketide synthase extender units. *Nat prod rep,* **26,** 90-114.

Chen, A. Y., Schnarr, N. A., Kim, C. Y., Cane, D. E. & Khosla, C. (2006). Extender unit and acyl carrier protein specificity of ketosynthase domains of the 6-deoxyerythronolide B synthase. *J Am Chem Soc,* **128,** 3067-3074.

Chen, Y. F., Kelly, E. E., Masluk, R. P., Nelson, C. L., Cantu, D. C. & Reilly, P. J. (2011). Structural classification and properties of ketoacyl synthases. *Protein Sci,* **20,** 1659-1667.

Cheng, Y. Q., Tang, G. L. & Shen, B. (2003). Type I polyketide synthase requiring a discrete acyltransferase for polyketide biosynthesis. *Proc Natl Acad Sci U S A,* **100,** 3149-54.

Clardy, J., Fischbach, M. A. & Walsh, C. T. (2006). New antibiotics from bacterial natural products. *Nat Biotechnol,* **24,** 1541-50.

Clarridge, J. E. (2004). Impact of 16s rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.,* **17,** 840-62.

Clubb, R. T., Thanabal, V. & Wagner, G. (1992). A new 3D HN(CA)HA experiment for obtaining fingerprint H-(N)-H-alpha cross peaks in ^{15}N-labeled and ^{13}C-labeled proteins. *Journal of Biomolecular NMR,* **2,** 203-210.

Clugston, S. L., Sieber, S. A., Marahiel, M. A. & Walsh, C. T. (2003). Chirality of peptide bond-forming condensation domains in nonribosomal peptide synthetases: The C_5 domain of tyrocidine synthetase is a DCL catalyst. *Biochemistry,* **42,** 12095-104.

Conti, E., Stachelhaus, T., Marahiel, M. A. & Brick, P. (1997). Structural basis for the activation of phenylalanine in the nonribosomal biosynthesis of gramicidin S. *The EMBO journal,* **16,** 4174-83.

Cooper, M. A. & Shlaes, D. (2011). Fix the antibiotics pipeline. *Nature,* **472,** 32.

Cornilescu, G., Delaglio, F. & Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR,* **13,** 289-302.

Cornilescu, G., Marquardt, J. L., Ottiger, M. & Bax, A. (1998). Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *Journal of the American Chemical Society,* **120,** 6836-6837.

Cortes, J., Haydock, S. F., Roberts, G. A., Bevitt, D. J. & Leadlay, P. F. (1990). An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of *Saccharopolyspora erythraea*. *Nature,* **348,** 176-8.

Crosby, J. & Crump, M. P. (2012). The structural role of the carrier protein--active controller or passive carrier. *Nat prod rep,* **29,** 1111-37.

Crump, M. P., Crosby, J., Dempsey, C. E., Parkinson, J. A., Murray, M., Hopwood, D. A. & Simpson, T. J. (1997). Solution structure of the actinorhodin polyketide synthase acyl carrier protein from *Streptomyces coelicolor* A3(2). *Biochemistry,* **36,** 6000-8.

Das, A. & Khosla, C. (2009). Biosynthesis of aromatic polyketides in bacteria. *Accounts Chem Res,* **42,** 631-639.

References

Davies, J. (2006). Where have all the antibiotics gone? *Can J Infect Dis Med Microbiol,* **17,** 287-90.

Davies, J. & Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev,* **74,** 417-33.

Day, J. B. & Mantle, P. G. (1982). Biosynthesis of radiolabeled verruculogen by *Penicillium simplicissimum*. *Appl Environ Microbiol,* **43,** 514-516.

De Alba, E., Tjandra, N. (2004). Residual dipolar couplings in protein structure determination. In: Downing, K. **2**. *Protein NMR techniques*. New Jersey: Humana Press, 89-106.

Del Vecchio, F., Petkovic, H., Kendrew, S. G., Low, L., Wilkinson, B., Lill, R., Cortes, J., Rudd, B. A., Staunton, J. & Leadlay, P. F. (2003). Active-site residue, domain and module swaps in modular polyketide synthases. *J Ind Microbiol Biotechnol,* **30,** 489-94.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRpipe - a multidimensional spectral processing system based on unix pipes. *J Biomol NMR,* **6,** 277-293.

Dhillon, N., Hale, R. S., Cortes, J. & Leadlay, P. F. (1989). Molecular characterization of a gene from *Saccharopolyspora erythraea* (*Streptomyces-erythraeus*) which is involved in erythromycin biosynthesis. *Mol Microbiol,* **3,** 1405-1414.

Dieckmann, R., Lee, Y. O., Van Liempt, H., Von Dohren, H. & Kleinkauf, H. (1995). Expression of an active adenylate-forming domain of peptide synthetases corresponding to acyl-coa-synthetases. *FEBS lett,* **357,** 212-6.

Doekel, S. & Marahiel, M. A. (2000). Dipeptide formation on engineered hybrid peptide synthetases. *Chem biol,* **7,** 373-84.

Domagk, G. (1935). Chemotherapy of bacterial infections. *Angewandte Chemie,* **48,** 0657-0667.

Donadio, S., Staver, M. J., Mcalpine, J. B., Swanson, S. J. & Katz, L. (1991). Modular organization of genes required for complex polyketide biosynthesis. *Science,* **252,** 675-679.

Dorrestein, P. C., Yeh, E., Garneau-Tsodikova, S., Kelleher, N. L. & Walsh, C. T. (2005). Dichlorination of a pyrrolyl-*S*-carrier protein by FADH(2)-dependent halogenase PltA during pyoluteorin biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America,* **102,** 13843-13848.

Du, L., Cheng, Y. Q., Ingenhorst, G., Tang, G. L., Huang, Y. & Shen, B. (2003). Hybrid peptide-polyketide natural products: Biosynthesis and prospects towards engineering novel molecules. *Genet Eng (N Y),* **25,** 227-67.

Du, L., Sanchez, C. & Shen, B. (2001). Hybrid peptide-polyketide natural products: Biosynthesis and prospects toward engineering novel molecules. *Metab Eng,* **3,** 78-95.

Du, L. C. & Lou, L. L. (2010). PKS and NRPS release mechanisms. *Nat Prod Rep,* **27,** 255-278.

Dunn, B. J. & Khosla, C. (2013). Engineering the acyltransferase substrate specificity of assembly line polyketide synthases. *J R Soc Interface,* **10**.

Dunn, B. J., Watts, K. R., Robbins, T., Cane, D. E. & Khosla, C. (2014). Comparative analysis of the substrate specificity of trans- versus cis-acyltransferases of assembly line polyketide synthases. *Biochemistry,* **53,** 3796-806.

References

Dutta, S., Whicher, J. R., Hansen, D. A., Hale, W. A., Chemler, J. A., Congdon, G. R., Narayan, A. R. H., Hakansson, K., Sherman, D. H., Smith, J. L. & Skiniotis, G. (2014). Structure of a modular polyketide synthase. *Nature,* **510,** 512-+.

Dutton, C. J., Banks, B. J. & Cooper, C. B. (1995). Polyether ionophores. *Nat Prod Rep,* **12,** 165-81.

Dutton, C. J., Gibson, S. P., Goudie, A. C., Holdom, K. S., Pacey, M. S., Ruddock, J. C., Bu'lock, J. D. & Richards, M. K. (1991). Novel avermectins produced by mutational biosynthesis. *J Antibiot (Tokyo),* **44,** 357-65.

Dutton, C. J., Hooper, A. M., Leadlay, P. F. & Staunton, J. (1994). Avermectin biosysthesis - intact incorporation of a diketide chain-assembly intermediate into the polyketide macrocyclic ring. *Tetrahedron Letters,* **35,** 327-330.

Edwards, A. L., Matsui, T., Weiss, T. M. & Khosla, C. (2014). Architectures of whole-module and bimodular proteins from the 6-deoxyerythronolide B synthase. *J Mol Biol,* **426,** 2229-45.

Ehmann, D. E., Shaw-Reid, C. A., Losey, H. C. & Walsh, C. T. (2000a). The EntF and EntE adenylation domains of *Escherichia coli* enterobactin synthetase: Sequestration and selectivity in acyl-AMP transfers to thiolation domain cosubstrates. *Proc Natl Acad Sci U S A,* **97,** 2509-14.

Ehmann, D. E., Trauger, J. W., Stachelhaus, T. & Walsh, C. T. (2000b). Aminoacyl-SNACS as small-molecule substrates for the condensation domains of nonribosomal peptide synthetases. *Chemistry & Biology,* **7,** 765-772.

Eppelmann, K., Stachelhaus, T. & Marahiel, M. A. (2002). Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry,* **41,** 9718-9726.

Engelke, J. & Rüterjans, H. (1999). Recent developments in studying the dynamics of protein structures from ^{15}N and ^{13}C relaxation time measurements. In: Krishna, N. R. & Berliner, L. J. *Biological Magnetic Resonance Volume 17; Structure computation and dynamics in protein NMR.*New York City: Springer Publishing, 357-418.

Farmer, B. T., Venters, R. A., Spicer, L. D., Wittekind, M. G. & Muller, L. (1992). A refocused and optimized HNCA - increased sensitivity and resolution in large macromolecules. *Journal of Biomolecular NMR,* **2,** 195-202.

Faul, M. M. & Huff, B. E. (2000). Strategy and methodology development for the total synthesis of polyether ionophore antibiotics. *Chemical reviews,* **100,** 2407-74.

Findlow, S. C., Winsor, C., Simpson, T. J., Crosby, J. & Crump, M. P. (2003). Solution structure and dynamics of oxytetracycline polyketide synthase acyl carrier protein from *Streptomyces rimosus*. *Biochemistry,* **42,** 8423-33.

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

Fischbach, M. A. & Walsh, C. T. (2006). Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic, machinery, and mechanisms. *Chem Rev,* **106,** 3468-96.

Fisher, J. F., Meroueh, S. O. & Mobashery, S. (2005). Bacterial resistance to β-lactam antibiotics: Compelling opportunism, compelling opportunity. *Chemical Reviews,* **105,** 395-424.

Fleming, A. (1929). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of *B. influenzae*. *British Journal of Experimental Pathology,* **10,** 226-236.

References

Fleming, K., Matthews, S. (2004) Media for studies of partially aligned states. In: Downing, K. **2**. *Protein NMR techniques*. New Jersey: Humana Press, 79-88.

Fossi, M., Linge, J., Labudde, D., Leitner, D., Nilges, M. & Oschkinat, H. (2005a). Influence of chemical shift tolerances on NMR structure calculations using ARIA protocols for assigning noe data. *J Biomol NMR,* **31,** 21-34.

Fossi, M., Oschkinat, H., Nilges, M. & Ball, L. J. (2005b). Quantitative study of the effects of chemical shift tolerances and rates of SA cooling on structure calculation from automatically assigned noe data. *Journal of Magnetic Resonance,* **175,** 92-102.

Fossi, M., Oschkinat, H., Nilges, M. & Ball, L. J. (2005c). Quantitative study of the effects of chemical shift tolerances and rates of sa cooling on structure calculation from automatically assigned noe data. *J Magn Reson,* **175,** 92-102.

Frykman, S., Leaf, T., Carreras, C. & Licari, P. (2001). Precursor-directed production of erythromycin analogs by *Saccharopolyspora erythraea*. *Biotechnology and Bioengineering,* **76,** 303-310.

Fujii, I. (2009). Heterologous expression systems for polyketide synthases. *Nat Prod Rep,* **26,** 155-69.

Fujii, I., Mori, Y., Watanabe, A., Kubo, Y., Tsuji, G. & Ebizuka, Y. (1999). Heterologous expression and product identification of *Colletotrichum lagenarium* polyketide synthase encoded by the PKS1 gene involved in melanin biosynthesis. *Biosci Biotechnol Biochem,* **63,** 1445-52.

Gallagher, J. R. & Prigge, S. T. (2010). *Plasmodium falciparum* acyl carrier protein crystal structures in disulfide-linked and reduced states and their prevalence during blood stage growth. *Proteins-Structure Function and Bioinformatics,* **78,** 575-588.

Gallimore, A. R. (2009). The biosynthesis of polyketide-derived polycyclic ethers. *Nat Prod Rep,* **26,** 266-80.

Garneau, S., Dorrestein, P. C., Kelleher, N. L. & Walsh, C. T. (2005). Characterization of the formation of the pyrrole moiety during clorobiocin and coumermycin A1 biosynthesis. *Biochemistry,* **44,** 2770-80.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003). Expasy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res,* **31,** 3784-3788.

Gause, G. F. & Brazhnikova, M. G. (1944). Gramicidin S and its use in the treatment of infected wounds. *Nature,* **154,** 703-703.

Gehret, J. J., Gu, L. C., Gerwick, W. H., Wipf, P., Sherman, D. H. & Smith, J. L. (2011). Terminal alkene formation by the thioesterase of curacin a biosynthesis structure of a decarboxylating thioesterase. *Journal of Biological Chemistry,* **286,** 14445-14454.

Gehring, A. M., Bradley, K. A. & Walsh, C. T. (1997). Enterobactin biosynthesis in *Escherichia coli*: Isochorismate lyase (EntB) is a bifunctional enzyme that is phosphopantetheinylated by EntD and then acylated by EntE using ATP and 2,3-dihydroxybenzoate. *Biochemistry,* **36,** 8495-8503.

Geoghegan, K. F., Dixon, H. B., Rosner, P. J., Hoth, L. R., Lanzetti, A. J., Borzilleri, K. A., Marr, E. S., Pezzullo, L. H., Martin, L. B., Lemotte, P. K., Mccoll, A. S., Kamath, A. V. & Stroh, J. G. (1999). Spontaneous alpha-n-6-phosphogluconoylation of a "His tag" in *Escherichia coli*: The cause of extra mass of 258 or 178 Da in fusion proteins. *Anal Biochem,* **267,** 169-84.

References

Giessen, T. W. & Marahiel, M. A. (2012). Ribosome-independent biosynthesis of biologically active peptides: Application of synthetic biology to generate structural diversity. *Febs Letters,* **586,** 2065-2075.

Ginzinger, S. W., Gerick, F., Coles, M. & Heun, V. (2007). Checkshift: Automatic correction of inconsistent chemical shift referencing. *Journal of Biomolecular NMR,* **39,** 223-227.

Glinski, M., Urbanke, C., Hornbogen, T. & Zocher, R. (2002). Enniatin synthetase is a monomer with extended structure: Evidence for an intramolecular reaction mechanism. *Archives of Microbiology,* **178,** 267-273.

Gokhale, R. S., Tsuji, S. Y., Cane, D. E. & Khosla, C. (1999). Dissecting and exploiting intermodular communication in polyketide synthases. *Science,* **284,** 482-5.

Grunewald, J. & Marahiel, M. A. (2006). Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides. *Microbiology and Molecular Biology Reviews,* **70,** 121-146.

Grzesiek, S. & Bax, A. (1992). An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. *Journal of Magnetic Resonance,* **99,** 201-207.

Guex, N., Peitsch, M. C. & Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis,* **30,** S162-S173.

Gully, D. & Bouveret, E. (2006). A protein network for phospholipid synthesis uncovered by a variant of the tandem affinity purification method in *Escherichia coli*. *Proteomics,* **6,** 282-293.

Gully, D., Moinier, D., Loiseau, L. & Bouveret, E. (2003). New partners of acyl carrier protein detected in *Escherichia coli* by tandem affinity purification. *Febs Letters,* **548,** 90-96.

Guy, J. E., Whittle, E., Moche, M., Lengqvist, J., Lindqvist, Y. & Shanklin, J. (2011). Remote control of regioselectivity in acyl-acyl carrier protein-desaturases. *Proc Natl Acad Sci U S A,* **108,** 16594-16599.

Hahn, M. & Stachelhaus, T. (2004). Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains. *Proceedings of the National Academy of Sciences of the United States of America,* **101,** 15585-90.

Hans, M., Hornung, A., Dziarnowski, A., Cane, D. E. & Khosla, C. (2003). Mechanistic analysis of acyl transferase domain exchange in polyketide synthase modules. *Journal of the American Chemical Society,* **125,** 5366-5374.

Hansen, M. R., Mueller, L. & Pardi, A. (1998). Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nature Structural Biology,* **5,** 1065-1074.

Henkin, J. A., Jennings, M. E., Matthews, D. E. & Vigoreaux, J. O. (2004). Mass processing--an improved technique for protein identification with mass spectrometry data. *Journal of Biomolecular Techniques : JBT,* **15,** 230-7.

Hertweck, C., Luzhetskyy, A., Rebets, Y. & Bechthold, A. (2007). Type II polyketide synthases: Gaining a deeper insight into enzymatic teamwork. *Natural Product Reports,* **24,** 162-190.

Hirel, P. H., Schmitter, J. M., Dessen, P., Fayat, G. & Blanquet, S. (1989). Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino-acid. *Proceedings of the*

National Academy of Sciences of the United States of America, **86,** 8247-8251.

Holm, L. & Rosenstrom, P. (2010). Dali server: Conservation mapping in 3D. *Nucleic Acids Research,* **38,** W545-W549.

Hooper, D. C. (1999). Mechanisms of fluoroquinolone resistance. *Drug Resistance Updates,* **2,** 38-55.

Hopwood, D. A. (1997). Genetic contributions to understanding polyketide synthases. *Chemical Reviews,* **97,** 2465-2497.

Hubbard, B. K. & Walsh, C. T. (2003). Vancomycin assembly: Nature's way. *Angewandte Chemie-International Edition,* **42,** 730-765.

Hur, G. H., Vickery, C. R. & Burkart, M. D. (2012). Explorations of catalytic domains in nonribosomal peptide synthetase enzymology. *Natural Product Reports,* **29,** 1074-1098.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. & Omura, S. (2003). Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol,* **21,** 526-31.

Ikura, M., Kay, L. E. & Bax, A. (1990). A novel-approach for sequential assignment of ^{1}H, ^{13}C, and ^{15}N spectra of larger proteins - heteronuclear triple-resonance 3-Dimensional NMR-spectroscopy - application to calmodulin. *Biochemistry,* **29,** 4659-4667.

Ireland, R. E., Habich, D. & Norbeck, D. W. (1985). The convergent synthesis of polyether ionophore antibiotics - the synthesis of the monensin spiroketal. *Journal of the American Chemical Society,* **107,** 3271-3278.

Jacobsen, J. R., Keatinge-Clay, A. T., Cane, D. E. & Khosla, C. (1998). Precursor-directed biosynthesis of 12-ethyl erythromycin. *Bioorganic & Medicinal Chemistry,* **6,** 1171-1177.

Johnson, M. N. R., Londergan, C. H. & Charkoudian, L. K. (2014). Probing the phosphopantetheine arm conformations of acyl carrier proteins using vibrational spectroscopy. *Journal of the American Chemical Society,* **136,** 11240-11243.

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

Jung, Y. S. & Zweckstetter, M. (2004). MARS -- robust automatic backbone assignment of proteins. *Journal of biomolecular NMR,* **30,** 11-23.

Kalverda, A. P., Thompson, G. S., Vogel, A., Schroder, M., Bowie, A. G., Khan, A. R. & Homans, S. W. (2009). Poxvirus K7 protein adopts a Bcl-2 fold: Biochemical mapping of its interactions with human DEAD box RNA helicase DDX3. *J Mol Biol,* **385,** 843-53.

Kamps-Hughes, N., Quimby, A., Zhu, Z. Y. & Johnson, E. A. (2013). Massively parallel characterization of restriction endonucleases. *Nucleic Acids Research,* **41**.

Karamanos, T. K., Kalverda, A. P., Thompson, G. S. & Radford, S. E. (2014). Visualization of transient protein-protein interactions that promote or inhibit amyloid assembly. *Mol Cell,* **55,** 214-26.

Kao, C. M., Katz, L. & Khosla, C. (1994). Engineered biosynthesis of a complete macrolactone in a heterologous host. *Science,* **265,** 509-512.

Kao, C. M., Mcpherson, M., Mcdaniel, R. N., Fu, H., Cane, D. E. & Khosla, C. (1997). Gain of function mutagenesis of the erythromycin polyketide

synthase. 2. Engineered biosynthesis of eight-membered ring tetraketide lactone. *Journal of the American Chemical Society,* **119,** 11339-11340.

Kao, C. M., Mcpherson, M., Mcdaniel, R. N., Fu, H., Cane, D. E. & Khosla, C. (1998). Alcohol stereochemistry in polyketide backbones is controlled by the beta-ketoreductase domains of modular polyketide synthases. *Journal of the American Chemical Society,* **120,** 2478-2479.

Kato, Y., Bai, L. Q., Xue, Q., Revill, W. P., Yu, T. W. & Floss, H. G. (2002). Functional expression of genes involved in the biosynthesis of the novel polyketide chain extension unit, methoxymalonyl-acyl carrier protein, and engineered biosynthesis of 2-desmethyl-2-methoxy-6-deoxyerythronolide B. *Journal of the American Chemical Society,* **124,** 5268-5269.

Katsu, T., Kobayashi, H. & Fujita, Y. (1986). Mode of action of gramicidin S on *Escherichia-coli* membrane. *Biochimica Et Biophysica Acta,* **860,** 608-619.

Katsuyama, Y. & Ohnishi, Y. (2012). Type III polyketide synthases in microorganisms. *Natural Product Biosynthesis by Microorganisms and Plant, Pt A,* **515,** 359-377.

Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. (1990). 3-Dimensional triple-resonance NMR-spectroscopy of isotopically enriched proteins. *Journal of Magnetic Resonance,* **89,** 496-514.

Keating, T. A., Marshall, C. G., Walsh, C. T. & Keating, A. E. (2002). The structure of VibH represents nonribosomal peptide synthetase condensation, cyclization and epimerization domains. *Nature structural biology,* **9,** 522-6.

Keating, T. A. & Walsh, C. T. (1999). Initiation, elongation, and termination strategies in polyketide and polypeptide antibiotic biosynthesis. *Current Opinion in Chemical Biology,* **3,** 598-606.

Keatinge-Clay, A. (2008). Crystal structure of the erythromycin polyketide synthase dehydratase. *Journal of Molecular Biology,* **384,** 941-953.

Keatinge-Clay, A. T. (2007). A tylosin ketoreductase reveals how chirality is determined in polyketides. *Chemistry & biology,* **14,** 898-908.

Keatinge-Clay, A. T. & Stroud, R. M. (2006). The structure of a ketoreductase determines the organization of the β-carbon processing enzymes of modular polyketide synthases. *Structure,* **14,** 737-48.

Keeler, J. (2002). Understanding NMR spectroscopy. *Understanding NMR spectroscopy*.

Kessler, N., Schuhmann, H., Morneweg, S., Linne, U. & Marahiel, M. A. (2004). The linear pentadecapeptide gramicidin is assembled by four multimodular nonribosomal peptide synthetases that comprise 16 modules with 56 catalytic domains. *Journal of Biological Chemistry,* **279,** 7413-7419.

Kevin Ii, D. A., Meujo, D. a. F. & Hamann, M. T. (2009). Polyether ionophores: Broad-spectrum and promising biologically active molecules for the control of drug-resistant bacteria and parasites. *Expert Opinion on Drug Discovery,* **4,** 109-146.

Khosla, C., Gokhale, R. S., Jacobsen, J. R. & Cane, D. E. (1999). Tolerance and specificity of polyketide synthases. *Annual Review of Biochemistry,* **68,** 219-253.

Khosla, C., Herschlag, D., Cane, D. E. & Walsh, C. T. (2014). Assembly line polyketide synthases: Mechanistic insights and unsolved problems. *Biochemistry,* **53,** 2875-2883.

Khosla, C., Tang, Y., Chen, A. Y., Schnarr, N. A. & Cane, D. E. (2007). Structure and mechanism of the 6-deoxyerythronolide B synthase. *Annu Rev Biochem,* **76,** 195-221.

Khosla, C. & Zawada, R. J. X. (1996). Generation of polyketide libraries via combinatorial biosynthesis. *Trends in Biotechnology,* **14,** 335-341.

Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L. & Schwede, T. (2009). The SWISS-MODEL repository and associated resources. *Nucleic Acids Research,* **37,** D387-D392.

Koglin, A., Lohr, F., Bernhard, F., Rogov, V. V., Frueh, D. P., Strieter, E. R., Mofid, M. R., Guntert, P., Wagner, G., Walsh, C. T., Marahiel, M. A. & Dotsch, V. (2008). Structural basis for the selectivity of the external thioesterase of the surfactin synthetase. *Nature,* **454,** 907-11.

Koglin, A., Mofid, M. R., Lohr, F., Schafer, B., Rogov, V. V., Blum, M. M., Mittag, T., Marahiel, M. A., Bernhard, F. & Dotsch, V. (2006). Conformational switches modulate protein interactions in peptide antibiotic synthetases. *Science,* **312,** 273-6.

Kohli, R. M., Trauger, J. W., Schwarzer, D., Marahiel, M. A. & Walsh, C. T. (2001). Generality of peptide cyclization catalyzed by isolated thioesterase domains of nonribosomal peptide synthetases. *Biochemistry,* **40,** 7099-7108.

Kohli, R. M. & Walsh, C. T. (2003). Enzymology of acyl chain macrocyclization in natural product biosynthesis. *Chemical Communications,* 297-307.

Komatsu, M., Uchiyama, T., Omura, S., Cane, D. E. & Ikeda, H. (2010). Genome-minimized *Streptomyces* host for the heterologous expression of secondary metabolism. *Proc Natl Acad Sci U S A,* **107,** 2646-51.

König, A., Schwecke, T., Molnar, I., Böhm, G. A., Lowden, P. A., Staunton, J. & Leadlay, P. F. (1997). The pipecolate-incorporating enzyme for the biosynthesis of the immunosuppressant rapamycin--nucleotide sequence analysis, disruption and heterologous expression of RapP from *Streptomyces hygroscopicus*. *Eur J Biochem,* **247,** 526-34.

Kopp, F. & Marahiel, M. A. (2007). Macrocyclization strategies in polyketide and nonribosomal peptide biosynthesis. *Natural Product Reports,* **24,** 735-749.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S. K., Codani, J. J., Connerton, I. F., Cummings, N. J., Daniel, R. A., Denizot, F., Devine, K. M., Dusterhoft, A., Ehrlich, S. D., Emmerson, P. T., Entian, K. D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S. Y., Glaser, P., Goffeau, A., Golightly, E. J., Grandi, G., Guiseppi, G., Guy, B. J., Haga, K., Haiech, J., Harwood, C. R., Henaut, A., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M. F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerrblanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S. M., Levine, A., Liu, H., Masuda, S., Mauel, C., Medigue, C., Medina, N., Mellado, R. P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., Oreilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S. H., Parro, V., Pohl, T. M., Portetelle, D., Porwollik, S., Prescott, A. M., Presecan, E., Pujic, P., Purnelle, B., et al. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature,* **390,** 249-256.

References

Labeda, D. P. (1987). Transfer of the type strain of *Streptomyces erythraeus* (Waksman 1923) Waksman and Henrici 1948 to the genus *Saccharopolyspora* Lacey and Goodfellow 1975 as *Saccharopolyspora erythraea* sp-nov, and designation of a neotype strain for *Streptomyces erythraeus*. *International Journal of Systematic Bacteriology,* **37,** 19-22.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature,* **227,** 680-5.

Lambalot, R. H., Gehring, A. M., Flugel, R. S., Zuber, P., Lacelle, M., Marahiel, M. A., Reid, R., Khosla, C. & Walsh, C. T. (1996). A new enzyme superfamily - the phosphopantetheinyl transferases. *Chemistry & biology,* **3,** 923-36.

Lau, J., Fu, H., Cane, D. E. & Khosla, C. (1999). Dissecting the role of acyltransferase domains of modular polyketide synthases in the choice and stereochemical fate of extender units. *Biochemistry,* **38,** 1643-51.

Li, C., Roege, K. E. & Kelly, W. L. (2009). Analysis of the indanomycin biosynthetic gene cluster from *Streptomyces Antibioticus* NRRL 8167. *Chembiochem,* **10,** 1064-72.

Li, C., Wen, A., Shen, B., Lu, J., Huang, Y. & Chang, Y. (2011). Fastcloning: A highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. *BMC Biotechnol,* **11,** 92.

Li, Q., Khosla, C., Puglisi, J. D. & Liu, C. W. (2003). Solution structure and backbone dynamics of the holo form of the frenolicin acyl carrier protein. *Biochemistry,* **42,** 4648-57.

Lin, S., Van Lanen, S. G. & Shen, B. (2009). A free-standing condensation enzyme catalyzing ester bond formation in c-1027 biosynthesis. *Proc Natl Acad Sci U S A,* **106,** 4183-8.

Linne, U. & Marahiel, M. A. (2000). Control of directionality in nonribosomal peptide synthesis: Role of the condensation domain in preventing misinitiation and timing of epimerization. *Biochemistry,* **39,** 10439-47.

Linne, U. & Marahiel, M. A. (2004). Reactions catalyzed by mature and recombinant nonribosomal peptide synthetases. *Methods Enzymol,* **388,** 293-315.

Liou, G. F. & Khosla, C. (2003). Building-block selectivity of polyketide synthases. *Current Opinion in Chemical Biology,* **7,** 279-284.

Liou, G. F., Lau, J., Cane, D. E. & Khosla, C. (2003). Quantitative analysis of loading and extender acyltransferases of modular polyketide synthases. *Biochemistry,* **42,** 200-7.

Lipmann, F. (1980). Bacterial production of antibiotic polypeptides by thiol-linked synthesis on protein templates. *Adv Microb Physiol,* **21,** 227-66.

Lipsitz, R. S. & Tjandra, N. (2004). Residual dipolar couplings in NMR structure analysis. *Annual Review of Biophysics and Biomolecular Structure,* **33,** 387-413.

Liu, C. M., Hermann, T. E., Liu, M., Bull, D. N., Palleroni, N. J., Prosser, B. L., Westley, J. W. & Miller, P. A. (1979). X-14547a, a new ionophorous antibiotic produced by *Streptomyces antibioticus* NRRL 8167. Discovery, fermentation, biological properties and taxonomy of the producing culture. *The Journal of antibiotics,* **32,** 95-9.

Liu, T., Cane, D. E. & Deng, Z. (2009). The enzymology of polyether biosynthesis. *Methods Enzymol,* **459,** 187-214.

Long, P. F., Wilkinson, C. J., Bisang, C. P., Cortes, J., Dunster, N., Oliynyk, M., Mccormick, E., Mcarthur, H., Mendez, C., Salas, J. A., Staunton, J. &

Leadlay, P. F. (2002). Engineering specificity of starter unit selection by the erythromycin-producing polyketide synthase. *Mol Microbiol,* **43,** 1215-25.

Loquet, A., Sgourakis, N. G., Gupta, R., Giller, K., Riedel, D., Goosmann, C., Griesinger, C., Kolbe, M., Baker, D., Becker, S. & Lange, A. (2012). Atomic model of the type III secretion system needle. *Nature,* **486,** 276-79.

Luo, L., Burkart, M. D., Stachelhaus, T. & Walsh, C. T. (2001). Substrate recognition and selection by the initiation module PheATE of gramicidin S synthetase. *J Am Chem Soc,* **123,** 11208-18.

Maier, T., Leibundgut, M. & Ban, N. (2008). The crystal structure of a mammalian fatty acid synthase. *Science,* **321,** 1315-22.

Marahiel, M. A. & Essen, L. O. (2009). Nonribosomal peptide synthetases: Mechanistic and structural aspects of essential domains. *Complex Enzymes in Microbial Natural Product Biosynthesis, Part A: Overview Articles and Peptides,* **458,** 337-51.

Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. (1997). Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chemical Reviews,* **97,** 2651-2673.

Marion, D., Driscoll, P. C., Kay, L. E., Wingfield, P. T., Bax, A., Gronenborn, A. M. & Clore, G. M. (1989a). Overcoming the overlap problem in the assignment of 1H-NMR spectra of larger proteins by use of 3-dimensional heteronuclear ^{1}H-^{15}N hartmann-hahn multiple quantum coherence and nuclear overhauser multiple quantum coherence spectroscopy - application to interleukin-1-beta. *Biochemistry,* **28,** 6150-6156.

Marion, D., Kay, L. E., Sparks, S. W., Torchia, D. A. & Bax, A. (1989b). 3-dimensional heteronuclear NMR of ^{15}N -labeled proteins. *Journal of the American Chemical Society,* **111,** 1515-1517.

Markwick, P. R. L., Malliavin, T. & Nilges, M. (2008). Structural biology by NMR: Structure, dynamics, and interactions. *Plos Computational Biology,* **4**.

Marques, M. A., Citron, D. M. & Wang, C. C. (2007). Development of tyrocidine a analogues with improved antibacterial activity. *Bioorganic & Medicinal Chemistry,* **15,** 6667-6677.

Marsden, A. F., Wilkinson, B., Cortes, J., Dunster, N. J., Staunton, J. & Leadlay, P. F. (1998). Engineering broader specificity into an antibiotic-producing polyketide synthase. *Science,* **279,** 199-202.

Martin, J. F., Gutierrez, S., Fernandez, F. J., Velasco, J., Fierro, F., Marcos, A. T. & Kosalkova, K. (1994). Expression of genes and processing of enzymes for the biosynthesis of penicillins and cephalosporins. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology,* **65,** 227-243.

Masoudi, A., Raetz, C. R. H., Zhou, P. & Pemble, C. W. (2014). Chasing acyl carrier protein through a catalytic cycle of lipid a production. *Nature,* **505,** 422-+.

Matthews, B. W. (1968). Solvent content of protein crystals. *Journal of Molecular Biology,* **33,** 491-7.

May, J. J., Kessler, N., Marahiel, M. A. & Stubbs, M. T. (2002). Crystal structure of DhbE, an archetype for aryl acid activating domains of modular nonribosomal peptide synthetases. *Proceedings of the National Academy of Sciences of the United States of America,* **99,** 12120-12125.

Mccoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). Phaser crystallographic software. *Journal of Applied Crystallography,* **40,** 658-674.

Mcdaniel, R., Ebert-Khosla, S., Hopwood, D. A. & Khosla, C. (1993). Engineered biosynthesis of novel polyketides. *Science,* **262,** 1546-50.

Mcdaniel, R., Kao, C. M., Fu, H., Hevezi, P., Gustafsson, C., Betlach, M., Ashley, G., Cane, D. E. & Khosla, C. (1997). Gain-of-function mutagenesis of a modular polyketide synthase. *Journal of the American Chemical Society,* **119,** 4309-4310.

Mcdaniel, R., Thamchaipenet, A., Gustafsson, C., Fu, H., Betlach, M. & Ashley, G. (1999). Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. *Proc Natl Acad Sci U S A,* **96,** 1846-51.

Mcdaniel, R., Welch, M. & Hutchinson, C. R. (2005). Genetic approaches to polyketide antibiotics. 1. *Chemical Reviews,* **105,** 543-558.

Meier, J. L. & Burkart, M. D. (2009). The chemical biology of modular biosynthetic enzymes. *Chemical Society reviews,* **38,** 2012-45.

Menninger, J. R. & Otto, D. P. (1982). Erythromycin, carbomycin, and spiramycin inhibit protein-synthesis by stimulating the dissociation of peptidyl-transfer RNA from ribosomes. *Antimicrobial Agents and Chemotherapy,* **21,** 811-818.

Menzella, H. G., Reid, R., Carney, J. R., Chandran, S. S., Reisinger, S. J., Patel, K. G., Hopwood, D. A. & Santi, D. V. (2005). Combinatorial polyketide biosynthesis by *de novo* design and rearrangement of modular polyketide synthase genes. *Nature Biotechnology,* **23,** 1171-6.

Mesleh, M. F., Veglia, G., Desilva, T. M., Marassi, F. M. & Opella, S. J. (2002). Dipolar waves as NMR maps of protein structure. *J Am Chem Soc,* **124,** 4206-7.

Mittenhuber, G., Weckermann, R. & Marahiel, M. A. (1989). Gene-cluster containing the genes for tyrocidine synthetases-1 and synthetases-2 from *Bacillus brevis* - evidence for an operon. *Journal of Bacteriology,* **171,** 4881-4887.

Molnar, I., Aparicio, J. F., Haydock, S. F., Khaw, L. E., Schwecke, T., Konig, A., Staunton, J. & Leadlay, P. F. (1996). Organisation of the biosynthetic gene cluster for rapamycin in *Streptomyces hygroscopicus*: Analysis of genes flanking the polyketide synthase. *Gene,* **169,** 1-7.

Molnar, I., Schupp, T., Ono, M., Zirkle, R., Milnamow, M., Nowak-Thompson, B., Engel, N., Toupet, C., Stratmann, A., Cyr, D. D., Gorlach, J., Mayo, J. M., Hu, A., Goff, S., Schmid, J. & Ligon, J. M. (2000). The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones A and B from *Sorangium cellulosum* so CE90. *Chemistry & biology,* **7,** 97-109.

Moore, B. S. & Hertweck, C. (2002). Biosynthesis and attachment of novel bacterial polyketide synthase starter units. *Natural product reports,* **19,** 70-99.

Mootz, H. D. & Marahiel, M. A. (1997). The tyrocidine biosynthesis operon of *Bacillus brevis*: Complete nucleotide sequence and biochemical characterization of functional internal adenylation domains. *Journal of Bacteriology,* **179,** 6843-6850.

Moseley, H. N. B. & Montelione, G. T. (1999). Automated analysis of NMR assignments and structures for proteins. *Current Opinion in Structural Biology,* **9,** 635-642.

Mueller, S., Coleman, J. R. & Wimmer, E. (2009). Putting synthesis into biology: A viral view of genetic engineering through *de novo* gene and genome synthesis. *Chem Biol,* **16,** 337-47.

References

Murli, S., Macmillan, K. S., Hu, Z. H., Ashley, G. W., Dong, S. D., Kealey, J. T., Reeves, C. D. & Kennedy, J. (2005). Chemobiosynthesis of novel 6-deoxyerythronolide B analogues by mutation of the loading module of 6-deoxyerythronolide B synthase 1. *Appl Environ Microbiol,* **71,** 4503-4509.

Muto, A. & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America,* **84,** 166-169.

Newman, D. J. & Cragg, G. M. (2012). Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod,* **75,** 311-335.

Nguyen, C., Haushalter, R. W., Lee, D. J., Markwick, P. R. L., Bruegger, J., Caldara-Festin, G., Finzel, K., Jackson, D. R., Ishikawa, F., O'dowd, B., Mccammon, J. A., Opella, S. J., Tsai, S. C. & Burkart, M. D. (2014). Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. *Nature,* **505,** 427-431.

Nilges, M. (1995). Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J Mol Biol,* **245,** 645-60.

O'brien, D. P., Kirkpatrick, P. N., O'brien, S. W., Staroske, T., Richardson, T. I., Evans, D. A., Hopkinson, A., Spencer, J. B. & Williams, D. H. (2000). Expression and assay of an N-methyltransferase involved in the biosynthesis of a vancomycin group antibiotic. *Chemical Communications***,** 103-104.

Olejniczak, E. T., Xu, R. X. & Fesik, S. W. (1992). A 4D-HcCH-TOCSY experiment for assigning the side-chain ^{1}H-resonance and ^{13}C-resonance of proteins. *Journal of Biomolecular NMR,* **2,** 655-659.

Oliynyk, M., Brown, M. J., Cortes, J., Staunton, J. & Leadlay, P. F. (1996). A hybrid modular polyketide synthase obtained by domain swapping. *Chem Biol,* **3,** 833-9.

Oliynyk, M., Stark, C. B. W., Bhatt, A., Jones, M. A., Hughes-Thomas, Z. A., Wilkinson, C., Oliynyk, Z., Demydchuk, Y., Staunton, J. & Leadlay, P. F. (2003). Analysis of the biosynthetic gene cluster for the polyether antibiotic monensin in *Streptomyces cinnamonensis* and evidence for the role of MonB and MonC genes in oxidative cyclization. *Molecular Microbiology,* **49,** 1179-1190.

Ottiger, M., Delaglio, F. & Bax, A. (1998). Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J Magn Reson,* **131,** 373-8.

Park, S. R., Yoo, Y. J., Ban, Y. H. & Yoon, Y. J. (2010). Biosynthesis of rapamycin and its regulation: Past achievements and recent progress. *Journal of Antibiotics,* **63,** 434-441.

Peirú, S., Gramajo, H. C. & Menzella, H. G. (2010). Recombinant approaches to large polyketide molecules as potential drugs. *Drug Discovery Today: Technologies,* **7,** e105-e113.

Pelludat, C., Rakin, A., Jacobi, C. A., Schubert, S. & Heesemann, J. (1998). The yersiniabactin biosynthetic gene cluster of *Yersinia enterocolitica*: Organization and siderophore-dependent regulation. *Journal of Bacteriology,* **180,** 538-546.

Perlman, D. & Bodanszky, M. (1971). Biosynthesis of peptide antibiotics. *Annu Rev Biochem,* **40,** 449-64.

Pfeifer, B. A., Admiraal, S. J., Gramajo, H., Cane, D. E. & Khosla, C. (2001). Biosynthesis of complex polyketides in a metabolically engineered strain of *Escherichia coli*. *Science,* **291,** 1790-2.

References

Pfeifer, E., Pavela-Vrancic, M., Von Döhren, H. & Kleinkauf, H. (1995). Characterization of tyrocidine synthetase 1 (TyC1): Requirement of post-translational modification for peptide biosynthesis. *Biochemistry,* **34,** 7450-9.

Piel, J. (2010). Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat Prod Rep,* **27,** 996-1047.

Ploskon, E., Arthur, C. J., Evans, S. E., Williams, C., Crosby, J., Simpson, T. J. & Crump, M. P. (2008). A mammalian type I fatty acid synthase acyl carrier protein domain does not sequester acyl chains. *The Journal of biological chemistry,* **283,** 518-28.

Ploskon, E., Arthur, C. J., Kanari, A. L. P., Wattana-Amorn, P., Williams, C., Crosby, J., Simpson, T. J., Willis, C. L. & Crump, M. P. (2010). Recognition of intermediate functionality by acyl carrier protein over a complete cycle of fatty acid biosynthesis. *Chemistry & Biology,* **17,** 776-785.

Pressman, B. C. (1976). Biological applications of ionophores. *Annual Review of Biochemistry,* **45,** 501-530.

Pressman, B. C. & Fahim, M. (1982). Pharmacology and toxicology of the monovalent carboxylic ionophores. *Annual Review of Pharmacology and Toxicology,* **22,** 465-490.

Prompers, J. J., Groenewegen, A., Hilbers, C. W. & Pepermans, H. A. M. (1998). Two-dimensional NMR experiments for the assignment of aromatic side chains in ^{13}C-labeled proteins. *Journal of Magnetic Resonance,* **130,** 68-75.

Purcell, E. M., Torrey, H. C. & Pound, R. V. (1946). Resonance absorption by nuclear magnetic moments in a solid. *Physical Review,* **69,** 37-38.

Qiu, X. & Janson, C. A. (2004). Structure of *apo* acyl carrier protein and a proposal to engineer protein crystallization through metal ions. *Acta crystallographica. Section D, Biological crystallography,* **60,** 1545-54.

Quadri, L. E., Weinreb, P. H., Lei, M., Nakano, M. M., Zuber, P. & Walsh, C. T. (1998). Characterization of Sfp, a *Bacillus subtilis* phosphopantetheinyl transferase for peptidyl carrier protein domains in peptide synthetases. *Biochemistry,* **37,** 1585-95.

Raman, S., Lange, O. F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G. H., Ramelot, T. A., Eletsky, A., Szyperski, T., Kennedy, M. A., Prestegard, J., Montelione, G. T. & Baker, D. (2010). NMR structure determination for larger proteins using backbone-only data. *Science,* **327,** 1014-1018.

Ramelot, T. A., Rossi, P., Forouhar, F., Lee, H. W., Yang, Y. H., Ni, S. S., Unser, S., Lew, S., Seetharaman, J., Xiao, R., Acton, T. B., Everett, J. K., Prestegard, J. H., Hunt, J. F., Montelione, G. T. & Kennedy, M. A. (2012). Structure of a specialized acyl carrier protein essential for lipid A biosynthesis with very long-chain fatty acids in open and closed conformations. *Biochemistry,* **51,** 7239-7249.

Rangan, V. S. & Smith, S. (1997). Alteration of the substrate specificity of the malonyl-coa/adetyl-coa:Acyl carrier protein *S*-acyltransferase domain of the multifunctional fatty acid synthase by mutation of a single arginine residue. *Journal of Biological Chemistry,* **272,** 11975-11978.

Rangaswamy, V., Jiralerspong, S., Parry, R. & Bender, C. L. (1998). Biosynthesis of the *Pseudomonas* polyketide coronafacic acid requires monofunctional and multifunctional polyketide synthase proteins. *Proceedings of the National Academy of Sciences of the United States of America,* **95,** 15469-15474.

References

Raymond, K. N., Dertz, E. A. & Kim, S. S. (2003). Enterobactin: An archetype for microbial iron transport. *Proceedings of the National Academy of Sciences of the United States of America,* **100,** 3584-3588.

Reeves, C. D., Murli, S., Ashley, G. W., Piagentini, M., Hutchinson, C. R. & Mcdaniel, R. (2001). Alteration of the substrate specificity of a modular polyketide synthase acyltransferase domain through site-specific mutations. *Biochemistry,* **40,** 15464-70.

Rehm, T., Huber, R. & Holak, T. A. (2002). Application of NMR in structural proteomics: Screening for proteins amenable to structural analysis. *Structure,* **10,** 1613-1618.

Reimmann, C., Patel, H. M., Serino, L., Barone, M., Walsh, C. T. & Haas, D. (2001). Essential PchG-dependent reduction in pyochelin biosynthesis of *Pseudomonas aeruginosa*. *Journal of Bacteriology,* **183,** 813-820.

Relman, D. A. (1999). The search for unrecognized pathogens. *Science,* **284,** 1308-1310.

Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T. E. & Nilges, M. (2007). ARIA2: Automated noe assignment and data integration in NMR structure calculation. *Bioinformatics,* **23,** 381-382.

Rocha-Martin, J., Harrington, C., Dobson, A. D. & O'gara, F. (2014). Emerging strategies and integrated systems microbiology technologies for biodiscovery of marine bioactive compounds. *Mar Drugs,* **12,** 3516-59.

Roege, K. E. & Kelly, W. L. (2009). Biosynthetic origins of the ionophore antibiotic indanomycin. *Org Lett,* **11,** 297-300.

Rosano, G. L. & Ceccarelli, E. A. (2014). Recombinant protein expression in *Escherichia coli*: Advances and challenges. *Front Microbiol,* **5,** 172.

Roujeinikova, A., Baldock, C., Simon, W. J., Gilroy, J., Baker, P. J., Stuitje, A. R., Rice, D. W., Slabas, A. R. & Rafferty, J. B. (2002). X-ray crystallographic studies on butyryl-ACP reveal flexibility of the structure around a putative acyl chain binding site. *Structure,* **10,** 825-35.

Roujeinikova, A., Simon, W. J., Gilroy, J., Rice, D. W., Rafferty, J. B. & Slabas, A. R. (2007). Structural studies of fatty acyl-(acyl carrier protein) thioesters reveal a hydrophobic binding cavity that can expand to fit longer substrates. *Journal of Molecular Biology,* **365,** 135-145.

Roy, A., Kucukural, A. & Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols,* **5,** 725-738.

Ruan, X., Pereda, A., Stassi, D. L., Zeidner, D., Summers, R. G., Jackson, M., Shivakumar, A., Kakavas, S., Staver, M. J., Donadio, S. & Katz, L. (1997). Acyltransferase domain substitutions in erythromycin polyketide synthase yield novel erythromycin derivatives. *J Bacteriol,* **179,** 6416-25.

Rückert, M. a. O. G. (2000). Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *Journal of the American Chemical Society,* **122,** 7793-7797.

Rusnak, F., Sakaitani, M., Drueckhammer, D., Reichert, J. & Walsh, C. T. (1991). Biosynthesis of the *Escherichia coli* siderophore enterobactin - sequence of the EntF gene, expression and purification of EntF, and analysis of covalent phosphopantetheine. *Biochemistry,* **30,** 2916-2927.

Sambrook, J., Fritsch, E. F. & Maniatis, T. 1989. *Molecular cloning : A laboratory manual,* Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

References

Samel, S. A., Schoenafinger, G., Knappe, T. A., Marahiel, M. A. & Essen, L. O. (2007). Structural and functional insights into a peptide bond-forming bidomain from a nonribosomal peptide synthetase. *Structure,* **15,** 781-792.

Samel, S. A., Wagner, B., Marahiel, M. A. & Essen, L. O. (2006). The thioesterase domain of the fengycin biosynthesis cluster: A structural base for the macrocyclization of a nonribosomal lipopeptide. *Journal of Molecular Biology,* **359,** 876-889.

Scaglione, J. B., Akey, D. L., Sullivan, R., Kittendorf, J. D., Rath, C. M., Kim, E. S., Smith, J. L. & Sherman, D. H. (2010). Biochemical and structural characterization of the tautomycetin thioesterase: Analysis of a stereoselective polyketide hydrolase. *Angewandte Chemie-International Edition,* **49,** 5726-5730.

Schetter, B. & Mahrwald, R. (2006). Modern aldol methods for the total synthesis of polyketides. *Angewandte Chemie-International Edition,* **45,** 7506-7525.

Schneider, A. & Marahiel, M. A. (1998). Genetic evidence for a role of thioesterase domains, integrated in or associated with peptide synthetases, in nonribosomal peptide biosynthesis in bacillus subtilis. *Archives of Microbiology,* **169,** 404-410.

Schneider, T. L., Shen, B. & Walsh, C. T. (2003). Oxidase domains in epothilone and bleomycin biosynthesis: Thiazoline to thiazole oxidation during chain elongation. *Biochemistry,* **42,** 9722-30.

Schoenafinger, G., Schracke, N., Linne, U. & Marahiel, M. A. (2006). Formylation domain: An essential modifying enzyme for the nonribosomal biosynthesis of linear gramicidin. *Journal of the American Chemical Society,* **128,** 7406-7407.

Schwarzer, D., Finking, R. & Marahiel, M. A. (2003). Nonribosomal peptides: From genes to products. *Nat Prod Rep,* **20,** 275-87.

Schwecke, T., Aparicio, J. F., Molnar, I., Konig, A., Khaw, L. E., Haydock, S. F., Oliynyk, M., Caffrey, P., Cortes, J., Lester, J. B., Bohm, G. A., Staunton, J. & Leadlay, P. F. (1995). The biosynthetic gene-cluster for the polyketide immunosuppressant rapamycin. *Proceedings of the National Academy of Sciences of the United States of America,* **92,** 7839-7843.

Schwieters, C. D., Kuszewski, J. J. & Clore, G. M. (2006). Using xplor-nih for NMR molecular structure determination. *Progress in Nuclear Magnetic Resonance Spectroscopy,* **48,** 47-62.

Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. (2003). The xplor-nih NMR molecular structure determination package. *Journal of Magnetic Resonance,* **160,** 65-73.

Sezonov, G., Joseleau-Petit, D. & D'ari, R. (2007). *Escherichia coli* physiology in Luria-Bertani broth. *J Bacteriol,* **189,** 8746-9.

Sgourakis, N. G., Lange, O. F., Dimaio, F., Andre, I., Fitzkee, N. C., Rossi, P., Montelione, G. T., Bax, A. & Baker, D. (2011). Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *Journal of the American Chemical Society,* **133,** 6288-6298.

Sgourakis, N. G., Natarajan, K., Ying, J., Vogeli, B., Boyd, L. F., Margulies, D. H. & Bax, A. (2014). The structure of mouse cytomegalovirus m04 protein obtained from sparse NMR data reveals a conserved fold of the m02-m06 viral immune modulator family. *Structure,* **22,** 1263-73.

References

Shen, Y. & Bax, A. (2013). Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J Biomol NMR,* **56,** 227-41.

Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. (2009a). Talos+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of biomolecular NMR,* **44,** 213-23.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G. H., Eletsky, A., Wu, Y. B., Singarapu, K. K., Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D. & Bax, A. (2008). Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci U S A,* **105,** 4685-4690.

Shen, Y., Vernon, R., Baker, D. & Bax, A. (2009b). *De novo* protein structure generation from incomplete chemical shift assignments. *J Biomol NMR,* **43,** 63-78.

Sieber, S. A., Linne, U., Hillson, N. J., Roche, E., Walsh, C. T. & Marahiel, M. A. (2002). Evidence for a monomeric structure of nonribosomal peptide synthetases. *Chemistry & Biology,* **9,** 997-1008.

Sieber, S. A. & Marahiel, M. A. (2003). Learning from nature's drug factories: Nonribosomal synthesis of macrocyclic peptides. *Journal of bacteriology,* **185,** 7036-43.

Sieber, S. A. & Marahiel, M. A. (2005). Molecular mechanisms underlying nonribosomal peptide synthesis: Approaches to new antibiotics. *Chemical reviews,* **105,** 715-38.

Skold, O. (2000). Sulfonamide resistance: Mechanisms and trends. *Drug Resistance Updates,* **3,** 155-160.

Smith, G. A., Smith, K. A. & Williams, D. H. (1975). Structural studies on antibiotic vancomycin - evidence for presence of modified phenyglycine and β-hydroxytyrosine units. *Journal of the Chemical Society-Perkin Transactions 1,* **2108-2115.**

Smith, S. & Tsai, S. C. (2007). The type I fatty acid and polyketide synthases: A tale of two megasynthases. *Natural Product Reports,* **24,** 1041-1072.

Srisailam, S., Lukin, J. A., Yee, L., Semesi, A. & Arrowsmith, C. H. (2006). Solution structure of acyl carrier protein from *Nitrosomonas europaea*. *Proteins-Structure Function and Bioinformatics,* **64,** 800-803.

Stachelhaus, T., Huser, A. & Marahiel, M. A. (1996). Biochemical characterization of peptidyl carrier protein (PCP), the thiolation domain of multifunctional peptide synthetases. *Chemistry & biology,* **3,** 913-21.

Stachelhaus, T. & Marahiel, M. A. (1995). Modular structure of genes encoding multifunctional peptide synthetases required for nonribosomal peptide synthesis. *FEMS Microbiol Lett,* **125,** 3-14.

Stachelhaus, T., Mootz, H. D., Bergendahl, V. & Marahiel, M. A. (1998). Peptide bond formation in nonribosomal peptide biosynthesis. Catalytic role of the condensation domain. *J Biol Chem,* **273,** 22773-81.

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

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

References

Stachelhaus, T. & Walsh, C. T. (2000). Mutational analysis of the epimerization domain in the initiation module pheate of gramicidin S synthetase. *Biochemistry,* **39,** 5775-87.

Starai, V. J. & Escalante-Semerena, J. C. (2004). Acetyl-coenzyme A synthetase (AMP forming). *Cellular and Molecular Life Sciences,* **61,** 2020-2030.

Staunton, J., Caffrey, P., Aparicio, J. F., Roberts, G. A., Bethell, S. S. & Leadlay, P. F. (1996). Evidence for a double-helical structure for modular polyketide synthases. *Nat Struct Biol,* **3,** 188-92.

Staunton, J. & Weissman, K. J. (2001). Polyketide biosynthesis: A millennium review. *Nat Prod Rep,* **18,** 380-416.

Staunton, J. & Wilkinson, B. (1997). Biosynthesis of erythromycin and rapamycin. *Chemical Reviews,* **97,** 2611-2629.

Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol,* **185,** 60-89.

Sundermann, U., Bravo-Rodriguez, K., Klopries, S., Kushnir, S., Gomez, H., Sanchez-Garcia, E. & Schulz, F. (2013). Enzyme-directed mutasynthesis: A combined experimental and theoretical approach to substrate recognition of a polyketide synthase. *ACS Chemical Biology,* **8,** 443-450.

Tang, C., Ghirlando, R. & Clore, G. M. (2008). Visualization of transient ultra-weak protein self-association in solution using paramagentic relaxation enhancement. *Journal of the American Chemical Society*, **130**, 4048-4056.

Tang, L., Fu, H. & Mcdaniel, R. (2000). Formation of functional heterologous complexes using subunits from the picromycin, erythromycin and oleandomycin polyketide synthases. *Chem Biol,* **7,** 77-84.

Tang, Y., Chen, A. Y., Kim, C.-Y., Cane, D. E. & Khosla, C. (2007). Structural and mechanistic analysis of protein interactions in module 3 of the 6-deoxyerythronolide B synthase. *Chemistry & Biology,* **14,** 931-943.

Tang, Y., Kim, C. Y., Mathews, Ii, Cane, D. E. & Khosla, C. (2006). 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,** 11124-9.

Tanovic, A., Samel, S. A., Essen, L. O. & Marahiel, M. A. (2008). Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science,* **321,** 659-663.

Teng, Q. (2005). Structural biology: Practical NMR applications. *Structural Biology: Practical NMR Applications*. New York City: Springer publishing.

Thomas, I., Martin, C. J., Wilkinson, C. J., Staunton, J. & Leadlay, P. F. (2002). Skipping in a hybrid polyketide synthase. Evidence for ACP-to-ACP chain transfer. *Chem Biol,* **9,** 781-7.

Timms, N., Windle, C. L., Polyakova, A., Ault, J. R., Trinh, C. H., Pearson, A. R., Nelson, A. & Berry, A. (2013). Structural insights into the recovery of aldolase activity in *N*-acetylneuraminic acid lyase by replacement of the catalytically active lysine with γ-thialysine by using a chemical mutagenesis strategy. *Chembiochem,* **14,** 474-81.

Tjandra, N. & Bax, A. (1997a). Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science,* **278,** 1111-1114.

Tjandra, N. & Bax, A. (1997b). Large variations in ^{13}C(alpha) chemical shift anisotropy in proteins correlate with secondary structure. *Journal of the American Chemical Society,* **119,** 9576-9577.

Tjandra, N. & Bax, A. (1997c). Measurement of dipolar contributions to 1JCH splittings from magnetic-field dependence of J modulation in two-dimensional NMR spectra. *J Magn Reson,* **124,** 512-5.

Tjandra, N., Grzesiek, S. & Bax, A. (1996). Magnetic field dependence of nitrogen-proton J splittings in ^{15}N-enriched human ubiquitin resulting from relaxation interference and residual dipolar coupling. *Journal of the American Chemical Society,* **118,** 6264-6272.

Tobias, J. W., Shrader, T. E., Rocap, G. & Varshavsky, A. (1991). The n-end rule in bacteria. *Science,* **254,** 1374-1377.

Tolia, N. H. & Joshua-Tor, L. (2006). Strategies for protein coexpression in *Escherichia coli*. *Nature methods,* **3,** 55-64.

Torbet, J. & Maret, G. (1979). Fibers of highly oriented Pf1 bacteriophage produced in a strong magnetic-field. *Journal of Molecular Biology,* **134,** 843-845.

Tran, L., Broadhurst, R. W., Tosin, M., Cavalli, A. & Weissman, K. J. (2010). Insights into protein-protein and enzyme-substrate interactions in modular polyketide synthases. *Chemistry & biology,* **17,** 705-16.

Trauger, J. W., Kohli, R. M., Mootz, H. D., Marahiel, M. A. & Walsh, C. T. (2000). Peptide cyclization catalysed by the thioesterase domain of tyrocidine synthetase. *Nature,* **407,** 215-218.

Tsai, S. C. & Ames, B. D. (2009). Structural enzymology of polyketide synthases. *Complex Enzymes in Microbial Natural Product Biosynthesis, Part B: Polyketides, Aminocoumarins and Carbohydrates,* **459,** 17-47.

Tsai, S. C., Lu, H. X., Cane, D. E., Khosla, C. & Stroud, R. M. (2002). Insights into channel architecture and substrate specificity from crystal structures of two macrocycle-forming thioesterases of modular polyketide synthases. *Biochemistry,* **41,** 12598-12606.

Tsai, S. C., Miercke, L. J., Krucinski, J., Gokhale, R., Chen, J. C., Foster, P. G., Cane, D. E., Khosla, C. & Stroud, R. M. (2001). Crystal structure of the macrocycle-forming thioesterase domain of the erythromycin polyketide synthase: Versatility from a unique substrate channel. *Proc Natl Acad Sci U S A,* **98,** 14808-13.

Tseng, C. C., Bruner, S. D., Kohli, R. M., Marahiel, M. A., Walsh, C. T. & Sieber, S. A. (2002). Characterization of the surfactin synthetase C-terminal thioesterase domain as a cyclic depsipeptide synthase. *Biochemistry,* **41,** 13350-13359.

Tuan, J. S., Weber, J. M., Staver, M. J., Leung, J. O., Donadio, S. & Katz, L. (1990). Cloning of genes involved in erythromycin biosynthesis from *Saccharopolyspora erythraea* using a novel actinomycete - *Escherichia coli* cosmid. *Gene,* **90,** 21-9.

Turgay, K., Krause, M. & Marahiel, M. A. (1992). 4 homologous domains in the primary structure of GrsB are related to domains in a superfamily of adenylate-forming enzymes. *Molecular Microbiology,* **6,** 529-546.

Unden, G., Becker, S., Bongaerts, J., Holighaus, G., Schirawski, J. & Six, S. (1995). O_2-sensing and O_2-dependent gene regulation in facultatively anaerobic bacteria. *Arch Microbiol,* **164,** 81-90.

Upadhyay, S. K., Misra, A., Srivastava, R., Surolia, N., Surolia, A. & Sundd, M. (2009). Structural insights into the acyl intermediates of the *Plasmodium*

falciparum fatty acid synthesis pathway: The mechanism of expansion of the acyl carrier protein core. *The Journal of biological chemistry,* **284,** 22390-400.

Vagin, A. & Teplyakov, A. (1997). Molrep: An automated program for molecular replacement. *Journal of Applied Crystallography,* **30,** 1022-1025.

Velkov, T., Thompson, P. E., Nation, R. L. & Li, J. (2010). Structure-activity relationships of polymyxin antibiotics. *Journal of Medicinal Chemistry,* **53,** 1898-1916.

Villiers, B. R. M. & Hollfelder, F. (2009). Mapping the limits of substrate specificity of the adenylation domain of TycA. *Chembiochem,* **10,** 671-682.

Vilotijevic, I. & Jamison, T. F. (2009). Epoxide-opening cascades in the synthesis of polycyclic polyether natural products. *Angewandte Chemie-International Edition,* **48,** 5250-5281.

Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, P., Ulrich, E. L., Markley, J. L., Ionides, J. & Laue, E. D. (2005). The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins-Structure Function and Bioinformatics,* **59,** 687-696.

Waksman, S. A., Schatz, A. & Reynolds, D. M. (2010). Production of antibiotic substances by actinomycetes. *Ann N Y Acad Sci,* **1213,** 112-24.

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

Walsh, C. T., Chen, H., Keating, T. A., Hubbard, B. K., Losey, H. C., Luo, L., Marshall, C. G., Miller, D. A. & Patel, H. M. (2001). Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. *Current opinion in chemical biology,* **5,** 525-34.

Walsh, C. T., Garneau-Tsodikova, S. & Howard-Jones, A. R. (2006). Biological formation of pyrroles: Nature's logic and enzymatic machinery. *Natural product reports,* **23,** 517-31.

Walsh, C. T., Gehring, A. M., Weinreb, P. H., Quadri, L. E. & Flugel, R. S. (1997). Post-translational modification of polyketide and nonribosomal peptide synthases. *Current opinion in chemical biology,* **1,** 309-15.

Walsh, C. T., Liu, J., Rusnak, F. & Sakaitani, M. (1990). Molecular studies on enzymes in chorismate metabolism and the enterobactin biosynthetic-pathway. *Chemical Reviews,* **90,** 1105-1129.

Wang, H., Liu, N., Xi, L., Rong, X., Ruan, J. & Huang, Y. (2011). Genetic screening strategy for rapid access to polyether ionophore producers and products in *Actinomycetes*. *Appl Environ Microbiol,* **77,** 3433-42.

Watanakunakorn, C. (1981). The antibacterial action of vancomycin. *Reviews of Infectious Diseases,* **3,** S210-S215.

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

Weber, J. M., Leung, J. O., Maine, G. T., Potenz, R. H. B., Paulus, T. J. & Dewitt, J. P. (1990). Organization of a cluster of erythromycin genes in *Saccharopolyspora erythraea*. *Journal of Bacteriology,* **172,** 2372-2383.

Weber, T., Baumgartner, R., Renner, C., Marahiel, M. A. & Holak, T. A. (2000). Solution structure of PCP, a prototype for the peptidyl carrier domains of modular peptide synthetases. *Structure,* **8,** 407-18.

References

Weissman, K. J. (2004). Polyketide biosynthesis: Understanding and exploiting modularity. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences,* **362,** 2671-90.

Weissman, K. J. & Leadlay, P. F. (2005). Combinatorial biosynthesis of reduced polyketides. *Nat Rev Microbiol,* **3,** 925-36.

Weissman, K. J. & Muller, R. (2008). Protein-protein interactions in multienzyme megasynthetases. *Chembiochem : a European journal of chemical biology,* **9,** 826-48.

Whicher, J. R., Dutta, S., Hansen, D. A., Hale, W. A., Chemler, J. A., Dosey, A. M., Narayan, A. R. H., Hakansson, K., Sherman, D. H., Smith, J. L. & Skiniotis, G. (2014). Structural rearrangements of a polyketide synthase module during its catalytic cycle. *Nature,* **510,** 560-564.

Williamson, R. A., Carr, M. D., Frenkiel, T. A., Feeney, J. & Freedman, R. B. (1997). Mapping the binding site for matrix metalloproteinase on the N-terminal domain of the tissue inhibitor of metalloproteinases-2 by NMR chemical shift perturbation. *Biochemistry,* **36,** 13882-13889.

Winter, G. (2010). Xia2: An expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography,* **43,** 186-190.

Wishart, D. S., Arndt, D., Berjanskii, M., Tang, P., Zhou, J. & Lin, G. (2008). CS23D: A web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Research,* **36,** W496-W502.

Wright, G. D. (2007). The antibiotic resistome: The nexus of chemical and genetic diversity. *Nature Reviews Microbiology,* **5,** 175-86.

Wuthrich, K., Wider, G., Wagner, G. & Braun, W. (1982). Sequential resonance assignments as a basis for determination of spatial protein structures by high-resolution proton nuclear magnetic resonance. *Journal of Molecular Biology,* **155,** 311-319.

Xu, G. Y., Tam, A., Lin, L., Hixon, J., Fritz, C. C. & Powers, R. (2001). Solution structure of *B. subtilis* acyl carrier protein. *Structure,* **9,** 277-87.

Xue, Q., Ashley, G., Hutchinson, C. R. & Santi, D. V. (1999). A multiplasmid approach to preparing large libraries of polyketides. *Proceedings of the National Academy of Sciences of the United States of America,* **96,** 11740-5.

Yamazaki, T., Formankay, J. D. & Kay, L. E. (1993). 2-Dimensional NMR experiments for correlating $^{13}C\beta$ and $^{1}H\delta/\varepsilon$ chemical-shifts of aromatic residues in ^{13}C-labeled proteins via scalar couplings. *Journal of the American Chemical Society,* **115,** 11054-11055.

Yang, Z. H. & Attygalle, A. B. (2007). LC/MS characterization of undesired products formed during iodoacetamide derivatization of sulfhydryl groups of peptides. *Journal of Mass Spectrometry,* **42,** 233-243.

Yocum, R. R., Rasmussen, J. R. & Strominger, J. L. (1980). The mechanism of action of penicillin - penicillin acylates the active-site of *Bacillus-stearothermophilus* D-alanine carboxypeptidase. *Journal of Biological Chemistry,* **255,** 3977-3986.

Yonus, H., Neumann, P., Zimmermann, S., May, J. J., Marahiel, M. A. & Stubbs, M. T. (2008). Crystal structure of DltA impications for the reaction mechanism of nonribosomal peptide synthetase adenylation domains. *Journal of Biological Chemistry,* **283,** 32484-32491.

Zhang, D., Nair, M. G., Murry, M. & Zhang, Z. (1997). Insecticidal activity of indanomycin. *J Antibiot (Tokyo),* **50,** 617-20.

References

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *Bmc Bioinformatics,* **9**.

Zheng, J., Gay, D. C., Demeler, B., White, M. A. & Keatinge-Clay, A. T. (2012). Divergence of multimodular polyketide synthases revealed by a didomain structure. *Nature Chemical Biology*.

Zheng, J. & Keatinge-Clay, A. T. (2011). Structural and functional analysis of C2-type ketoreductases from modular polyketide synthases. *Journal of molecular biology,* **410,** 105-17.

Zheng, J. T., Taylor, C. A., Piasecki, S. K. & Keatinge-Clay, A. T. (2010). Structural and functional analysis of a-type ketoreductases from the amphotericin modular polyketide synthase. *Structure,* **18,** 913-922.

Zornetzer, G. A., Fox, B. G. & Markley, J. L. (2006). Solution structures of spinach acyl carrier protein with decanoate and stearate. *Biochemistry,* **45,** 5217-27.

Zuiderweg, E. R. P. & Fesik, S. W. (1989). Heteronuclear 3-Dimensional NMR-spectroscopy of the inflammatory protein C5A. *Biochemistry,* **28,** 2387-2391.

Zweckstetter, M. & Bax, A. (2000). Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. *Journal of the American Chemical Society,* **122,** 3791-3792.