Biosynthesis and discovery of compounds with activity against *Acinetobacter baumannii*

by

Daniel Griffiths

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Supervisors: Dr. Józef R. Lewandowski, Prof. Gregory L. Challis

Department of Chemistry

September 2016

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisors, Józef Lewandowski and Greg Challis, for providing the opportunity to work on engaging multidisciplinary projects, and for introducing me to the field of natural product biosynthesis. Together, you have provided sound guidance, enthusiasm, encouragement, an encyclopedic knowledge of natural products, humour, relentless optimism, patience and tolerance. I would like to thank all Challis group and Lewandowski group members (past and present) for creating a positive and cooperative working environment, including Lona Alkhalaf, Daniel Zabala-Alvarez, Matthew Jenner, Douglas Roberts, Simone Kosol, Joleen Masschelein, Chuan Huang, Yousef Dashti, Emzo De Los Santos, Angelo Gallo, Sławomir Potocki, Ruby Awodi, Shanshan Zhou, Rebin Salih, Gideon Idowu, Jade Ronan, Joshua Cartwright, Rakesh Saroay, Richard Gibson, Marrianne Costa, Xinyun Jian, Christian Hobson, Chris Perry, Alma Svatoš, Matt Beech, Panward Prasongpholchai.

Thanks to Joleen Masschelein for training and collaboration during LC-MS analysis of condensation assays, and vibroxin production, extraction and purification. To Douglas Roberts and Matthew Jenner for their persistent enthusiasm, and much of my chemistry and natural product biosynthesis education, through informative and outlandish conversations. To Paulina Sydor for providing the biochemical basis of enacyloxin chain release. To Vilmos Fülöp, Dean Rea, Sheryl-Tsai, Tim Valentic for the crystal structure of the condensation domain. To Lona Alkhalaf for proof-reading and feedback on this thesis. To MOAC/MAS DTC and the EPSRC for funding. To the members of my DTC cohort. To Lijiang Song and Ivan Prokes for dutifully facilitating MS and NMR experiments, respectively. To the members of the Chemical Biology Facility.

Thank you to Frances, family and friends for support.

Declaration

I confirm that my thesis has been prepared in accordance with the University's guidelines on the presentation of a research thesis. The thesis has been composed by myself and has not been submitted in any previous application for any degree. The work presented in this thesis was undertaken by myself except where otherwise stated.

Abstract

Enacyloxin is a natural product from *Burkholderia ambifaria* that exhibits antimicrobial activity against the multi-drug resistant pathogen *Acinetobacter baumannii*. The biosynthesis of enacyloxin involves a hybrid modular polyketide synthase (PKS)-nonribosomal peptide synthetase, performing sequential thiol-template catalysis. The process of chain release encompasses an unusual domain architecture that includes an acyl carrier protein domain, a non-elongating ketosynthase domain, a peptidyl carrier protein (PCP) domain and a standalone condensation (C) domain. Specific protein-protein interactions have been shown to facilitate chain release, by the identification, and characterisation, of a pair of communication-mediating domains across a PCP-C protein boundary. This class of communication-mediating domains was also identified within other biosynthetic pathways using a hidden Markov model, which demonstrates the widespread occurrence of this type of intermolecular interaction. A hybrid intermolecular interaction was formed using non-native communication-mediating domains from watasemycin and enacyloxin biosynthetic pathways, which demonstrate their inherent compatibility.

A search for enacyloxin analogues revealed a homologous biosynthetic gene cluster within rhizosphere-associated *Vibrio rhizosphaerae*. A novel enacyloxin analogue, vibroxin, was isolated for characterisation by high-resolution mass spectrometry and nuclear magnetic resonance spectroscopy. The relative stereochemistry of a cyclohexane moiety is consistent with that of enacyloxin, and the assignment of other stereo-centers are proposed from ketoreductase domains. Analysis of PKS proteins revealed a methyltransferase domain within the loading module, where isotopic feeding experiments support a dimethylated starter unit. Fragments of acyltransferase (AT) domains, identified within PKS proteins, indicate a transition state between a *cis*-AT PKS and a *trans*-AT PKS. Vibroxin exhibits potent antimicrobial activity against *Acinetobacter baumannii*, demonstrating the benefit of the genomics-driven approach for the discovery of novel bioactive compounds.

Abbreviations

α-KG	α-ketoglutarate
A	Adenylation
ACP	Acyl carrier protein
AT	Acyl transferase
Atd	*Trans*-AT docking
ATP	Adenosine triphosphate
Bcc	*Burkholderia* cepecia complex
bL	β-lactamase-like
BME	Basal medium Eagle
BSM	Basal salt medium
BSM2S	Basal salt medium with 2% NaCl
C	Condensation (or Carbon)
CAL	Coenzyme A ligase
C_d	Dual condensing and epimerising
C_{lcl}	Condensation domain that condenses two L-amino acids
CLSI	Clinical Laboratory and Standards Institute
COM	Communication-mediating
COM_{2C}	C-terminal type 2 COM
COM_{2N}	N-terminal type 2 COM
COSY	Correlation spectroscopy
CP	Carrier protein
Cy	Heterocyclisation
δ	Chemical shift (or ω)
DC	Glutamate decarboxylase
dd	Docking domain

DH	Dehydratase
DHCCA	Dihydroxycyclohexane carboxylic acid
dMT	Dimerisation subdomain of MT ('acyl-recruitment')
DNA	Deoxyribonucleic acid
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid
E	Epimerisation
E.I.C	Extracted ion chromatogram
EF-Tu	Elongation-Factor Thermo-unstable
ER	Enoyl reductase
ESKAPE	*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter cloacae*
$FADH_2$	Flavin adenine dinucleotide (hydroquinone form)
FAS	Fatty acid synthase
GE	General electric
GNAT	Malonyl transferase/decarboxylase (Gcn5-*N*-aceyltransferase family)
H	Halogenase (or Hydrogen)
HMBC	Heteronuclear multiple-bond correlation spectroscopy
HMM	Hidden Markov model
HPLC	High-performance liquid chromatography
HSQC	Heteronuclear single quantum coherence
IPTG	isopropyl β-D-1-thiogalactopyranoside
KAc	C-terminal KS-AT linker motif
KAn	N-terminal KS-AT linker motif
KLD	Kinase-Ligase-DpnI
KR	Ketoreductase
KS	Ketosynthase
KS^0	ACP *trans*-acylase (non-elongating KS)
KSc	C-terminal KS motif
KS^Q	Ketosynthase with active site glutamine (decarboxylating)

LB	Lysogeny broth
LC-HR/MS	Liquid-chromatography high-resolution/mass-spectrometry
MALDI-ToF	Matrix-assisted laser desorption/ionization-time of flight
MBC	Minimal bactericidal concentration
MH	Müller-Hinton
MIC	Minimum inhibitory concentration
MRSA	Methicillin-resistant *Staphylococcus aureus*
MT	Methyltransferase
MS	Mass spectrometry
NADPH	Reduced nicotinamide adenine dinucleotide phosphate
NEB	New England Biolabs
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
NRPS	Nonribosomal peptide synthetase
OD	Optical density
Ox	Oxidase
PCP	Peptidyl carrier protein
PCR	Polymerase chain reaction
PDB	Protein data bank
PKS	Polyketide synthase
PMSF	Phenylmethylsulfonyl fluoride
PPTase	4'-phosphopantetheinyl transferase
PQQ	Pyrroloquinoline quinone
SAH	*S*-adenosyl-homocysteine
SAM	*S*-adenosyl-methionine
SAT	Starter-unit:ACP transacylase
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
sfp	Pantetheinyl transferase from surfactin NRPS
sMT	Starter methyltransferase
SOC	Super optimal broth with Catabolite repression
TA	ACP *trans*-acylase

TCEP	Tris(2-carboxyethyl)phosphine
TE	Thioesterase
TFA	Trifluoroacetic acid
THCCA	Trihydroxycyclohexane carboxylic acid
TIM	Triosephosphate isomerase
TOCSY	Total correlation spectroscopy
TR	Thioester-reductase
TSB	Tryptic soy broth
TSB2S	Tryptic soy broth and 2% NaCl
TTC	2,3,5-triphenyltetrazolium chloride
U-^{15}N	Uniformly labelled with ^{15}N isotope
U-^{13}C,^{15}N	Uniformly labelled with ^{13}C and ^{15}N isotopes

Chapter 1

Introduction

1.1 Natural products

The discovery of novel antibiotics over the last 50 years has declined dramatically, whilst multidrug-resistant bacterial pathogens, like methicillin-resistant *Staphylococcus aureus*[1] (MRSA), *Pseudomonas aeruginosa*,[2] and *Acinetobacter baumannii*[3] have proliferated rapidly. This constitutes an ever-growing healthcare challenge, and is recognised as a global threat.[4] Without the development of novel antibiotics, and regulated prescription, many standard medical treatments, that we depend on, will fail. Currently, the majority of clinically important antimicrobial agents are derived from bacterial natural products.[5] The discovery of novel natural products will provide important drug-leads, to combat the rise of multidrug-resistant bacteria. A detailed understanding of natural product biosynthesis can provide motivation for rational mutasynthetic approaches to develop drug-leads, or to engineer biosynthetic pathways, developing novel compounds.

1.2 Biosynthesis using megasynthases

Many natural products are biosynthesised via enzymatic assembly lines. Similar to product assembly lines, where each worker has a specific component to sequentially add to a product, microorganisms contain enzymes, acting as workers,

which produce an antibiotic product. These enzymes recruit chemical building blocks (commonly malonate from malonyl CoA, or an amino acid), append them to the product, and pass the product to the following enzyme in a defined order, to manufacture an antibiotic. A thorough understanding of natural product assembly provides opportunities to venture towards rationally engineering these processes, to combat the challenge of multi-drug resistant microbial pathogens.

Two foundational classes of natural products are derived from the acetate pathway: fatty acids and polyketides, produced respectively by fatty acid synthases (FASs), and polyketide synthases (PKSs). Fatty acid synthesis is fundamental to the production of energy and cell structure, which designates it as a member of primary metabolism, being essential for life. FAS I is a dimeric multienzyme protein[6] that sequentially synthesises aliphatic acids. Each enzymatic domain from FAS is well-characterised both structurally and biochemically, from a diverse range of organisms.[6–9] Akin to FAS, PKSs can also resemble multienzyme proteins[10] or can be composed of independent enzymes acting in succession[11] or iteratively.[12] Alternatively, natural products that are synthesised from amino acids, can be produced by a nonribosomal peptide synthetase (NRPS).

1.3 Domain organisation and function

1.3.1 PKS domains

Type 1 PKSs are large, highly modular proteins, where enzymatic domains catalyse sequential reactions on carrier protein-bound intermediates.[10] A prerequisite for PKS biosynthesis is the post-translational modification of *apo*-acyl carrier proteins (ACPs) to their *holo* form, using a phosphopantetheinyltransferase (PPTase) as shown in Fig. 1.1a. Archetypical PKSs load malonyl or methylmalonyl extender units, based on the substrate specificity of acyl transferase (AT) domains, onto a downstream *holo*-ACP domain (Fig. 1.1bi). Ketosynthase (KS) domains act to elongate the polyketide by catalysing decarboxylative Claisen con-

densation of (methyl) malonyl-ACP and the upstream thioester (Fig. 1.1bii). In the case of an extender unit resulting in malonyl-ACP, methyltransferase (MT) domains can methylate the substrate after condensation (Fig. 1.1biii).[13] Additional domains can act on the growing natural product, resulting in changes at the β position; reduction by a ketoreductase (KR) domain (Fig. 1.1biv), subsequent dehydration by a dehydratase (DH) domain (Fig. 1.1bv), and enoyl reduction by an enoyl reductase (ER) domain (Fig. 1.1bvi). Chain release, from the PKS, is typically accomplished by a thioesterase (TE) domain, which can also facilitate macrocyclisation.

1.3.2 NRPS domains

The typical architecture of an NRPS module includes an adenylation (A) domain, which determines the amino acid specificity, and facilitates amino acid loading onto a peptidyl carrier protein (PCP) domain, through an ATP-dependent process. A condensation (C) domain condenses amino acids, to form an amide bond between the thioester of the growing chain from the previous module, and the amino group of the amino acid-PCP from its own module. Additional enzymes may act within an NRPS, modifying loaded amino acids. The most common of which is the epimerization (E) domain, that epimerizes at the α position of the proximal PCP-bound amino acid. NRPSs can also form hybrid biosynthetic pathways with PKSs, either as distinct NRPS and PKS proteins, or as chimeric proteins.

1.3.3 PKS/NRPS product prediction

The sequential and combinatorial nature of PKSs and NRPSs allows the collection of domains into modules, corresponding to one round of chain elongation of a carrier protein-tethered intermediate. The nature of the elongation unit is determined by the specificity of AT domains in PKSs, and A domains in NRPS. For each type of loading domain, the substrate can be predicted based upon

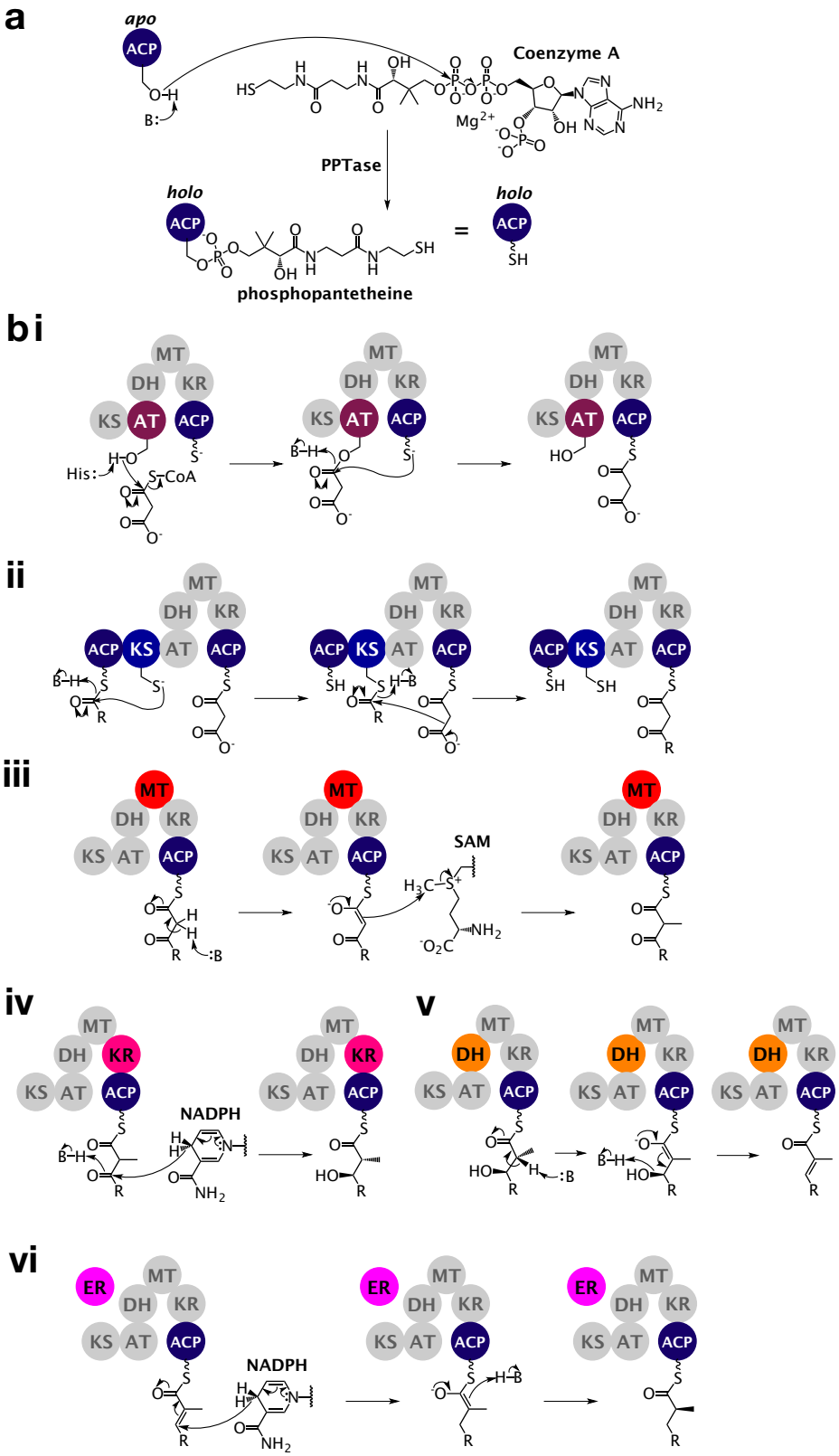

Figure 1.1: a) Conversion of *apo*-ACP to *holo*-ACP, b) mechanisms of PKS domains, i) acyltransferase, ii) ketosynthase, iii) carbon-methyl transferase, iv) ketoreductase, v) dehydratase, vi) enoyl reductase.

amino acids within their active site.[14,15] The auxiliary KR domain, governs the stereochemistry of the β-OH (and α-Me if present) at a thioester. The stereospecificity of any given KR domain can also be predicted based on active site amino acids.[16] For PKS modules containing a trio of KR, DH, and ER domains, the resulting stereochemistry of an α-Me (if present) can be putatively assigned, based on the presence of a conserved tyrosine amino acid[17] within the ER domain. Together, the predictions of enzyme specificity, classification, and their sequential occurrence within modules, provides a useful prediction of the core molecular structure of a PKS, or NRPS, derived natural product.

1.4 Docking and communication-mediating domains

The genetic organisation of PKS and NRPS genes typically adhere to a 'colinearity rule', whereby genes are arranged in agreement with the direction of the thiol-template stages of a biosynthetic pathway. This biosynthetic order needs to be maintained after translation, to ensure that a particular biosynthetic route is followed. Each multi-enzymatic protein must selectively interact with an enzyme corresponding to the subsequent stage of biosynthesis, passing a chemical intermediate, whilst avoiding other enzymes within any given biosynthetic pathway. A justification for thiol-template biosynthetic sequential order has been identified, in part, in the form of N- and C-terminal docking domains (dd) in PKSs, and communication-mediating (COM) domains in NRPSs.

1.4.1 Structures of docking and COM domains

To date, docking domains and COM domains have been structurally characterised from the biosynthetic pathways of curacin,[18] erythromycin,[19] pikromycin,[20] virginiamycin,[21] macrolactin,[22] tubulysin,[23] surfactin[24] (Fig. 1.2), and enacyloxin (chapter 2). N-terminal docking domains, found at ACP-KS module bound-

aries, typically form a coiled-coil structure. The C-terminal binding partner also consists of one, or two, helices that bind across the dimeric interface. Creating hybrid pairs of docking domains, and mutants have revealed indications of docking domain selectivity, via key electrostatic interactions.[18] The selectivity of docking domains, is most apparent in a study that measures binding between each docking domain from erythromycin PKS and pikromycin PKS. It shows that the binding affinity of native docking domain pairs interact with binding affinities near 100 μM, whereas no binding was observed with non-native pairings of docking domains.[20] Interestingly, the addition of an adjacent KS domain has been reported to provide a twofold increase in the affinity of the intermolecular interaction, and addition of a whole module provides a ninefold increase.[20]

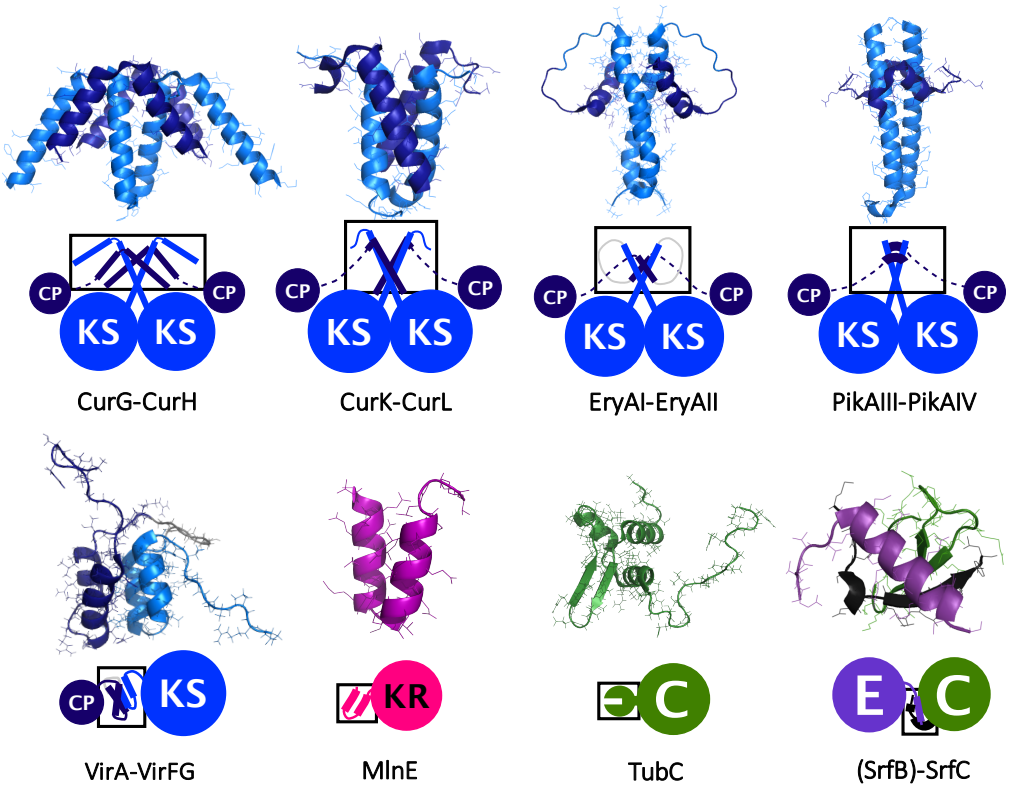

Figure 1.2: Structures of docking domains and COM domains from the following PDB coordinates: 4MYY, 4MYZ, 1PZR, 3F5H, 2N5D, 5D2E (single protomer shown), 2JUG (single protomer shown from a dimeric species), 2VSQ (mimic of a COM domain interaction).

Another type of docking domain structure, found across split-module PKS

boundaries, possesses a helix-turn-helix structure (*e.g.* VirA-VirFG and MlnE), which forms a four-helical bundle in its heterodimeric complex. This class of docking domain appears across split modules, with different types of PKS enzymes found adjacently. The low micromolar dissociation constant measured for this class of docking domain, support the capability of mediating the association of neighbouring domains.[22] Truncation and swapping docking domains show their importance for forming protein complexes, and their pairwise portability.[22]

A structure of an N-terminal COM domain is reported from SrfC within the surfactin NRPS, which forms a β-sheet structure with the adjacent condensation domain. A mimic of the corresponding C-terminal interaction partner, within SrfB, forms an α-helix that makes contacts with the N-terminal COM domain, as well as the condensation domain. At a PCP-C protein boundary (TubB-TubC from tubulysin NRPS), an isolated N-terminal interaction domain was structurally characterised, which revealed a distinctively different class of COM domain (herein referred to as 'type 2') exhibiting a $\alpha\beta\beta\alpha\alpha$ fold.

1.5 NMR characterisation of protein domains

Small protein domains from PKSs and NRPSs are ideal candidates for structural characterisation by protein NMR spectroscopy. Docking domains,[19,21] COM domains,[23] and carrier proteins,[25] represent types of small biosynthetic protein domains that have been structurally characterised by NMR. Carrier proteins display inherent conformational flexibility, making them less desirable for other common biophysical methods, such as X-ray crystallography. Protein NMR spectroscopy can uncover atom specific details of protein domains, describing protein interactions in solution, by monitoring changes in the electromagnetic environment of atoms during a domain-domain interaction. This can provide a molecular-level description of significant interactions, during the sequential stages of enacyloxin chain release.

Protein NMR spectroscopy typically relies upon observing ^{1}H, ^{15}N and ^{13}C

nuclei. To improve sensitivity and enable the use of multidimensional experiments, proteins are often isotopically labelled with ^{15}N and ^{13}C isotopes. One of the most frequently used solution NMR experiments is the two-dimensional ^{1}H-^{15}N heteronuclear single quantum coherence spectroscopy (HSQC) sequence, where each ^{1}H-^{15}N correlation signal is assigned to a specific amino acid residue in the primary sequence of the protein. The HSQC spectrum provides a snapshot of the electromagnetic environment for each ^{1}H-^{15}N bond within a protein, in the form of chemical shift values. Since the chemical shifts depend on their local environment, perturbations, or changes in cross peak intensity, in the presence of a binding partner can be used to map interaction sites with other proteins.

1.6 Enacyloxin PKS-NRPS

PKSs and NRPSs have been identified in bacteria, fungi, plants and some animal genomes.[26,27] Recently, the proteobacteria *Burkholderia* has been identified as a promising candidate for the discovery of bioactive compounds, based on the genome-guided discovery of natural product biosynthetic gene clusters.[28] The *Burkholderia* genus consists of over forty Gram-negative species that occupy remarkably diverse ecological niches, through the production of secondary metabolites.[29] A group of seventeen closely related Gram-negative bacteria, so-called *Burkholderia cepacia* complex (Bcc), are known to demonstrate biopesticidal interactions against fungal diseases in plants,[30] biodegrade pollutants,[31,32] and have the ability to cause devastating lung infections in individuals with cystic fibrosis.[33] Screening a large collection of Bcc strains has identified a strong activity against *Acinetobacter baumannii*, which exhibits resistance to most antibiotics and is one of the most problematic pathogens in healthcare institutions globally. This antimicrobial activity has been identified from *Bukholderia ambifaria*, and the anti-Gram-negative activity was mapped to a biosynthetic gene cluster using transposon mutagenesis.[34] Purification and structure elucidation of the metabolic products identified them as the antibiotic enacyloxin IIa and *iso*-enacyloxin IIa.[34]

The proposed biosynthetic pathway for enacyloxin IIa is outlined in Fig. 1.3. It consists of a large modular PKS, NRPS domains, a pathway for a shikimate derived compound, and tailoring enzymes. The loading module in protein Bamb_5925 consists of a MT domain, a GNAT domain (malonyl transferase/decarboxylase from the Gcn5 histone acyltransferase family) and an ACP domain, which initiates polyketide assembly with a propionyl starter moiety. Modular extension continues in a largely sequential manner for ten rounds of chain extension from PKS proteins belonging to the *cis*-AT PKS phylogeny, despite many of these modules lacking an AT domain within their modules. Polyketide chain release is preceded by a PKS module within Bamb_5919, belonging to the *trans*-AT PKS phylogeny, which does not perform chain elongation. The natural product intermediate is proposed to be passed from the ACP of Bamb_5919, to the adjacent non-elongating ketosynthase (KS^0), that lacks the necessary catalytic histidine for chain elongation. The chain transfer continues from the KS^0 domain to a PCP domain within Bamb_5917, where specific protein-protein interactions are postulated to contribute to the directionality of this process. Chain release is accomplished by the condensation of the 3'-hydroxyl group of a shikimate derived dihydroxy-cyclohexane carboxylic acid (DHCCA), with the PCP-bound polyketide. The biosynthetic pathway for the DHCCA compound includes the proteins Bamb_5912 Bamb_5913, Bamb_5914, Bamb_5916 and Bamb_5918. The condensation reaction is catalysed by an NRPS-like C domain within Bamb_5915. The released intermediate then becomes the substrate for a series of tailoring enzymes (Bamb_5927, Bamb_5928, Bamb_5931, Bamb_5930 and Bamb_5932).

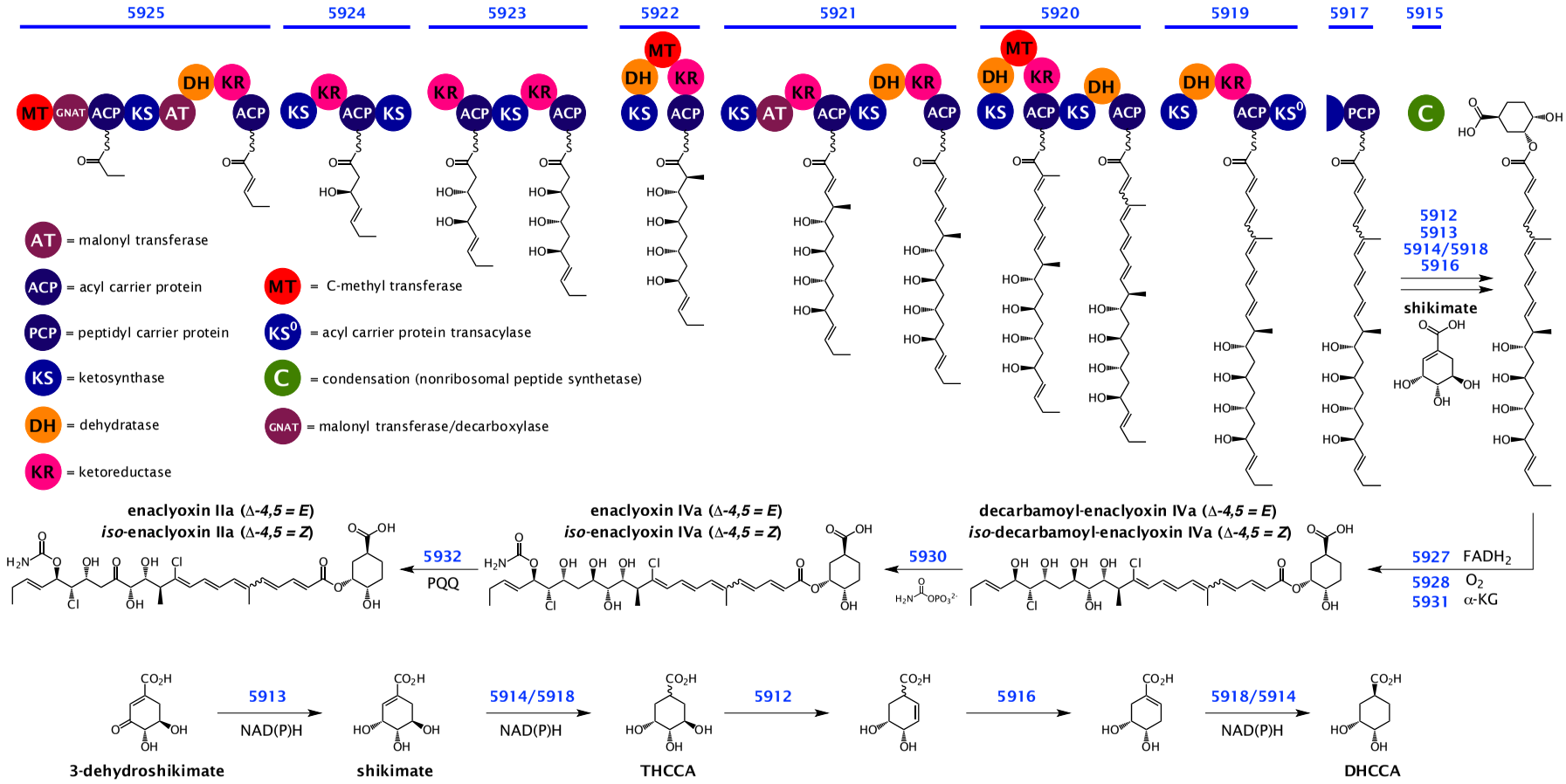

Figure 1.3: Proposed biosynthetic pathway for enacyloxin[34] (adapted and updated).[35]

1.7 Objectives

1.7.1 Objective I

The primary objective is to elucidate the biophysical basis for the unusual chain release mechanism from the enacyloxin biosynthetic pathway. This process involves two surplus enzymes, in terms of biochemistry, so it is proposed that a series of compatible interactions take place between ACP, KS^0, PCP and C protein domains. This could explain the necessity for this unusual domain architecture. Elucidating the molecular basis for product release from the enacyloxin pathway, will facilitate the application of this unusual enzymology to synthetic biology approaches, for novel antibiotic production. A major challenge for engineering biosynthetic pathways is developing a molecular-level understanding of domain-domain and protein-protein interactions to faciliate efficient catalysis. The description of interactions during enacyloxin chain release could describe rules of interactions that could facilitate bioengineering novel biosynthetic pathways, and describe unidentified features from characterised biosynthetic pathways.

1.7.2 Objective II

The secondary objective is to search for enacyloxin analogues, exploit any bioactive properties and to investigate the biosynthesis of enacyloxin analogues. A genomics-driven approach will be used to extensively search databases of sequenced organisms as experimental leads. Resulting organisms that contain components related to the enacyloxin biosynthetic gene cluster, will be acquired, cultured and their metabolic profile will be surveyed for the production of enacyloxin analogues. The structure, spectrum of activity, and biosynthetic pathway of any enacyloxin-analogues will be investigated. Any beneficial bioactivity of enacyloxin analogues could provide promising drug-leads, for chemical synthesis or mutasynthesis.

1.8 Synopsis

This body of work encompasses three result chapters on the theme of the biosynthesis of enacyloxins. Chapter 2 probes an unusual chain release process during the biosynthesis of enacyloxin, where specific protein-protein interactions are proposed to facilitate biosynthesis. Investigating protein interaction by NMR leads to the discovery of an interaction site at the C-terminus of a PCP domain. A crystal structure of the condensing enzyme revealed an additional N-terminal communication-mediating (COM) domain. Together, this provided a full account of the intermolecular interaction during chain release between PCP and C domains. Following the discovery of the COM interaction within the enacyloxin biosynthetic pathway, an HMM was developed to search for similar motifs within other biosynthetic pathways (chapter 3). Many biosynthetic pathways were identified using this HMM, and a hybrid COM interaction was investigated at a non-native COM boundary from watasemycin NRPS and enacyloxin PKS-NRPS. Other types of docking domains were also investigated using HMMs, which led to the identification of a novel KS-DH protein boundary feature.

The discovery of an enacyloxin analogue is described in chapter 4. A genome-mining approach revealed an enacyloxin-like biosynthetic gene cluster within *Vibrio rhizosphaerae*. The production of an enacyloxin analogue, referred to as vibroxin, was purified from extracts of solid cultures, and characterised. Some of the differences between the molecular structures of vibroxin and enacyloxin could be predicted from the differences in biosynthetic gene clusters. One significant difference between enacyloxin and vibroxin is a dimethylation of the starter unit during vibroxin biosynthesis, which was investigated by isotope feeding experiments. Rigorous inspection of the PKS proteins, revealed fragments of AT domains, suggesting intermediate stages between *cis*-acting AT domains and *trans*-acting AT domains. The spectrum of activity of vibroxin is tested against a set of pathogens, demonstrating its antibiotic potential.

Chapter 2

Enacyloxin PKS-NRPS chain release

2.1 Chain release process

During the biosynthesis of polyketides and non-ribosomal peptides, chain release is ordinarily accomplished by a TE domain. This domain is dedicated to the hydrolysis of the final thiol-bound intermediate, and often mediates macrocyclisation. In contrast, within the enacyloxin biosynthetic pathway, four enzymes are involved in the formation of an ester bond during chain release. This unusual system includes an ACP domain, a non-elongating KS domain (KS^0), a PCP domain and a stand-alone C domain. The KS^0 domain lacks a catalytic histidine residue involved in the mechanism for chain elongation. Thus, the KS^0 domain is proposed to act as an adapter between ACP and PCP domains, acting as a *trans*-acylase (TA), passing the polyketide intermediate from the ACP domain to the PCP domain. Subsequently, the catalytic C domain is then proposed to selectively attack the 3'-OH of dihydroxy-cyclohexane carboxylic acid (DHCCA), which then forms an ester with the PCP-bound polyketide (Fig. 2.1a). However, the C domain has also been shown to accept a number of other intermediates within the DHCCA biosynthetic pathway, including shikimate.[35] ($1'R$,$3'R$,$4'S$)-

DHCCA does not appear to be a substrate for the C domain, but (1′*S*,3′*R*,4′*S*)-DHCCA is accepted by the C domain.[35]. For the remainder of this thesis, the substrate of the C domain will be referred to as (1′*S*,3′*R*,4′*S*)-DHCCA, or shikimate, where stated. One alternative pathway describes the condensation of (1′*S*,3′*R*,4′*S*)-DHCCA with the ACP-bound polyketide, instead of the PCP-bound polyketide (Fig. 2.1b). However, previous work has shown that *in vitro* assays with acetyl-ACP, (1′*S*,3′*R*,4′*S*)-DHCCA and the C domain do not perform catalysis.[36] It is proposed that specific domain interactions play an important role in this unusual chain release process, requiring four enzymes to produce one ester bond. The biophysical basis of such specific interactions was investigated, using a range of biophysical techniques.

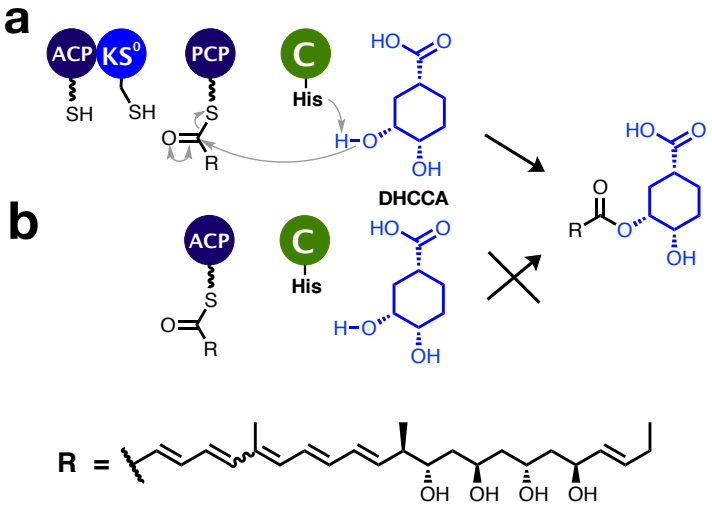

Figure 2.1: Stages preceding chain release in enacyloxin PKS-NRPS. a) The proposed chain release pathway, b) an equivalent biosynthetic pathway without KS^0 and PCP domains.

2.2 Protein NMR spectroscopy

2.2.1 Sample preparation

To uncover specific details of the protein-protein interactions the domains involved in the chain release process in enaclyoxin PKS: ACP (encoded by gene

Bamb_5919), KS^0 (encoded by gene Bamb_5919), PCP (encoded by gene Bamb_5917) and C (encoded by gene Bamb_5915) were reconstituted *in vitro* with N-terminal His6-tags to facilitate purification. For the reminder of Chapter 2 and for the sake of simplicity the designations ACP, PCP, C and KS0 will refer to the above mentioned constructs unless specified otherwise.

In order to facilitate NMR studies and whenever required, ^{15}N and ^{13}C were incorporated in proteins by using isotopically labeled precursors as the sole sources of carbon and nitrogen during recombinant overproduction of proteins in *E. coli*. Proteins were purified using nickel affinity column. The SDS-PAGE of the purified domains are presented in Fig. 2.2. Because KS^0 and C domains are very large, and thus challenging for solution NMR (75 and 60 kDa respectively), only smaller and amenable to solution NMR ACP and PCP domain where isotopically labelled and observed in ^{15}N and ^{13}C-filtered experiments. For protein-protein interaction experiments, carrier proteins were uniformly isotopically labelled with ^{15}N (*i.e.* [U-^{15}N]PCP and [U-^{15}N]ACP), and the larger domains, KS^0 and C, were not isotopically enriched.

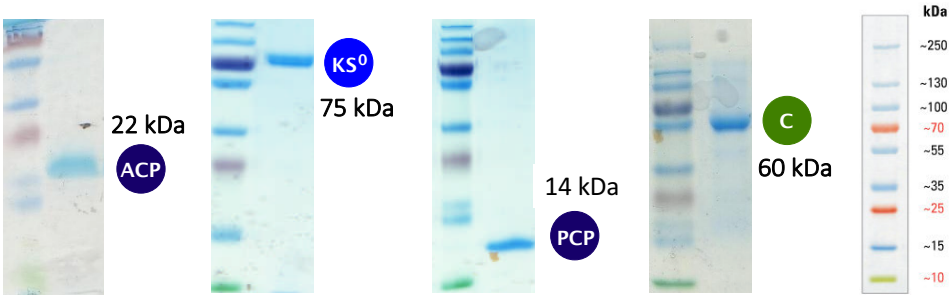

Figure 2.2: SDS-PAGE of purified domains from enacyloxin PKS.

2.2.2 Carrier protein interactions with partner domains

Protein-protein interactions, or the formation of protein-protein complexes, can be characterised by following changes in 2D ^{1}H-^{15}N HSQC spectra of uniformly ^{15}N-labelled carrier protein upon addition of a larger domain, e.g. natural abundance C or KS^0 domains (Fig. 2.3). In the case of fast chemical exchange the

changes of positions (or chemical shifts) of the sites involved in binding inform on the interaction. In the case of slow chemical exchange the interaction manifests as a decrease of the intensity of the peak for free site and increase of intensity of the peak for the bound site as a function of interacting partner concentration. As the size of the complex increases the decrease in the overall rotational diffusion correlation time leads to improvement of transverse relaxation and associated line broadening. For very large complexes, as it is in our case, the signals of the carrier protein in the complex may become broadened beyond detection.

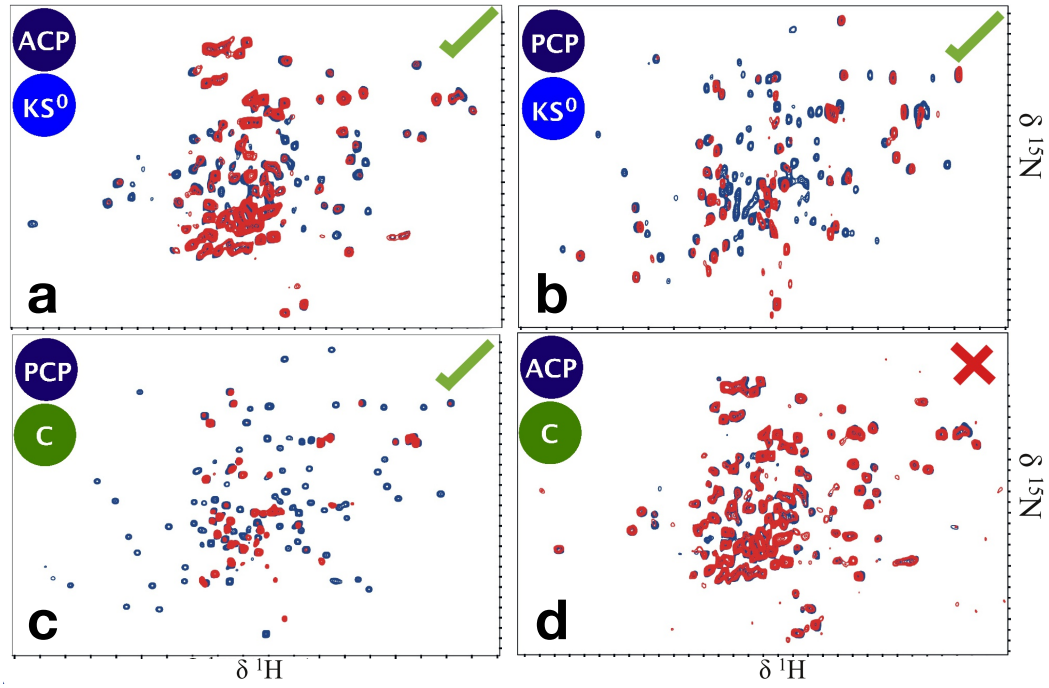

Figure 2.3: ^{1}H-^{15}N HSQC spectra of carrier proteins in isolation (blue), and with the addition of a partner domain at a 1:1 ratio (red). a) ACP-KS^{0}, b) PCP-KS^{0}, c) PCP-C, d) ACP-C.

Within the HSQC spectra of protein complexes ACP-KS^{0}, PCP-KS^{0} and PCP-C (Fig 2.3a-c), there are significant changes in the HSQC spectra of carrier proteins following the addition of a larger partner domain. This indicates an interaction between them. On the contrary, the ACP-C complex HSQC spectrum shows no significant changes when compared with those observed for the other domain-domain experiments (Fig 2.3d). This difference in behaviour of PCP and ACP suggests that the ACP lacks a compatible interaction site for the C domain.

The residue assignment of PCP was then undertaken, to uncover the specific details of the interaction site on PCP.

2.2.3 Residue assignment of PCP

The buffer conditions for NMR samples were screened for stability of PCP, including changes to pH (Appendix Fig. 7.1, 7.2) and temperature (Appendix Fig. 7.3, 7.4). A suite of three-dimensional NMR experiments could then be performed (Section 5.4.2), to sequentially deduce the assignment of [U-^{13}C,^{15}N]PCP. The backbone assignment of PCP was achieved for 96% of backbone amides (Fig. 2.4, Appendix Table 7.1), excluding the N-terminal purification tag. The dispersion of the hydrogen chemical shifts are characteristic of a well-folded protein, but the random coil deviations, proximal to the C-terminus, are indicative of an unstructured peptide, with a high degree of disorder (Fig. 2.5).

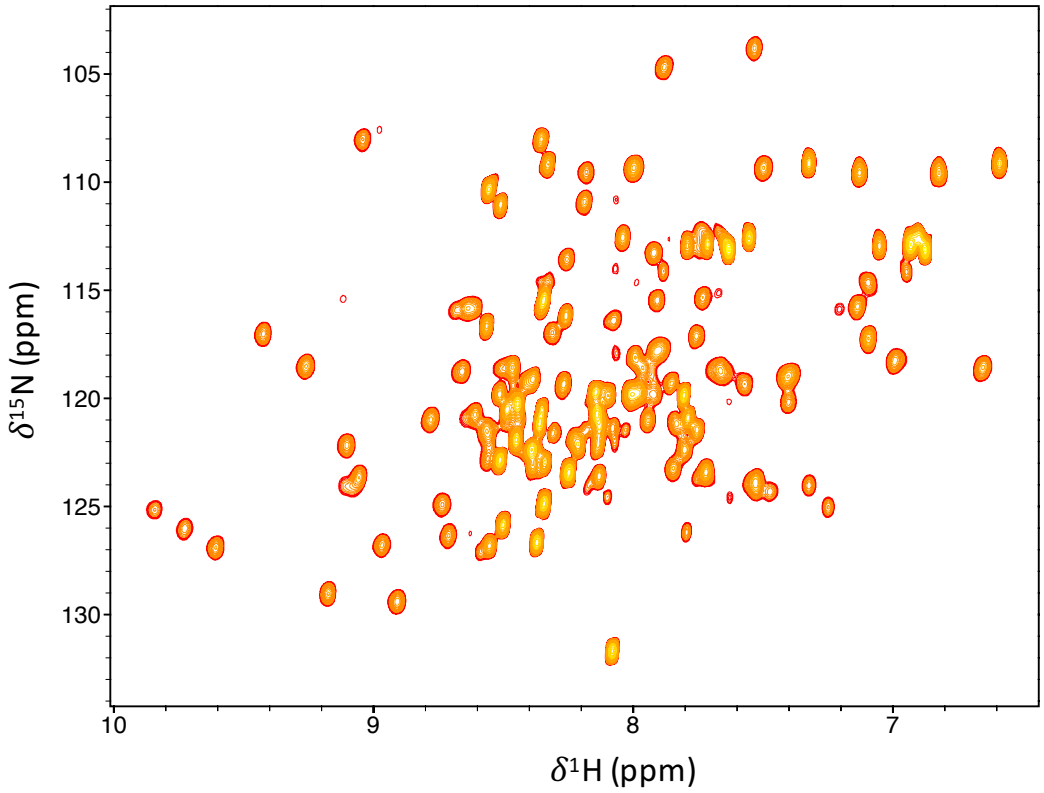

Figure 2.4: HSQC of PCP where the amide assignment is listed in Appendix Table 7.1.

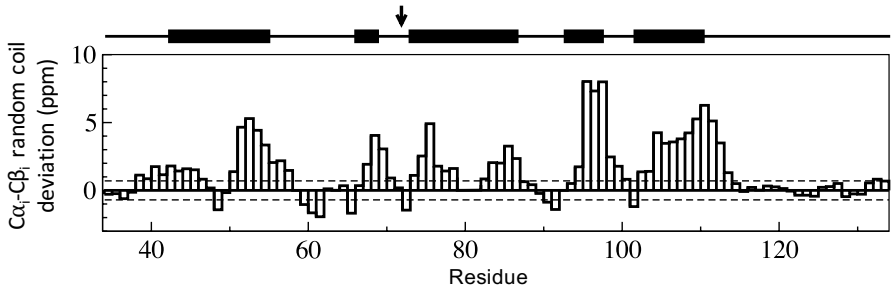

Figure 2.5: Random coil chemical shift deviations for PCP based on the difference in Cα_i and Cβ_i chemical shift values for each residue, i. Predicted secondary structure is represented above the bar chart, where bold lines indicate helices and the active site serine residue is identified with an arrow.

2.2.4 C-terminal interaction site

The aforementioned HSQC spectra, that monitor interactions between PCP and C domains, display peak broadening, where the C-terminal region displays the most intense signals, even with a relatively high concentration of the C domain. The high intensities of these correlations correspond to a region of the protein that does not conform to the core helical bundle, observed in other carrier protein structures. In fact, the chemical shifts of this C-terminal extension are consistent with an unstructured, or disordered, peptide. This lack of secondary structure is congruous with the observed high signal intensities. The decay in each amide correlation signal, during interaction of the PCP and C domain, can be observed across the PCP protein to identify interaction sites during complex formation. As well as the core PCP domain, significant changes were observed at the disordered C-terminus of PCP (Fig. 2.6a-e) from about residue 125 to residue 134, with increased amounts of C domain. This suggests that the C-terminal region is important for the complex formation of the PCP and C domains, and consequentially facilitating ester formation, during chain release. This C-terminal domain, downstream of PCP, shall be referred to as a C-terminal communication-mediating domain (COM_{2C}).

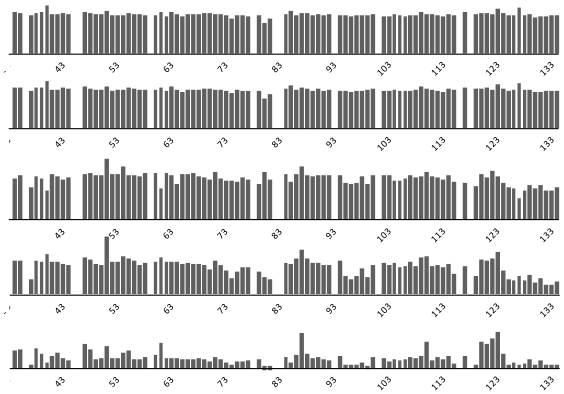

Figure 2.6: Normalised signal intensities for PCP-COM_{2C} amide HSQC correlations upon titration with natural abundance C domain, for the following C:PCP ratios: a) 1:16, b) 1:8, c) 1:4, d) 1:2, e) 1:1.

2.3 *In vitro* bioassays

2.3.1 Construction of mutants

The role of the COM_{2C} domain can be investigated using assays that depend upon the COM interaction. To evaluate the biochemical importance of the COM domains, nine amino acids were removed from the C-terminus of PCP-COM_{2C} to form PCP. This was achieved by site-directed mutagenesis, where a premature stop codon was substituted on the PCP-COM_{2C} expression plasmid. The truncated protein, PCP, was recombinantly overproduced and purified (Fig. 2.7a). Mass spectrometry of PCP (Fig. 2.7bi) and PCP-COM_{2C} (Fig. 2.7bii) provide ions consistent with predicted molecular weights, 14466.3 Da and 13403.1 Da respectively.

2.3.2 Condensation assay

A condensation assay was developed using: acetyl-PCPs (Bamb_5917), the C domain (Bamb_5915), and shikimate, as a mimic of DHCCA (Fig.2.8a). This setup

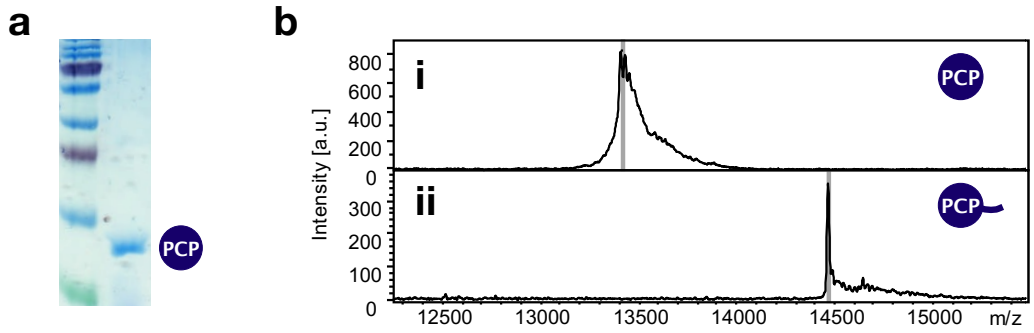

Figure 2.7: Truncated mutant PCP. a) SDS-PAGE of PCP, b) MALDI-ToF spectra of i) PCP truncated mutant, ii) PCP-COM_{2C}.

could be used to verify the role of the COM_2 boundary during biosynthesis, in particular the effect of COM_{2C}. A mimic of the wild-type system, acetyl-PCP-COM_{2C} with the C domain and shikimate, yielded a product consistent with acetylated shikimate, monitored by LC-MS. The extracted ion chromatogram corresponding to the sodiated adduct of acetylated shikimate displays three species with different retention times (Fig.2.8b). This could be due to acetylation of each of the three hydroxyls on the cyclohexene during condensation, or acetyl rearrangement thereafter. The truncated mutant PCP was used to replace the PCP-COM_{2C} protein in the assay, to investigate the biochemical significance of the COM_{2C} domain. No product formation was observed for assays containing the COM_{2C} truncated PCP (Fig.2.8bii). The interpretation of this result is that the breakdown of the COM_{2C}-dependent intermolecular interaction prevents efficient complex formation, a prerequisite for catalysis in this case. The assay with the full length PCP-COM_{2C} was then repeated with the addition of a tenfold excess of COM_{2C} peptide (11 amino acid length chemically synthesised by Insight Biotechnologies). From this assay, no product was detected (Fig.2.8biii). Competitive inhibition of the protein-protein interaction is proposed to account for significantly decreasing catalysis upon the addition of supplementary COM_{2C} peptides. Together these results support the biochemical significance of the COM_{2C} domain during chain release of enacyloxin.

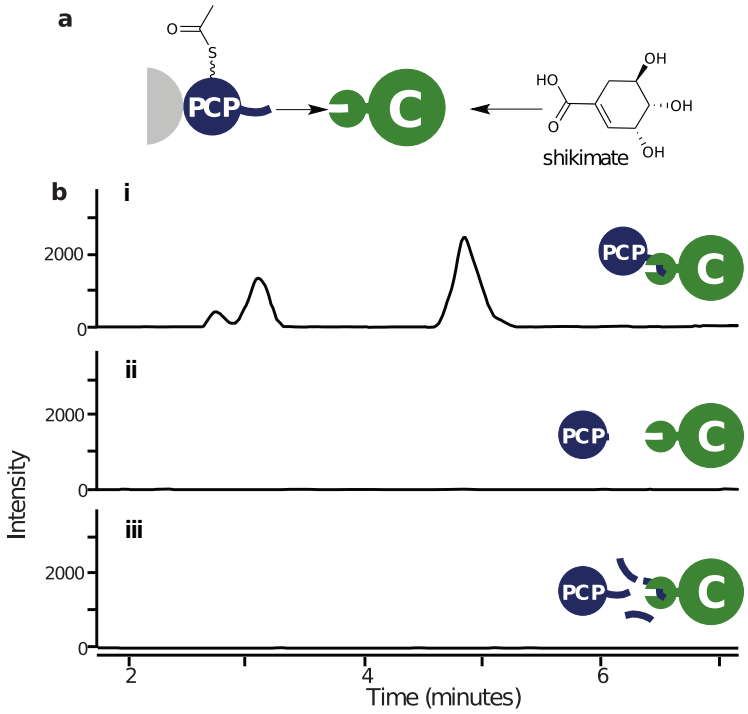

Figure 2.8: a) Schematic representation of the condensation assay. b) E.I.C corresponding to the sodiated adduct of acetylated shikimate for the assay with the C domain, shikimate and either, i) acetyl-PCP-COM_{2C}, ii) truncated acetyl-PCP, iii) acetyl-PCP-COM_{2C} with a tenfold excess of COM_{2C} peptide.

2.4 Crystal structure of the condensing enzyme

In order to determine the possible binding site of COM_{2C}, the structure of the C domain from enacyloxin PKS-NRPS was solved. This was part of a collaboration between the University of Warwick and UC Irvine, which included Timothy Valentic, Sheryl Tsai, Paulina Sydor, Dean Rea and Vilmos Fülöp. This structure deviates from other NRPS-like C domain structures, by a loop above the active site, and the addition of an N-terminal domain.

2.4.1 N-terminal appendage

The additional N-terminal domain consists of an $\alpha\beta\beta\alpha\alpha$ structure, and appears recurrently within NRPS proteins, typically identified upstream of condensation or heterocyclisation domains using homology searches and motif finders. An isolated structure of a homologous domain was solved from tubulysin NRPS[23]

(Fig. 2.9a), and its function was proposed to perform an intermolecular interaction across a protein boundary. Subsequently, this type of domain will be herein referred to as a 'type 2' N-terminal communication mediating (COM_{2N}) domain. This will distinguish it from 'type 1' COM domains, regularly identified between condensation domain and epimerisation domain boundaries (see Chapter 3 for further details).[37]

The crystal structure of COM_{2N}-C from enacyloxin PKS-NRPS, provides the first structural account of the relative positioning of a COM_{2N} domain, to its catalytic neighbouring domain (Fig. 2.9b). A short linker allows the COM_{2N} domain to situate near the C-terminal subdomain of the condensing enzyme, positioned adjacently to the entrance of the downstream substrate. Amino acids Thr2 and Tyr51 from the COM_{2N} domain engage in polar contacts with Glu380 and Ser383 from the C domain respectively, which stabilise the localisation of COM_{2N}.

2.4.2 Loop above the active site

The structure of the C domain contains some variations from other NRPS C domain structures, namely between Ser418 and Ser453. Within this region, a loop extends above the C-terminal subdomain, forming little secondary structure. In other structures this region typically forms a β-sheet with the N-terminal subdomain with additional helical components, displaying an overall more structured fold. These differences could arise from the isolated nature of this particular condensation domain, as its placement differs from the archetypal location within an NRPS module. Whilst the upstream PCP delivers a polyketide to the accepting face of the condensing enzyme, access to the donating face is proposed only to facilitate a relatively small substrate (*e.g.* DHCCA), not a typical PCP-bound amino acid.

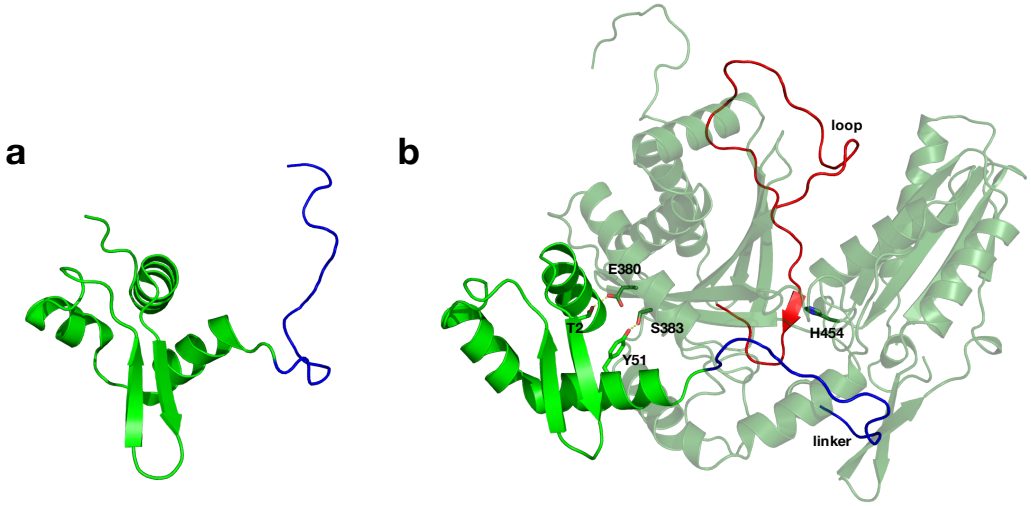

Figure 2.9: a) COM_{2N} structure from tubulysin NRPS.[23] b) COM_{2N}-C didomain structure from enacyloxin PKS-NRPS. Linker between COM_{2N} domain and C domain is shown in blue. The atypical loop within the C domain is shown in red.

2.5 Protein-protein complex

A crystal structure of a PCP-E didomain[38] (Fig. 2.10a) from gramicidin S NRPS provides a reference for modelling the PCP-C complex from enacyloxin PKS-NRPS, as E domains and C domains possess the same overall fold. The PCP domain from enacyloxin PKS-NRPS was overlayed on the PCP-E structure, and the COM_{2N}-C domains were aligned to the relative position of the E domain (Fig. 2.10b).

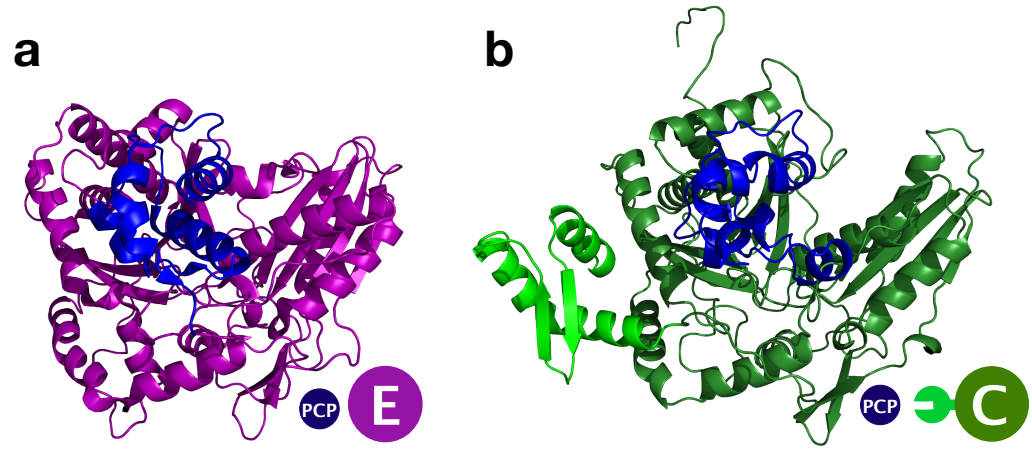

Figure 2.10: a) PCP-E didomain structure from gramacidin NRPS.[38] b) Alignment of the COM_{2N}-C domains and PCP from enacyloxin PKS-NRPS to the relative positions of the PCP-E complex structure.

Models of the interaction between COM_{2N} and COM_{2C} domains have been proposed from tubulysin NRPS[23] (Fig. 2.11ai) and myxothiazol NRPS[39] (Fig. 2.11aii). Using the structure of COM_{2N}, and the partner interaction site COM_{2C}, a protein-peptide molecular dynamics simulation was performed to objectively investigate the COM_{2C}-COM_{2N} domain interaction during the biosynthesis of enacyloxin. Small amounts of helical propensity were observed for the COM_{2C} peptide during simulations, but it remained largely devoid of any secondary structure generally. Four clusters of solutions of the COM_{2N} and COM_{2C} complex are represented in Fig. 2.11b. Each of the protein-peptide complex models represent clusters of hundreds of states, where the COM_{2C} peptide samples proximal regions of the COM_{2N} surface. These solutions may represent a 'fuzzy' complex, whereby degenerate multiplicity exists, as is commonly observed for interaction sites involving disordered peptides.[40,41] It is unlikely that the proposed interaction sites on the COM_{2N} can be simultaneously occupied by the COM_{2C} domain, due to their relatively large expanse. The COM domain complex solutions provide an insight into the role of the hydrophobic amino acids on COM_{2C}. There are two pockets, adjacent to the β-hairpin, that locate the hydrophobic residues (namely valine and isoleucine) on COM_{2C}.

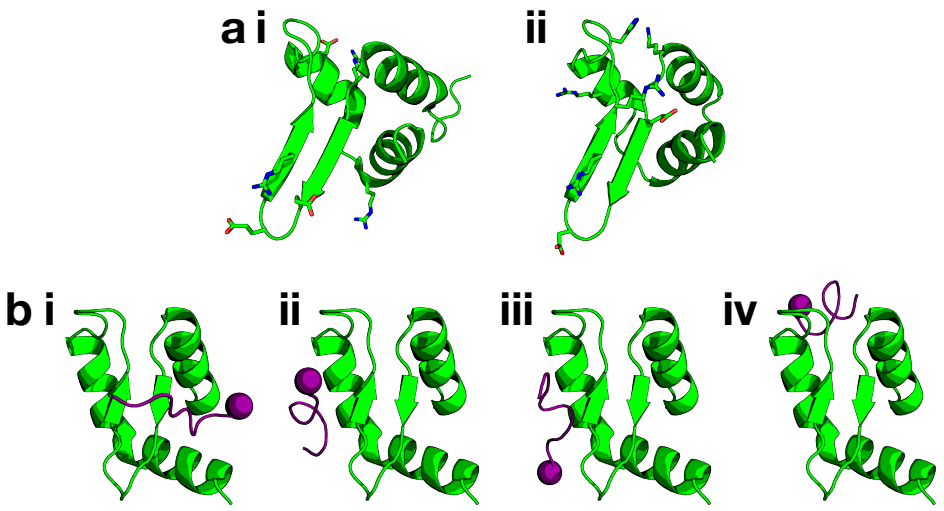

Figure 2.11: a) COM_{2N} domain interaction sites (sticks) proposed from i) a COM_{2N} from tubulysin NRPS,[23] ii) a COM_{2N} domain homology model from myxothiazol NRPS.[39] bi-iv) Solutions from simulations of the COM_{2C} domain (purple, C-terminus represented as sphere) with the COM_{2N} domain from enacyloxin PKS-NRPS.

Coupling the PCP-C model, and the COM_{2N}-COM_{2C} models, provides a representation of the entire complex (Fig. 2.12). It describes intermolecular communication, and the interactions of adjacent enzymes responsible for catalytic activity during enacyloxin chain release. A model solution for the COM domain-complex was linked to the C-terminus of the core PCP domain by building in a loop, which was subsequently subjected to energy minimisation. The PCP-COM_{2C} and COM_{2N}-C complex model describes the proposed interactions, and a schematic of the states it must accommodate during biosynthesis. After translation, the intermolecular interactions that are required between the PCP and C domain, are proposed to be facilitated by a pair of compatible COM domains. This ensures a particular biosynthetic pathway is followed, and acts to increase the effective concentration of the PCP domain with respect to the C domain. Following COM domain interactions, the PCP is proposed to deliver a polyketide intermediate to the C domain, for condensation with a shikimate derived compound. The *holo*-PCP domain then needs to leave the C domain so that the

terminal thiol of *holo*-PCP can be made acessible, to load another molecule of the polyketide from the KS^0 domain, before redelivery to the C domain for ester formation with DHCCA.

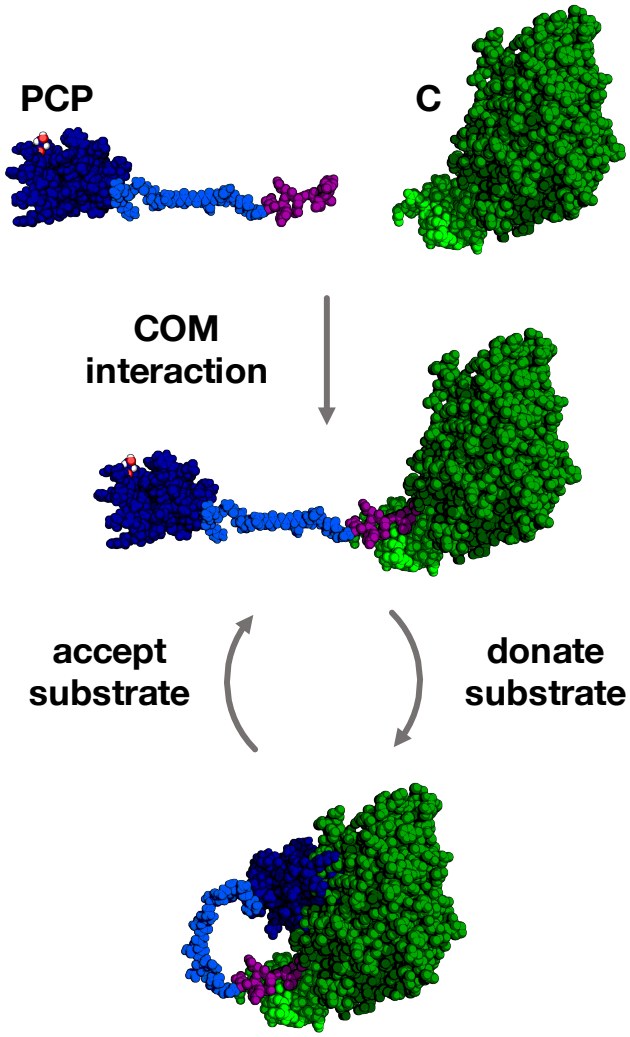

Figure 2.12: Representation of the domain oragnisation of PCP-COM_{2C} and COM_{2N}-C from enacyloxin PKS-NRPS. COM domains are propsed to direct intermolecular interactions, where the PCP domain accepts polyketides from the previous module to donate to the C domain in a recursive manner.

2.6 Conclusion and future work

Protein interactions have been shown to play an important role in the unusual chain termination process during enacyloxin biosynthesis. Interactions between

domains have been observed by protein NMR spectroscopy, and are consistent with the proposed biosynthetic pathway. Biophysical investigation of the chain release process has uncovered a pair of COM domains, which form specific intermolecular interactions across the PCP and C protein boundaries. The residue assignment of PCP, and PCP-C mixing experiments, led to the identification of a C-terminal COM domain, exhibiting high levels of disorder. This COM domain has been shown to be necessary for efficient catalysis, performing a fundamental intermolecular interaction across a protein-protein boundary.

The structure of the C domain revealed an additional N-terminal COM domain, and its position relative to the catalytic C domain. The nature of the COM domain interaction was simulated, which provided clusters of statistical solutions. These states were subsequently evaluated based upon the NMR interaction data with the C-terminal binding partner. However, an experimentally determined structure of the protein complex, or an experimentally verified description of the interactions sites on the N-terminal COM domain interface, would provide additional insight to this particular intermolecular interaction.

The COM domain interaction has led to an updated description of chain release within enacyloxin biosynthesis, and is proposed to drive complex formation to facilitate catalysis. This COM domain interaction site allows the quaternary structure to be maintained, whilst permitting accessibility for the PCP domain to locate the active sites of both the C domain, and the KS^0 domain. Given the significance of the COM domain interaction within enacyloxin biosynthesis, further experiments could investigate the exploitation of this interaction to rationally engineer intermolecular interactions to develop hybrid biosynthetic pathways. Furthermore, to test the promiscuity of the enzymes across COM domain boundaries, creating a chimeric protein of ACP-COM_{2C} (instead of PCP-COM_{2C}) would present valuable insight to the tolerance of ACP and C domains, and ventures for engineering COM domain intermolecular interactions to biosynthesise novel chemical entities. In order to discover the prevalence of this form of COM

domain interaction, new tools could be developed for their annotation within biosynthetic pathways.

Chapter 3

Intermolecular docking domains and communication-mediating domains

3.1 Identification of docking and COM domains

Given the large number of known PKS and NRPS proteins, docking domains and COM domains are scarcely identified within them. The development of new tools to identify these terminal domains will provide a more detailed understanding of intermolecular interactions within PKSs and NRPSs.

3.1.1 Profile hidden Markov models for dd_1 and COM_1

Docking domains and COM domains can be identified using profile hidden Markov models (pHMMs), which consist of a statistical description of a protein family. Currently, antiSMASH[42] utilises two pHMMs that describe N-terminal and C-terminal docking domains (denoted dd_{1N} and dd_{1C} respectively) and two pHMMs for COM domains (COM_{1N} and COM_{1C}). The pHMM for COM_{1N} contains a region that can be considered part of a condensation or cyclisation domain from a structural perspective (Fig. 3.1). This region of the pHMM leads to a large num-

ber of results that could be considered potential false positive annotations. As a consequence, it is commonplace to identify COM_{1N} domains upstream of condensation domains that do not lie at a terminus. The pHMM for the corresponding binding partner, COM_{1C}, contains a small region that can be attributed to an epimerisation (E) domain. Despite this, COM_{1C} domains can also be identified following PCP domains, at PCP and condensation domain protein boundaries (*e.g.* VpsA-VpsB from vancomycin NRPS).

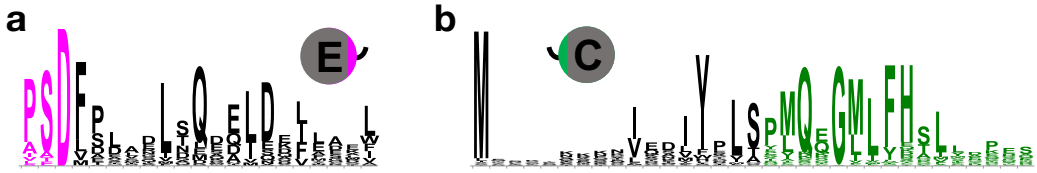

Figure 3.1: HMM logos for a) COM_{1C} and b) COM_{1N} HMMs from antiSMASH.

Analogous to COM domains, docking domains have been identified within a number of PKSs. These have also been identified using pHMMs for dd_{1N} and dd_{1C}, corresponding to the N-terminal and C-terminal interaction sites, respectively (Fig. 3.2). The pHMM that includes the dd_{1C} domain, also includes a region describing an additional dimerisation domain.[19] This region has been structurally characterised from eryAI, and is regularly unidentified within PKS proteins. This is a limitation of this relatively long pHMM. In contrast, the N-terminal binding partner, dd_{1N}, more accurately represents the profile of this domain, and is consistent with a structurally characterised dd_{1N} domain.

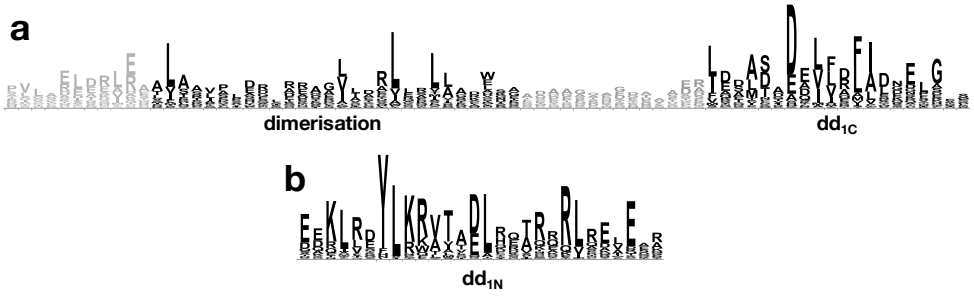

Figure 3.2: HMM logo for type 1 docking domains used by antiSMASH, a) C-terminal PKS dd, b) N-terminal PKS dd.

3.1.2 Novel HMMs for docking domains

A second class of docking domains (dd2) are based on protein termini identified at ACP and KS boundaries within curacin PKS (CurG-CurH and CurK-CurL).[18] These dd2 domains were also reported within 22 other biosynthetic pathways at ACP and KS protein boundaries.[18] This set of N-terminal and C-terminal docking domains (dd_{2N} and dd_{2C} respectively) was used to build two HMMs (dd_{2N}, dd_{2C}) to facilitate the identification of these domains.

The dd_{2N} and dd_{2C} HMMs were used to search the MIBiG database[43] for biosynthetic pathways that contain dd2 domains. This revealed a large number of dd2-containing biosynthetic pathways, which were previously unidentified. The results include biosynthetic pathways for: anabaenopeptin, ambruticin, burkholdac, chondrochloren, crocacin, ebelactone, enacyloxin, gephyronic acid, gulmirecin, leupyrrin, microsclerodermins, nannocystin, nostophycin, pellasoren, polyoxypeptin, puwainaphycin and spiruchostatin. Each of these examples exhibit class 2 docking domains, which are proposed to facilitate biosynthesis across ACP and KS boundaries. One exception is nostophycin PKS, where dd2 domains are not located at protein boundaries of NpnA. Within this protein, dd_{2N} and dd_{2C} domains are found in between the ACP from module 1 and the KS from module 2 (Fig. 3.4a). It is clear that docking domains between modules 1 and 2 are not required as they are part of the same protein. These modules may have originated from separate genes where the dd2 domains would have facilitated biosynthesis across a protein boundary, but mutations in the stop codon of the upstream gene could have led to a genetic merger, if the downstream gene was situated in the same reading frame.

3.1.3 Identifying unknown types of docking domain

Whilst HMMs for docking domains have revealed many biosynthetic pathways utilising these terminal features, there are many PKS and NRPS proteins where no docking domains, or COM domains, are identified. In many cases, unanno-

tated terminal gaps provide motivation to investigate patterns in these terminal regions, which could uncover novel types of docking domains. To address this, a database of PKS and NRPS amino acid sequences was curated from characterised biosynthetic pathways, and the termini were extracted to provide a list of N-terminal and C-terminal amino acid sequences for inspection. These sequences were analysed for novel motifs (recurring, fixed-length patterns) using MEME.[44] The resulting motifs for N-terminal PKS sequences are listed in Fig. 3.3.

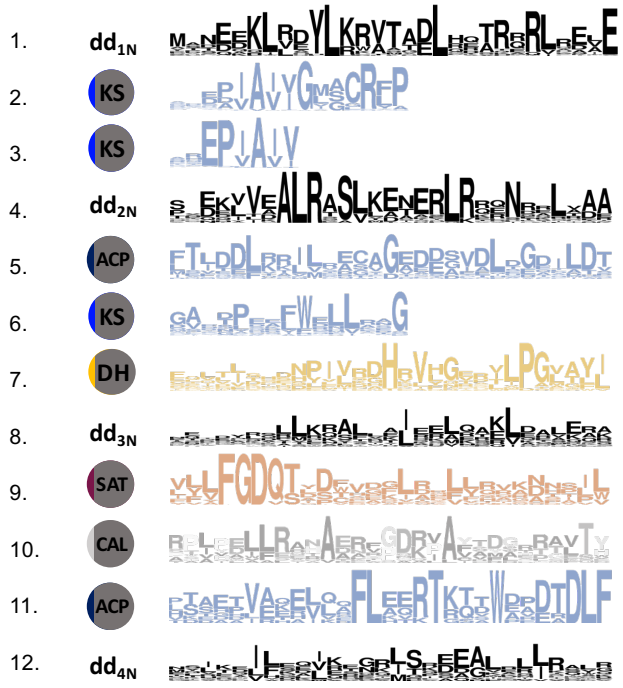

Figure 3.3: Motifs from N-terminal PKS sequences identified using MEME, listed by statistical significance. Each motif is labelled corresponding to its catalytic domain, or type of docking domain.

The N-terminal motifs could be attributed to docking domains as well as enzymatic domains that are situated directly at the N-terminus. KS and ACP domains from type II PKSs were identified by MEME, but did not contain terminal docking domains, as well as KS^Q domains involved in loading modules. Other N-terminal domains include starter-unit:ACP transacylase (SAT) domains from fungal biosynthetic pathways, and coenzyme A ligase (CAL) domains, conferring different forms of PKS loading strategies. The most significant motif was consistent with the well-studied type 1 docking domain, typically observed between

ACP and KS domains at protein boundaries.

The fourth most significant motif was used to develop a HMM to describe a docking domain with some similarity to type 2 docking domains (coined $dd_{2N-\alpha}$ and $dd_{2C-\beta}$). During instances where multiple HMMs identify a docking domain within the same amino acid sequence, the highest scoring HMM is chosen. The $dd_{2N-\alpha}$ and $dd_{2C-\beta}$ HMMs identified docking domains from 66 biosynthetic pathways (Appendix 7.3.1) at protein boundaries between ACP domains and KS domains. The eighth most significant motif described another pair of putative docking domains, which were converted into paired HMMs (dd_{3N} and dd_{3C}). These HMMs were used to identify 22 biosynthetic pathways (Appendix 7.3.2) from the MIBiG database, which putatively utilise dd3 domains. Most examples represent boundaries between at ACP and KS domains, but dd3 domains were also identified at PCP-KS and ACP-C protein boundaries. An example of an ACP-C boundary is BaeM-BaeN, from bacillaene PKS-NRPS, where a dd_{3N} domain was identified at the N-terminus of BaeN (or PksN), illustrated in Fig. 3.4b. Within the gobichelin biosynthetic pathway, a dd_{3C} domain was identified adjacent to a PCP at a PCP-C protein boundary, but no corresponding N-terminal binding partner was identified next to the condensation domain.

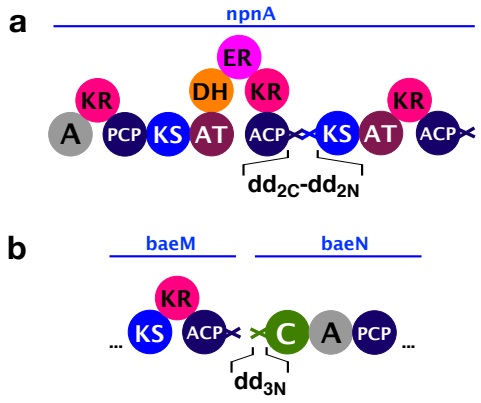

Figure 3.4: Examples of docking domains in unusual locations, a) nostophycin PKS with dd2 domains in *cis* b) bacillaene PKS-NRPS showing a PKS-like dd_{3N} domain adjacent to a condensation domain.

The twelfth motif was identified both at the N- and C-terminus of biosynthetic genes. The corresponding HMM (dd4) was used to identify this type of docking

domain from 18 biosynthetic pathways. The predicted structure of dd4 domains is a helix-turn-helix. When attempting to model the structure of dd4 domains based on the HMM consensus sequence, a structure of a N-terminal dd4 domain was found with a ketoreductase from macrolactin biosynthesis (this dd4 domain was also identified with the dd4 HMM).[22] Another structure of dd4 domains was later characterised from virginiamycin PKS, where a chimeric protein revealed a chimeric dd4-dd4 interaction over a protein boundary.[21] Apart from terminal d4 domains, three tandem d4 domains were identified within protein CurF from curacin PKS-NRPS (Fig. 3.5). This stage of biosynthesis has an unusual domain architecture including three tandem ACPs and four *trans*-acting enzymes. The N-terminus of CurF does not seem to accommodate a N-terminal docking domain, which may suggest that the d4-d4-d4 domains could facilitate biosynthesis across the protein boundary between CurA and CurF, and play a role in the recruitment of *trans*-acting enzymes.

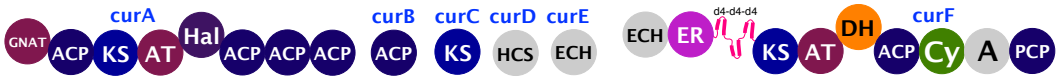

Figure 3.5: Domain organisation of curacin PKS-NRPS displaying tandem d4 domains located between the ER and KS domains.

3.1.4 (2*E*,4*Z*)-configuring dehydratase HMM

There were a number of N-terminal DH domains within PKS proteins, reflected in the seventh motif from MEME. These examples were unusual as the DH domains were identified directly at the N-terminus of proteins, without any vacancy for an additional N-terminal docking domain. Reviewing the N-terminal DH domains from characterised biosynthetic pathways revealed some common features. Each DH domain was within a module where the ACP-bound intermediate (or an adjacent ACP-bound intermediate) possessed (2*E*,4*Z*)-configured double bonds, and there was no DH domain within the previous module. The C-terminal region of the upstream protein was a *trans*-AT KS domain, but no *trans*-AT docking domains were identified. In fact, the region C-terminal to the KS did not contain

a motif associated with the C-terminus of a KS domain which could position the downstream module away from the upstream module. This may contribute to the DH domain acting on the ACP-bound intermediate from the upstream module, or equivalently, dehydration at the δ-OH after chain elongation. In each biosynthetic pathway, the amino acid sequence to the C-terminus of the KS domain possessed a relatively low pI.

An HMM was built to describe the N-terminal DH domains that do not possess a typical docking domain, in the form of an additional terminal appendage. The intention was to use this model to identify a class of intermolecular interactions between a C-terminal peptide and N-terminal DH domains. However, these domains were typically identified within modular PKS modules from *trans*-AT and *cis*-AT PKSs, located away from termini (Fig. 3.6). Examples of these DH domains can be found in the biosynthetic pathways for spirangien, thuggacin and chivosazole. For chivosazole PKS, the corresponding DH domains that are not located at the N-terminus, may arise as a consequence of a genetic merger resulting in two proteins becoming a single protein (as proposed within the nostophycin PKS). For example, in a pathway identified with the N-terminal DH HMM, the ACP-bound intermediate, within the corresponding module, contains a (2*E*,4*Z*)-configured diene. Within spirangien and thuggacin PKS, the modular architecture is archetypal of *cis*-AT modular PKSs (KS-AT-KR-ACP-KS-AT-DH-KR-ACP). The predicted extension of the ACP-bound intermediate arising from these two modules would possess a 4-OH, where the stereochemistry is defined by KR stereospecificity, and a 2*E*-configured double bond from the module containing a KR domain and a DH domain. The molecular structures of spirangien and thuggacin suggest that the modules containing a DH, identified by the aforementioned HMM, possess an intermediate that contains (2*E*,4*Z*)-configured dienes. This implies that the HMM that was initially developed to describe N-terminal DH domains, actually encompasses DH domains that could be involved in the stereospecific configuration of dienes. This could be utilised as a tool to assist

with the annotation of biosynthetic gene clusters, to predict the configuration of dienes for uncharacterised polyketides and polyketide-hybrids.

This HMM was also used to predict N-terminal (2*E*,4*Z*)-configuring DH domains from biosynthetic pathways for bacillaene, basiliskamide, bongkrekic acid, chivosazole, difficidin, elansolid, etnangien, kirromycin, macrolactin, sorangicin and tartrolon (Fig. 3.6). Each of these biosynthetic pathways has an unannotated C-terminal region, which is a putative interaction site on the upstream protein, at a *trans*-AT KS and a putative (2*E*,4*Z*)-configuring DH boundary. It is noted that an additional histidine amino acid (predicted to situate within the active site) is conserved within the proposed (2*E*,4*Z*)-configuring DH domains from *trans*-AT PKSs, that is not present in other DH domains, which most frequently possess a tyrosine amino acid at the corresponding position. This base may be involved in the diene synthase mechanism, in particular describing the additional dehydration event compared with other DH domains.

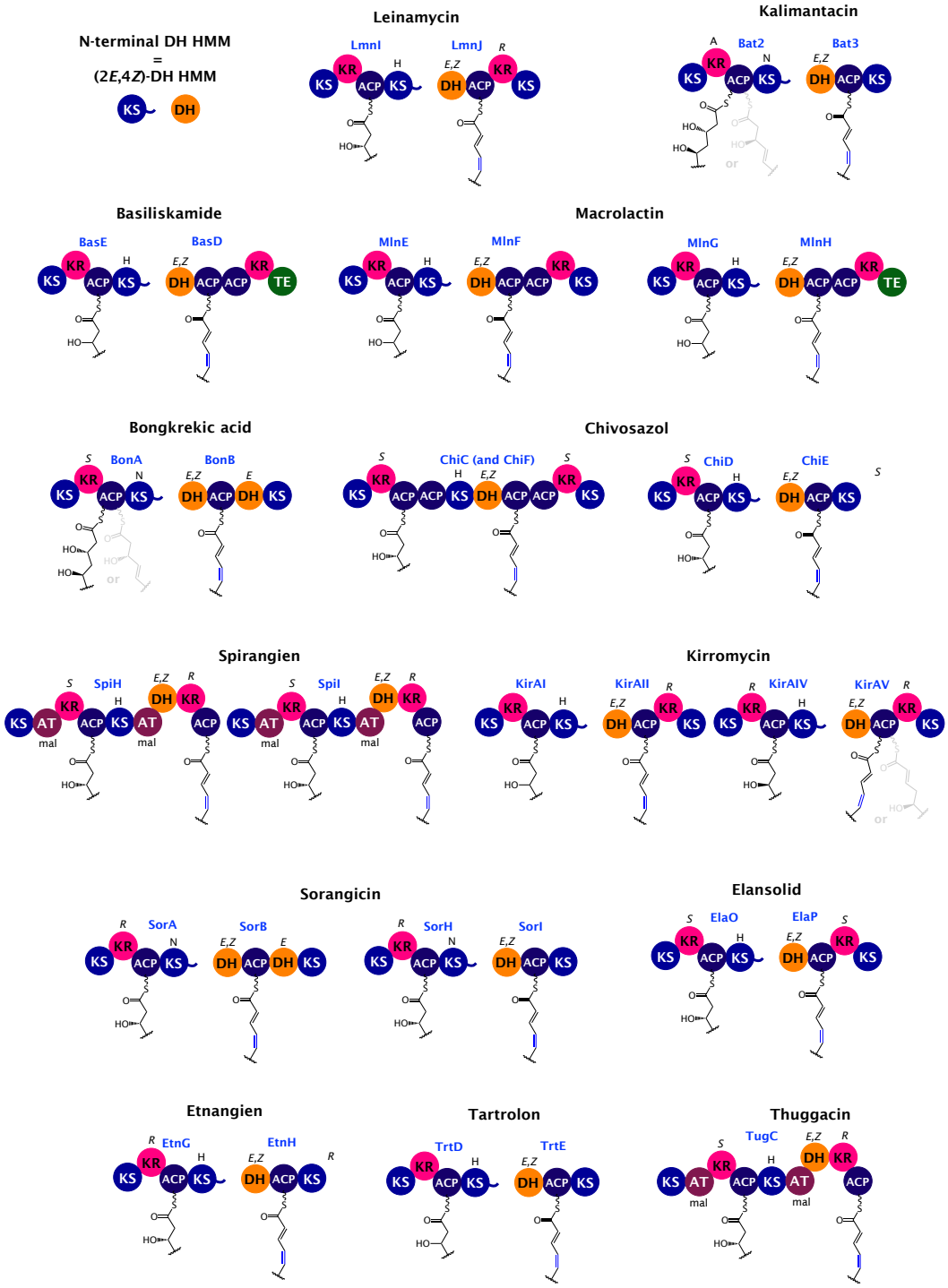

Figure 3.6: (2*E*,4*Z*)-configuring DH domains from characterised biosynthetic pathways, identified using the (2*E*,4*Z*)-DH HMM.

3.1.5 Identification of Type 2 COM domains

Two structural representations of a type 2 COM domain have been characterised, TubC from tubulysin NRPS, and Bamb_5915 from enacyloxin PKS-NRPS. Each of these examples describe intermolecular communication across a PCP-condensation domain boundary. Despite the relatively large size (90 amino acids) of a COM_{2N} domain, they cannot be identified using current bioinformatic methods based on domain conservation (e.g. conserved domain database, Pfam,[45] SMART,[46] COG,[47] TIGRFAM,[48] and HMMER.[49]) To address this, an HMM was developed using HMMER, which could allow the identification of previously unidentified COM_{2N} domains (Fig. 3.7b). The atlas PKS/NRPS database[50] was used as a test set of PKS and NRPS proteins, and 218 COM_{2N} domains were identified. These domains were identified within a range of genera (Appendix Fig. 7.5), and most frequently from the phyla of proteobacteria, cyanobacteria and actinobacteria. The interaction partner of the COM_{2N} domain is proposed to situate at the C-terminus of the previous stage of any biosynthetic pathway. To inspect these sequences, biosynthetic colinearity is assumed, and the upstream genes were used to identify the encoded C-terminal sequence. The C-terminal region was defined as all amino acids from the C-terminus that did not conform to any other domain, where boundaries were set by PKS and NRPS related pHMMs. Alignment of these terminal sequences demonstrates high conservation at the C-terminus represented in a sequence logo (Fig. 3.7a).[51] The most distinctive conservation are two glutamic acids and leucine forming a EExxL motif within the terminal six amino acids.

Whilst many COM_{2N} were identified within the atlas PKS/NRPS database, linking COM_{2N} domains to characterised biosynthetic pathways posed an additional challenge. Databases that provide a link between PKS and NRPS genes and their associated natural products assisted with this process (*e.g.* ClusterMine360[52] and MIBiG[43]), however the majority of COM_{2N} examples were identified by curating a bespoke database from literature searches of characterised

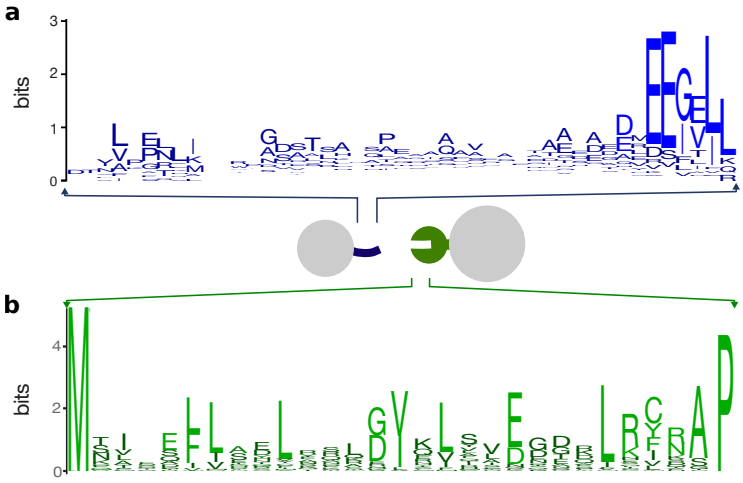

Figure 3.7: a) Sequence logo of putative COM_{2C} domains. b) HMM logo for a COM_{2N} domain.

biosynthetic pathways. Resulting compounds that utilise COM_{2N} domains during their proposed biosynthetic pathways are listed in Table. 3.1, along with their associated bioactivity. The abundance of COM_{2N} domains identified from characterised and cryptic biosynthetic pathways demonstrates their widespread use across a diverse set of genera.

Table 3.1: Fifty examples of COM_{2N}-containing pathways for bioactive natural products identified using a hidden Markov model for COM_{2N} domains.

Protein	Natural product	Species (example)	Bioactivity
AerB,J[53]	aeruginosin	*Microcystis aeruginosa*	protease inhibitor[54]
AjuL[55]	ajudazol	*Chondromyces crocatus*	antifungal[56]
ApdC,D[57]	anabaenopeptilide	*Anabaena* sp. 90	protease inhibitor[58]
AptD[59]	anabaenopeptin	*Anabaena* sp. 90	protease inhibitor[60]
BacB[61]	bacitracin	*Bacillus subtilis*	antibiotic[62]
BarG[63]	barbamide	*Lyngbya majuscula*	molluscicidal[64]
BlmVII[65]	bleomycin	*Streptomyces verticillus*	anti-tumour[66]
CmdC[67]	chondramide	*Chondromyces crocatus*	anti-cancer[68]
SCO7683[69]	coelibactin	*Streptomyces coelicolor*	-
ColG[70]	columbamide	*Moorea bouillonii*	cannabinomimetic[70]
CrpC[71]	cryptophycin	*Nostoc* sp. ATCC 53789	anti-tumour[72]
OciC[73]	cyanopeptin oci	*Planktothrix agardhii*	-
McnD,E[73]	cyanopeptolin	*Microcystis aeruginosa*	protease inhibitor[58]
CtaG[74]	cystothiazole	*Cystobacter fuscus*	antibiotic[75]
DidC[76]	didemnin	*Tistrella mobilis*	anti-tumour[77]
DkxG,J[78]	DKxanthene	*Myxococcus xanthus*	antioxidative[79]
Bamb_5915[34]	enacyloxin	*Burkholderia ambifaria*	antibacterial[80]
EtlC[81]	entolysin	*Pseudomonas entomophila*	hemolytic[81]
EpoB[82]	epothilone	*Sorangium cellulosum*	anticancer[83]
EqbG[84]	equibactin	*Streptococcus equi*	-
HctE,F[85]	hectochlorin	*Moorea porducens*	antifungal[86]
LeuA,E[87]	leupyrrin	*Sorangium cellulosum*	antifungal[88]

Table 3.1: (continued)

Protein	Natural product	Species (example)	Bioactivity
LynE,F[89]	lyngbyabellin	*Moorea bouillonii*	anticancer[90]
MassC[91]	massetolide	*Pseudomonas fluorescens*	antibiotic[91]
MelG[92]	melithiazol	*Melittangium lichenicola*	antifungal[93]
MicC,D[94]	microginin	*Microcystis aeruginosa*	treat hypertension[95]
Psm3b[96]	micropeptin	*Microcystis aeruginosa*	protease inhibitor[58]
MscI[97]	microsclerodermins	*Sorangium cellulosum*	antifungal[98]
MxaA[99]	myxalamid	*Myxococcus xanthus*	antibiotic[100]
MtaD,G[101]	myxothiazol	*Stigmatella aurantiaca*	antibiotic[102]
NcyE,F[103]	nannocystin	*Nannocyctis* sp. ST201196	anti-cancer[104]
NcpA[105]	nostocyclopeptide	*Nostoc* sp. ATCC53789	anticancer[106]
NosC[107]	nostopeptolide	*Nostoc punctiforme*	cell-differentiation[107]
OfaC[108]	orfamide	*Pseudomonas protegens*	insecticide[109]
PoaC[110]	poaeamide	*Pseudomonas poae*	antifungal[110]
PuwF[111]	puwainaphycin	*Cylindrospermum alatosporum*	anticancer[111]
PchF[112]	pyochelin	*Pseudomonas aeruginosa*	antifungal[113]
DepD[114]	romidepsin	*Chromobacterium violaceum*	anticancer[115]
SesB[116]	sessilin	*Pseudomonas* sp. CMR12a	anti-fungal[117]
SyfB[118]	syringafactin	*Pseudomonas syringae*	biosurfactant[119]
SypB[120]	syringopeptin	*Pseudomonas syringae*	antimicrobial[121]
TlmVII[122]	tallysomycin	*Streptoalloteichus hindustanus*	anticancer[123]
TaaB,C[124]	tolaasin	*Pseudomonas costantinii*	mushroom toxin[125]
TubC[23]	tubulysin	*Cystobacter* sp. SBCb004	anticancer[126]
VbxO	vibroxin	*Vibrio rhizosphaerae*	antibiotic (chapter 4)
ViscC[127]	viscosin	*Pseudomonas fluorescens*	antibiotic[128]
Sven_0517	watasemycin	*Streptomyces* sp. TP-A0597	antibiotic[129]
WlpC[130]	WLIP	*Pseudomonas putida*	antibacterial[130]
XtlC[131]	xantholysin	*Pseudomonas putida*	antimicrobial[131]
ZbmIX[132]	zorbamycin	*Streptomyces flavoviridis*	antibiotic[133]

The identification of many bioactive compounds using COM_{2N} domains provides an opportunity to exploit these features, in a similar manner to early experiments engineering COM_{1N} and COM_{1C} boundaries. This repertoire of natural products provide an opportunity to biosynthesise hybrid-compounds through genetic engineering, focused on intermolecular interactions provided by COM_2 interfaces. Whilst biosynthetic pathways are regularly divided into modules, corresponding to an elongation of a enzyme-bound intermediate, nature has divided biosynthetic pathways into proteins as biosynthetic components. Protein boundaries are regularly observed within an elongation module, where COM (or docking domains) facilitate substrate transfer, or enzyme activity, over a protein boundary. Representing COM_2-type boundaries on compounds highlights molecular components corresponding to protein boundaries (Fig. 3.8). This representation provides an opportunity to visualise natural products as molecular components

based upon protein boundaries with COM domains. Exploitation of these COM domains could facilitate the creation of novel hybrid-compounds, assembled in a combinatorial fashion, based upon a framework underpinning bioactivity.

3.2 Bioactivity of natural products utilising type 2 COM domains

The list of natural products utilising COM_2-junctions during their biosynthesis is extensive, and demonstrates a range of biological activity. Bacitracins are a family of commonly used antibiotics which interfere with cell wall and peptidoglycan biosynthesis. The COM_2-juction within bacitracin NRPS, can be mapped to an amide linking the macrocycle to the linear peptide region. It is plausible that COM_2-junctions could be exploited to exchange the linear component of bacitracin with another molecular constituent found upstream of a COM_2-junction from a difference biosynthetic pathway.

Romidepsin (also known as Istodax) contains a COM_2 junction during its biosynthesis, condensing valine with a polyketide. In 2009, romidepsin gained FDA approval for the treatment of cutaneous T-cell lymphomas after receiving a fast-track designation (Gloucester Pharmaceuticals). Later, in 2011, it was also approved for other peripheral T-cell lymphomas. It represents a noteworthy example of a natural product that exhibits beneficial pharmaceutical properties. Another class of anticancer drugs are epothilones, which employ COM domains during their biosynthesis. Epothilones prevent cancer cells from dividing by interfering with tubulin. Several analogues are in clinical development for various cancers. One epothilone analogue, ixabepilone was approved for the treatment of advanced breast cancer in 2007 (R-PHARM). The methyl located on the thiazole of epothilones, represents the COM_2-junction in the epothilone PKS-NRPS. The PKS, found upstream of the COM_2 boundary, contains an acetyl-PCP domain, which becomes condensed to cysteine from the downstream NRPS. By modifying

the PKS component of the biosynthetic pathway, the COM_2-junction could be utilised to generate novel epothilone analogues, where the methyl is exchanged for a larger polyketide, or hybrid polyketide-nonribosomal peptide molecule.

Large structural variations are represented across the collection of molecules that utilise COM_2 domains within their respective biosynthetic pathways. Some are relatively large cyclised lipopeptides, like syringopeptin, xantholysin, tolaasin, puwainaphycins, white line inducing principle, entolysin, massetolide and poaeamide. Each of these compounds show: variation in length, amino acid components, macro-cyclisation size and type, and the position of reactions mediated by COM_2 domains. Syringopeptin contains a particularly large protein (around 1.5 MDa), consisting of twelve extension modules encoded by a 40 kb gene. Such a large protein shows that some COM_2-junctions at NRPS-NRPS boundaries could potentially be replaced by a single protein (at least in *Pseudomonas syringae*). However, one advantage of using COM domains, compared with a single protein without a boundary, is the opportunity for COM crosstalk and sampling other potential biosynthetic pathways during evolution.

A common feature of COM_2-derived compounds is a thiazole at the COM_2-junction (*e.g.* myxothiazol, melithiazole, pyochelin, thiazostatin, watasemycin, coelibactin, equibactin, epothilone), or an oxazole (*e.g.* DKxantheme, leupyrrin). From the untapped potential of cryptic biosynthetic gene clusters, the most frequent COM_{2N}-containing genes are also predicted to produce thiazoles or oxazoles. Careful dereplication procedures will uncover the fuller extent of diversity within cryptic gene clusters and putative structures of natural products. Despite the relatively high frequency of PCP and cyclisation domains at COM_2 boundaries, many natural products contain thiazole and oxazoles without crossing COM boundaries. These features can be observed in natural products that utilize COM domains corresponding to positions elsewhere in a given compound (*e.g.* barbamide, coelibactin, cystothiazole, tubulysin).

The most common type of enzyme found adjacent to COM_{2C} and COM_{2N}

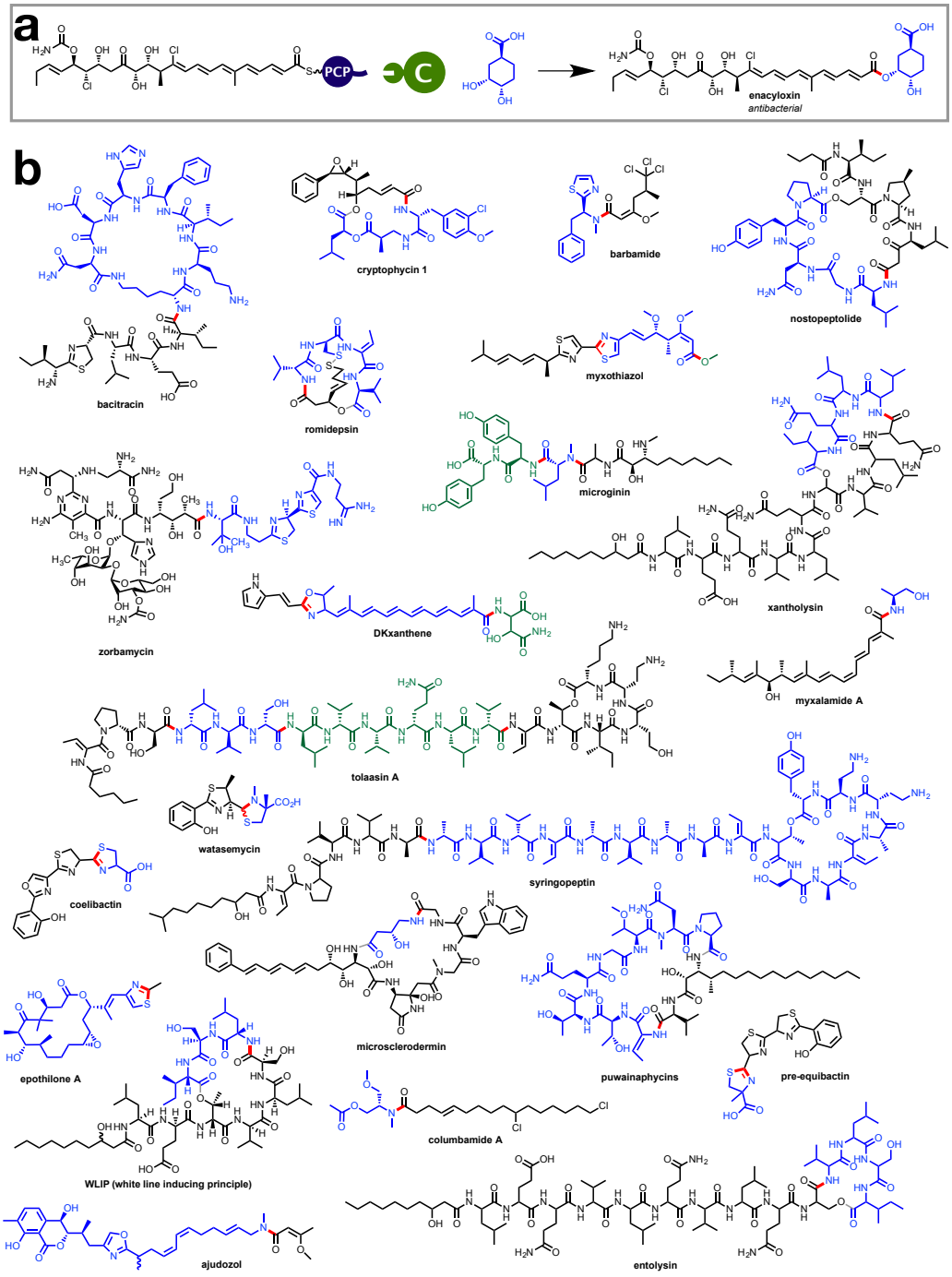

Figure 3.8: a) Enacyloxin IIa communication-mediated chain release. b) Examples of natural products biosynthesised utilising COM_2 domains. Red bold lines indicate the point of a COM_2 boundary during biosynthesis, which is also reflected in some colour changes for molecular constituents. Macrocyclisations, and other linkages, are indicated with grey bonds at boundaries with colour changes.

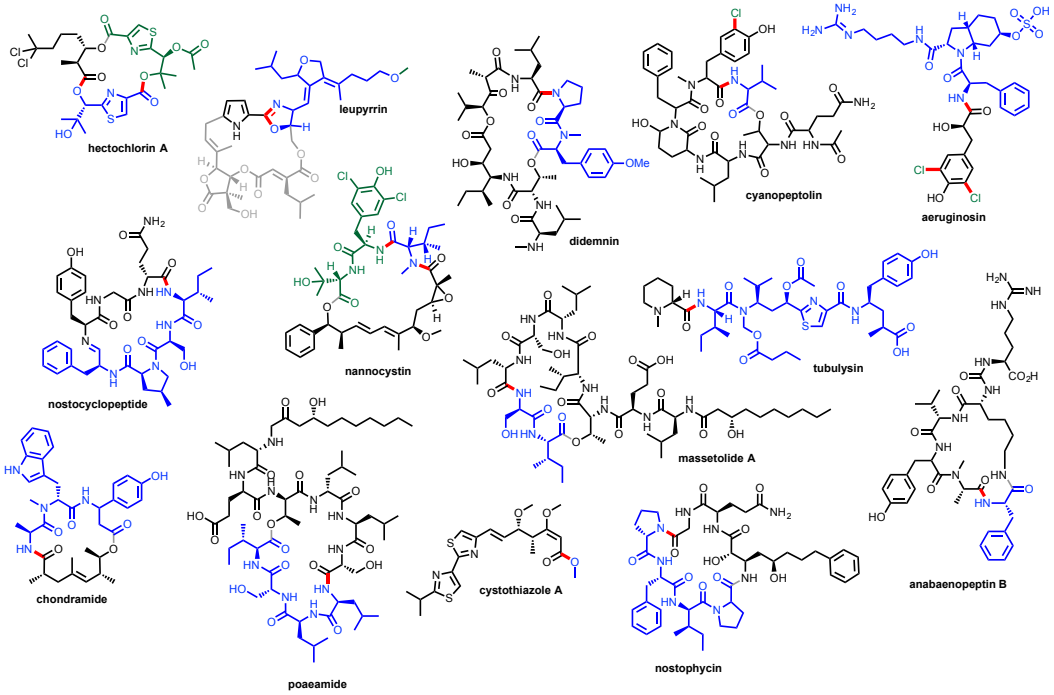

Figure 3.8 (cont.)

domains, within the atlas PKS/NRPS database, are PCP and heterocyclisation domains respectively. Other types of enzymes found bordering COM_2-junctions are shown in Fig. 3.9a. PCP domains can utilise COM_2 domains to pass a substrate to condensation domains (like in enacyloxin PKS-NRPS), and epimerising condensation domains. Thioester-reductase domains are identified downstream of some COM_{2N} domains, either as an isolated entity, or with an additional domain of unknown function (predicted to have a beta-lactamase-like fold). The COM_{2N} domain is situated in the centre of the protein in this case. PKS-NRPS hybrids can be formed over COM_2 boundaries, with ACP domains passing intermediates to condensation, or heterocyclisation, domains. These observations suggest a degree of promiscuity between ACP and condensation, or cyclisation, domains which could be exploited to engineer novel PKS-NRPS hybrid biosynthetic pathways.

Whilst the COM_{2C} motif is most frequently preceded by a carrier protein, it can also be found after oxidising enzymes (*e.g.* melithiazol and myxothiazol PKS-NRPSs). In these examples the modular setup is Cy-A-PCP-Ox-COM_{2C} followed

by COM_{2N}-Cy-A-Ox-PCP, where each oxidase acts on the cysteine-derived dihydrothiazoles within each module. The PCP-bound intermediate needs to locate the downstream cyclisation domain over the COM_2-junction despite the additional oxidase within the PCP-Ox-COM_{2C} module. This system raises some questions on the quaternary structure of these multi-enzyme complexes, and the relative positions of each enzyme around COM_2-junctions.

Whilst the atlas PKS/NRPS database was used to identified many COM_{2N} domains, the larger non-redundant protein database was used to explore variation in the domains acting across COM_2 boundaries. Using HMM searches, more exotic systems were identified that deviated from those found within the atlas PKS/NRPS database (Fig. 3.9b). In one example, a thioesterase (type I) was identified to putatively facilitate chain release over a COM_2-junction from an upstream carrier protein (Appendix Fig. 7.14). This is similar to the thioester-reductase example, and could potentially be exploited for the targeted release of carrier protein-bound intermediates, or similar bioengineering applications exploiting COM_2 boundaries.

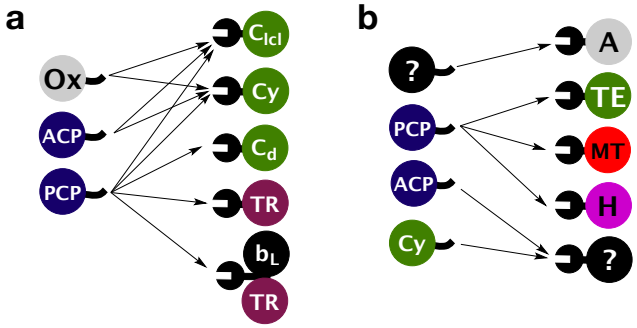

Figure 3.9: Examples of domains found adjacent to COM_{2N} domains. a) Domain pairs identified from atlas PKS/NRPS database, C_{lcl} - D-amino acid condensation domain, Cy - heterocyclisation domain, C_d - dual epimerising and condensing domain, TR - thioester-reductase, b_L - beta-lactamase-like, Ox - oxidase. b) domain pairs from searching the non-redundant protein database, H - halogenase, question marks denote domains of unknown function.

Two examples were found involving enzymes of unknown function based on homology to characterised enzymes (Appendix Fig. 7.9, 7.10). These proteins occur upstream of an adenylation domain, and downstream of a cyclisation domain

or ACP. One cryptic biosynthetic pathway was identified with a *trans*-acting COM_{2N}-cMT protein (Appendix Fig. 7.12). This hybrid PKS-NRPS has two COM_{2C} domains and three COM_{2N} domains. Assuming biosynthetic colinearity, the COM_{2N}-cMT protein shares a COM_{2C} domain with a condensation domain. It putatively acts to C-methylate a PCP-bound threonine residue, before condensation with N-methylated proline. Though this is an isolated cryptic example pathway, it represents the diversity of biochemistry facilitated by COM_2 domains.

There are two families of chlorinated compounds which contain COM_2 domains within their biosynthetic pathways, aeruginosins and cyanopeptolins. Each NRPS contains one COM_{2C} domain and two COM_{2N} domains, where one of the COM_{2N} domains is located N-terminal to the halogenase enzyme (Fig. 3.10). Assuming colinearity, the halogenase is proposed to utilise the COM_{2N} domain to chlorinate the phenyl PCP-bound intermediate from the downstream NRPS. Then, the upstream NRPS condenses the chlorinated intermediate, with the PCP-bound amino acid, from the downstream NRPS module. This would suggest some level of promiscuity of the COM_{2C} domain towards each COM_{2N} domain. To obtain a high proportion of chlorinated product the halogenase must perform its catalytic activity preferentially over the downstream condensation domain. Such information may be encoded in each COM_{2N} domain, in terms of binding affinity with COM_{2C}. The identification of the COM_{2N}-halogenase protein reveals a biophysical basis for chlorination activity acting on the PCP-bound intermediate, compared to the NRPS product after chain release.

3.2.1 COM_{2N} domain located away from a terminus

The hybrid PKS-NRPS, involved in the biosynthesis of the neurotoxin jamaicamide,[134] features two COM_{2N} domains (Fig. 3.11). A COM_{2C} motif can be found at the C-terminus of JamN, which is proposed to form an intermolecular interaction with the COM_{2N} domain identified in JamO. Another COM_{2N} is found in JamL, between the ACP and C domains, at the PKS-NRPS interface. An additional

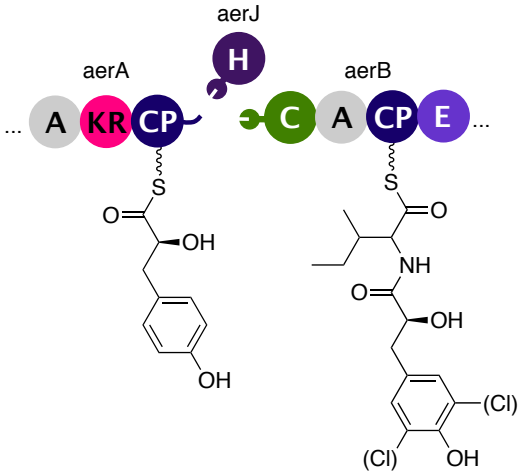

Figure 3.10: COM_2-boundary within aeruginosin NRPS showing a *trans*-acting COM_{2N}-halogenase.

COM_{2C} domain was not identified at the C-terminus of any protein within the jamaicamide biosynthetic gene cluster, and was also not identified directly upstream of the COM_{2N} domain in JamL. It is unclear whether the COM_{2N} domain within JamL performs an intermolecular interaction, or is a consequence of genetic recombination. The previously unidentified oxidase in JamL may utilise an interaction between COM_{2N} from JamL and the COM_{2C} from JamN to oxidise the ACP-bound intermediate on JamN. However, these stages of jamaicamide biosynthesis are poorly understood due to differences between the predicted molecular structure arising from this unusual domain architecture and the structure of jamaicamides.

3.3 Crosstalk at COM_2 junctions

The identification of putative COM_{2N} domains, from many biosynthetic pathways, provides an opportunity to exploit these intermolecular interactions to design novel products. Some COM_2 junctions may have a degree of inherent promiscuity, as shown in examples where two COM_{2N} domains share a single COM_{2C} domain. Whilst other examples, such as those with two COM_2-junctions, must prevent crosstalk, where a single biosynthetic route is followed.

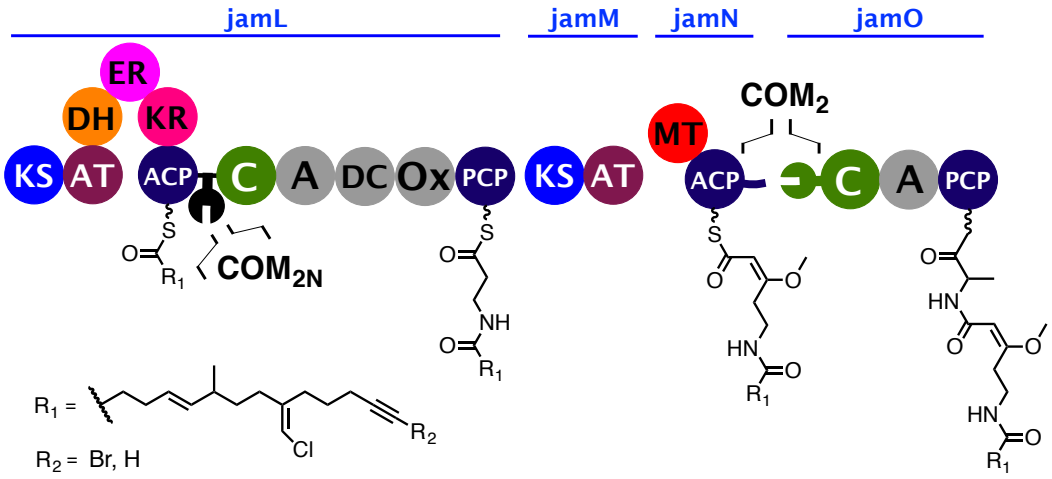

Figure 3.11: Biosynthesis of jamaicamide, showing COM domains. Intermediates represent those previously proposed,[134] but the domain annotations have been updated (DC - glutamate decarboxylase, Ox - oxidase).

3.3.1 Watasemycin NRPS

Two proteins within the watasemycin NRPS, Sven_5912 and Sven_5917, are proposed to contain a COM_2-junction. A COM_{2N} domain was identified within Sven_0517 from watasemycin NRPS, and subsequently a putative COM_{2C} domain was proposed in Sven_0512 (Fig. 3.12a). These NRPSs were used as a model system, to investigate the promiscuity of COM_2-junctions. By creating a hybrid COM_2-junction using the COM_{2C} domain from Sven_0512 NRPS, and the COM_{2N} domain from Bamb_5915 (of enacyloxin biosynthesis), a hybrid watasemycin-enacyloxin compound could be biosynthesised, assuming some substrate tolerance of downstream enzymes. This compound would consist of the intermediate bound to the C-terminal PCP of Sven_0512, condensed with the substrate of the enacyloxin condensing enzyme, (1′*R*,3′*R*,4′*S*)-DHCCA, at the 3′-OH position (Fig. 3.12b).

The work towards the creation of a watasemycin-enacyloxin compound, included chemical synthesis of precursors, and *in vitro* assays, which were performed in collaboration with Matthew Beech (chemical synthesis of watasemycin intermediate), Shanshan Zhou (cloning PCP-COM_{2C} from Sven_0512), Christian Hobson (coupling watasemycin intermediate to protected pantetheine) and Dr Matthew

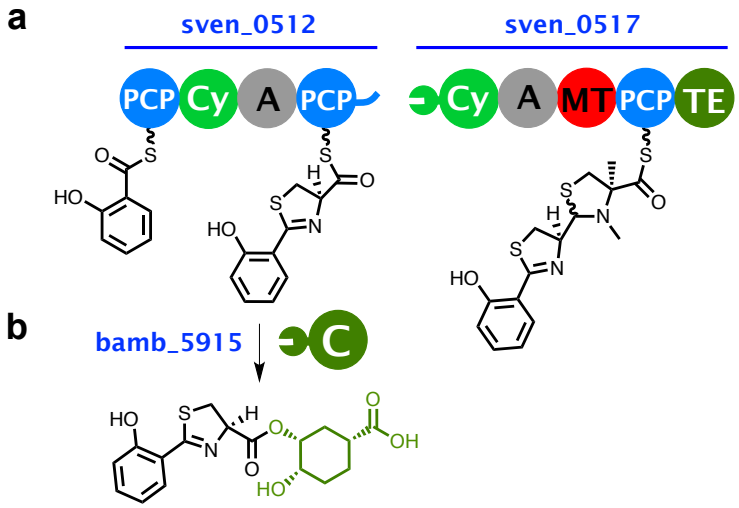

Figure 3.12: a) Watasemycin NRPS. b) Hypothetical watasemycin-enacyloxin hybrid compound.

Jenner (provided CoA biosynthetic enzymes and performed PCP loading assay). The free-acid of the PCP-bound intermediate, (*R*)-2-(2-hydroxyphenyl)-4,5-dihydrothiazole-4-carboxylic acid, was chemically synthesised and coupled to protected pantetheine, before deprotection. A set of three enzymes were used to form a CoA derivative from the pantetheine-coupled compound, a pantothenate kinase (PanK), a phosphopantetheine adenyltransferase (PPAT) and a dephosphocoenzyme A kinase (DPCK).[135] Forming the CoA derivative, provides a closer mimic of the substrate of phosphopantetheinyl-transferases (*e.g.* Sfp from surfactin NRPS). Loading a CoA derivative onto an apo-carrier protein can usually be accomplished using the relatively promiscuous Sfp enzyme with Mg^{2+} as a co-factor. Unfortunately, the CoA derivative of the PCP-bound intermediate could not be transfered onto the C-terminal PCP-COM_{2C} from Sven_0512 using Sfp. The relatively hindered α-position at the sterocenter of the thiazole is one possible explanation for the observed temperance of Sfp for this particular substrate.

3.3.2 Cross-talk assay

To investigate the promiscuity of COM_2 domain interactions, a condensation assay was designed using the PCP-COM_{2C} domain from watasemycin NRPS and

the COM_{2N}-C domain from enacyloxin biosynthesis. The assay is based on the transfer of an acetate unit across the hybrid communication boundary for condensation with shikimate, as the more complex watasemycin intermediate could not be loaded onto PCP-COM_{2C} using Sfp. Acetyl-PCP-COM_{2C} was observed to condense with shikimate via the COM_{2N}-C domains (Fig. 3.13a), exhibiting a similar profile E.I.C. as for the PCP-COM_{2C} from enacyloxin (Fig. 3.13b). The COM_{2C} domains share a common EYEEG subsequence from the terminal nine amino acids (SEYEEGVIR from enacyloxin PKS-NRPS and LVEYEEGEL from watasemycin NRPS), which may contribute to the inherent compatibility of this hybrid COM_2-junction.

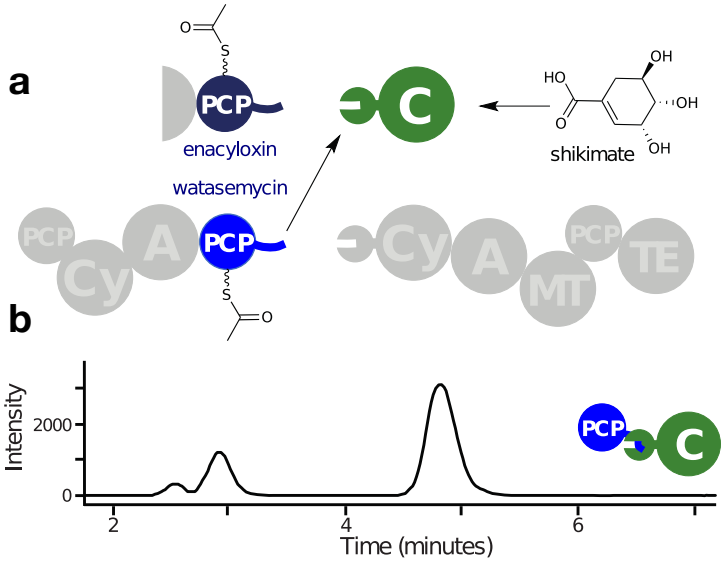

Figure 3.13: a) Crosstalk scheme with watasemycin and enacyloxin domains. b) E.I.C for acetylated shikimate from crosstalk assay.

3.3.3 Other carrier protein-condensation domain boundaries

The COM_{2N} HMM has revealed many instances of COM_2-junctions at PCP-to-condensation domain protein boundaries. However, some PCP and condensation domain boundaries do not utilise COM_2 domains. The VpsA-VpsB boundary within vancomycin NRPS, and homologous NRPSs, possess COM1 domains at PCP and condensation domain boundaries. In these examples the PCP domain

possesses an adjacent COM_{1C} domain, even though COM_{1C} domains are most commonly associated with an epimerisation domain at a boundary with a condensation domain. The relatively short length of COM_{1N} domains (around 17 amino acids), means that the COM_{1N} domain locates to the N-terminal subdomain of the condensing enzyme. If the binding partner is an epimerase-COM_{1C} protein, then this domain must also situate to the similar vicinity due to the short length of COM_{1C}. This must also hold true when the binding partner is PCP-COM_{1C}. In VpsA-VpsB within vancomycin NRPS, there appears to be a small linker of approximately 11 amino acids between the PCP and the COM_{1C} domain. This may help facilitate the location of the PCP to the entrance face of the downstream condensation domain, and the exit face of the upstream condensation domain. For comparison, COM_{2N} domains are located at the C-terminal subdomain of condensing enzymes, at boundaries between PCP-COM_{2C} and COM_{2N}-C domains (Fig. 3.14), where a 17 amino acid linker extends from the condensation domain to the COM_{2N} domain.

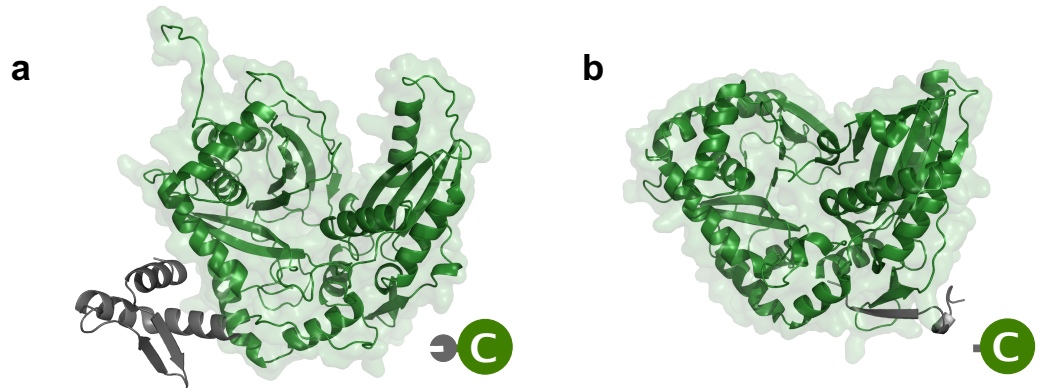

Figure 3.14: a) COM_{2N}-C structure from enacyloxin PKS-NRPS (Bamb_5915) b) COM_{1N}-C structure from surfactin NRPS.[24]

As well as COM_{1N} and COM_{2N} domains, PKS-like docking domains associated with ACP and KS protein boundaries can also be identified preceding condensation domains. In the bacillaene biosynthetic machinery, a PKS-NRPS boundary lies in between an ACP and a condensation domain. The condensation domain is proceeded by a region which was identified as a type 3 docking domain

using a pHMM. To date, this class of docking domain has only been reported before KS domains. This is the first instance of a PKS-like docking domain found preceding an NRPS condensation domain. This type of docking domain is relatively short and has a helical structure. It is expected that the higher-order architecture of this ACP-dd_{3C} and dd_{3N}-C complex would resemble the complex of PCP-COM_{1C} and COM_{1N}-C, despite a different type of intermolecular interaction.

3.4 Conclusion and further work

The development of HMMs to describe docking domains and COM domains have demonstrated the ubiquity of these appendages within modular PKSs and NRPSs. However, improved methods for their identification and classification, remains a challenge due to their small size. The dd2-class of docking domains have been identified within 22 biosynthetic pathways, in addition to those previously reported. This further demonstrates the widespread use of docking domains, and opportunities for exploiting this class of interaction sites. HMMs have uncovered instances of docking domains that are not located at protein termini. Experimental investigation of these internal docking domains may elucidate functional recruitment of *trans*-acting enzymes, or they may exist as a consequence of evolutionary development.

Inspection of terminal amino acid sequences from PKS and NRPS proteins provided motifs corresponding to docking domains, and components of domains that accommodate the terminal amino acid sequence. Amongst type II PKS domains and loading domains, four classes of docking domains were independently identified within PKSs, during searches for terminal motifs. Resulting HMMs facilitated the identification of each of these classes from databases of PKS and NRPS proteins. Extension of this method, to larger PKS and NRPS protein databases, and investigations of less frequent motifs, should elucidate novel forms of docking domains, and facilitate their annotation within modular PKSs and

NRPSs.

A novel form of intermolecular interaction has been proposed at boundaries between C-terminal *trans*-AT KS domains and N-terminal DH domains, which diverges from the paradigm of two terminal docking domains. A C-terminal peptide is proposed to interact directly with the N-terminal DH domain. In each example of this domain architecture, the KS domain does not contain a *trans*-AT docking domain, or an associated motif that binds to a KS domain before forming the following modular components. The consistent failure to identify this motif may suggest a breakdown in the quaternary structure of adjacent modules, as it appears to direct the downstream module away from the upstream module. This is one possible explanation for additional KR activity, where KR domains are only located in adjacent modules. Biophysical and biochemical experiments should provide additional insight into these features.

The classification of N-terminal DH domains, involved in this putative form of intermolecular interaction, serendipitously described DH domains that were not located at the N-terminus. However, each of these DH domains correlated with a (2*E*,4*Z*)-configured diene, situated at the ACP-bound intermediate within the same module. The identification of these DH domains, within archetypal *cis*-AT PKS domain organisations, suggests that this class of DH domain acts sequentially to dehydrate two hydroxyls, in a stereospecific manner. An HMM that describes this domain has predicted many (2*E*,4*Z*)-configured dienes within characterised PKS products, but its value in predicting these features within cryptic PKSs, is yet to demonstrate its full potential. The proposed function and mechanism of this class of DH domain would benefit from *in vitro* assays of substrate mimics to probe their native function.

The identification of a COM_{2N} domain in the enacyloxin PKS-NRPS led to the development of an HMM that described intermolecular interactions within a range of biosynthetic pathways. These results show a clear region of conservation on the corresponding C-terminal binding partner, but methods for its indepen-

dent identification remains a considerable challenge. The identification of type 2 COM boundaries may provide assistance to establish biosynthetic order within biosynthetic pathways, when the colinearity rule is not adhered to. The number of characterised natural products, biosynthesised using type 2 COM domains is extensive, demonstrating a range of beneficial activities from a diverse set of producing organisms. The number of cryptic biosynthetic gene clusters encoding COM domains represents even more diversity, in terms of catalysis across protein-protein boundaries. There are *trans*-acting thioesterases, thioester-reductases, halogenase, methyltransferases and enzymes of unknown function, to name a few. This discovery provides many opportunities for exploiting these attributes, and ventures to engineer novel biosynthetic pathways. The identification of NRPSs containing multiple type 2 COM boundaries, suggests a measure of selectivity is required to avoid native crosstalk. However, the exact nature of this selectivity remains an opportunity for experimental investigation, within biosynthetic pathways containing multiple COM domains.

Chapter 4

Genomics-driven discovery of an enacyloxin analogue, vibroxin

4.1 Genomics-driven drug discovery

The number of microorganisms with sequenced genomes has escalated, as DNA sequencing technology has become faster and inexpensive. Databases of genomes can be mined, by searching for biosynthetic gene clusters using a set of computational programs,[136] to evaluate their biosynthetic potential. Within any given gene cluster, genetic components can be assigned putative functions based upon homology to genes of known function. In the case of large modular PKS and NRPS genes, their corresponding amino acid sequences can be partitioned into catalytic domains. Analysis of these domains can be used to predict the specificity of enzymes, and thus infer putative core molecular structures of corresponding natural products.

4.2 Rhizosphere-associated *Vibrio*

Many Gram-negative bacteria from the *Vibrio* genus are renowned human pathogens, causing cholera[137] and food-borne infections[138] globally. Outbreaks of *Vibrio*-related food poisoning frequently occur in tropical coastal areas, and often orig-

inate from undercooked or raw shellfish,[139] as many *Vibrio* species live symbiotically within the organs of marine invertebrates[140] and fish.[141] Whilst the pathogenic properties of marine *Vibrio* sp. are relatively well-studied, it is emerging that some *Vibrio* sp. diverge from this paradigm, occupying alternative ecological niches, such as the plant rhizosphere within salt marshes.[142–144]

In 2007, a red-pigmented strain, *Vibrio rhizosphaerae* MSSRF3^{T}, was isolated from the rhizosphere region of mangrove-associated wild rice (*Porteresia coarctata* Tateoka) within the Pichavaram forest (India). It was the first *Vibrio* strain to show growth-promoting activity towards plants, as well as antibacterial activity against the phytopathogens *Xanthomonas oryzae*, *Erwinia carotovora* and *Pseudomonas syringae*.[145] This activity was observed in cell-free culture fractions, indicating that *V. rhizosphaerae* produces an extracellular bioactive compound. However, no such compound has so far been identified, and the underlying mechanism causing this beneficial bioactivity remains elusive.

4.3 Natural products from *Vibrio*

Marine *Vibrio* sp. produce an extensive collection of bioactive natural products[146] (Fig. 4.1), including anticancer, antibacterial, quorum sensing interference compounds and sodium channel inhibitors. To date, most natural products isolated from *Vibrio* sp. are derived from non-ribosomal peptides, or are N-containing molecules that are biosynthesised by NRPS-independent methods. Some well-known examples of compounds produced by *Vibrio* strains with potential as promising drug leads include prodigiosin, kahalalide F, and tetrodotoxin.

The red-pigmented antibiotic prodigiosin from *V. gazogenes*,[147] *V. ruber*[148] and *V. psychroerythreus*,[149] along with its cyclised prodigiosin analogue,[150,151] has demonstrated benefits for anticancer therapy.[152] Its drug variant, obatoclax, reached phase II clinical trials[153] for the treatment of leukemia, lymphoma, mastocytosis and myelofibrosis, but exhibits a low therapeutic window due to toxic effects.[154]

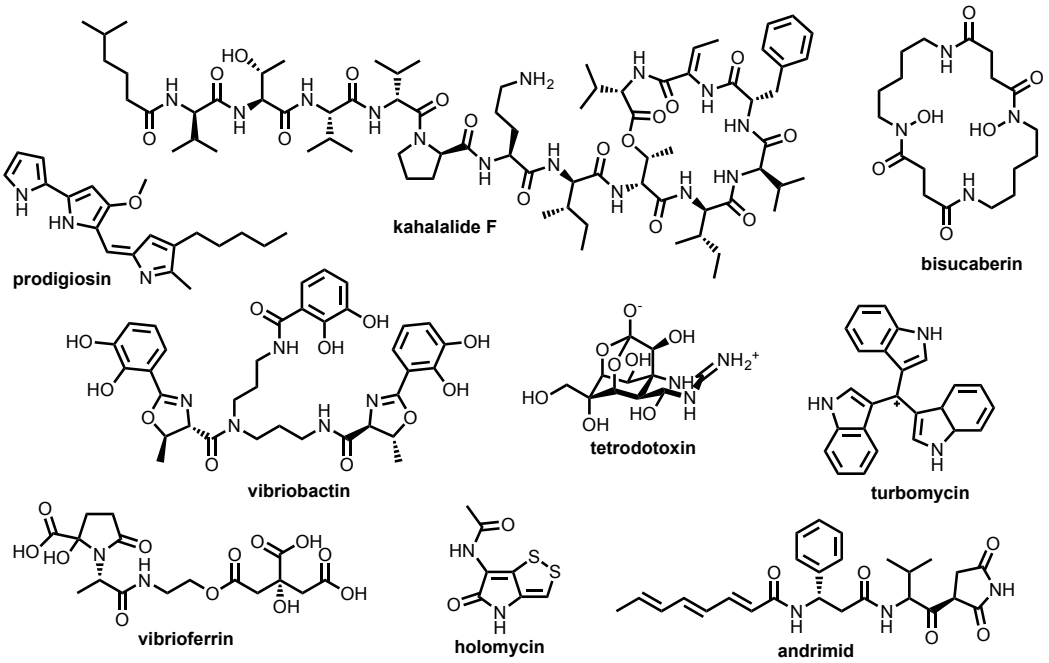

Figure 4.1: Examples of bioactive compounds isolated from *Vibrio* sp.

Kahalalides, isolated from *V. mediterranei*,[155] display an attractive spectrum of activity against cancer cell lines and AIDS-related opportunistic infections.[156] The compound kahalalide F reached phase II clinical trials for liver carcinoma, prostate and lung cancer,[156] and for treatment of patients suffering from severe cases of psoriasis (PharmaMar).

Perhaps the most infamous natural product from *Vibrio* is tetrodotoxin.[157,158] Some *Vibrio* sp. live symbiotically within pufferfish and produce the potent neurotoxin tetrodotoxin, making the pufferfish one of the most poisonous vertebrates in the world. Tetrodotoxin, extracted directly from pufferfish, is currently in phase II and phase III clinical trials for pain relief during chemotherapy treatments.[159]

Whilst marine *Vibrio* sp. have been shown to produce a range of bioactive compounds, rhizosphere-associated *Vibrio* sp. remains an underexploited source of antimicrobial metabolites. In an effort to discover novel bioactive compounds, the genome of *V. rhizosphaerae* was analysed for the presence of specialised metabolite biosynthetic gene clusters.

4.3.1 Genome-mining of *V. rhizosphaerae*

Forty seven scaffolds from the genome assembly of *V. rhizosphaerae* MSSRF3^{T} were processed using antiSMASH[42] to identify specialised metabolite biosynthetic gene clusters. A total of twelve clusters were identified: nine NRPSs, one ectoine pathway, one bacteriocin pathway and one *trans*-AT PKS. This number of biosynthetic gene clusters demonstrates the untapped potential for bioactive metabolite production within this strain.

Some of the NRPS clusters were incomplete, located at scaffold boundaries. While other NRPS clusters were very small, containing only one or two adenylation domains. The largest complete NRPS gene cluster was situated 4 kb from a *trans*-AT PKS cluster on scaffold KL543967. It consists of three large multimodular genes encoding six adenylation domains with predicted specificities for asp-gly-ser-ser-?-gly (Fig. 4.2). Genes encoding a putative dioxygenase and peptidase were located peripherally. However, the product of this biosynthetic gene cluster remains unknown.

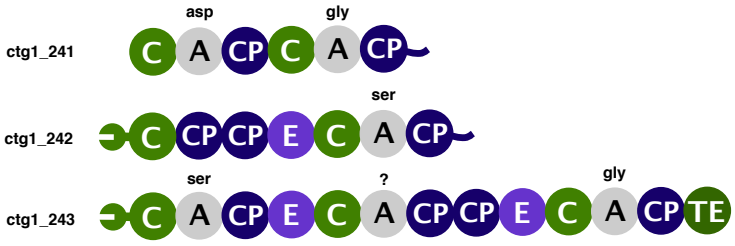

Figure 4.2: NRPS from scaffold KL543967 of *V. rhizosphaerae*.

The *trans*-AT PKS gene cluster (Fig. 4.7) from scaffold KL543967 of *V. rhizosphaerae* MSSRF3^{T} was found to have significant homology to the enacyloxin biosynthetic gene clusters from *B. ambifaria*[80] and *B. gladioli*[160] (Appendix Table 7.2). Thus, a search was conducted to uncover an enacyloxin-like natural product from *V. rhizosphaerae*, named vibroxin.

4.4 Vibroxin structure elucidation

The production of vibroxin by *V. rhizosphaerae* was not observed from cultures grown in liquid media (TSB2S and BSM2S), and only small amounts could be detected in ethyl acetate extracts from cultures grown on solid media. Enhanced production of vibroxin compounds was observed using glycerol as a carbon source instead of glucose (Fig. 4.3). Thus, BSM2S-glycerol solid medium was chosen for larger-scale (1 L) production of vibroxin. Vibroxin was purified from the ethyl acetate extracts of the agar by semi-preparative HPLC to give 0.3 mg purified vibroxin as a yellow/brown solid for structural characterisation by NMR spectroscopy and mass spectrometry.

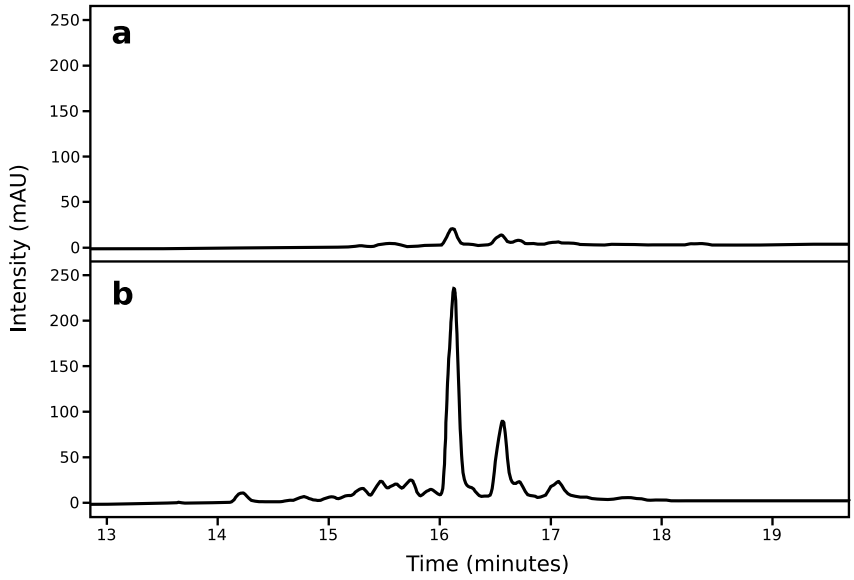

Figure 4.3: LC-MS analysis of ethyl acetate extracts from *V. rhizosphaerae* cultures, measuring absorbance intensity at 360 nm. Extracts were derived from cultures grown on equal amounts of BSM2S-agar with; a) glucose as a carbon source, b) glycerol as a carbon source.

The crude ethyl acetate extract of *V. rhizosphaerae* grown on BSM-agar contained a prominent red-pigmented compound, similarly to the culture itself. LC-HR/MS analysis of the extract showed the production of an analyte with an absorption maximum at 555 nm, and a molecular formula $C_{20}H_{25}N_3O$ (calculated for $C_{20}H_{26}N_3O^+$: 324.2070, found: 324.2092). This molecular formula is identical to that of the red pigmented prodigiosin.

Aside from the red-pigmented compound, LC-HR/MS analysis of the crude extract revealed the presence of one major and one minor metabolite with an absorption maximum around 360 nm (Fig. 4.3), indicative of the predicted pentaene subunit vibroxin (similarly to enacyloxin).

4.4.1 LC-HR/MS

The molecular formula of vibroxin was shown to be $C_{33}H_{47}ClO_9$ (m/z calculated for $C_{33}H_{47}ClNaO_9^+$: 645.2806, found: 645.2803) (Fig. 4.4). Additionally, in-source fragmentation of the $[M+H]^+$ ion yielded a daughter ion with m/z 605.2881, corresponding to dehydration of the parent ion. The high-resolution mass spectrum observed for vibroxin is consistent with the predicted isotopic distribution for a $C_{33}H_{47}ClNaO_9^+$ species (Fig. 4.4).

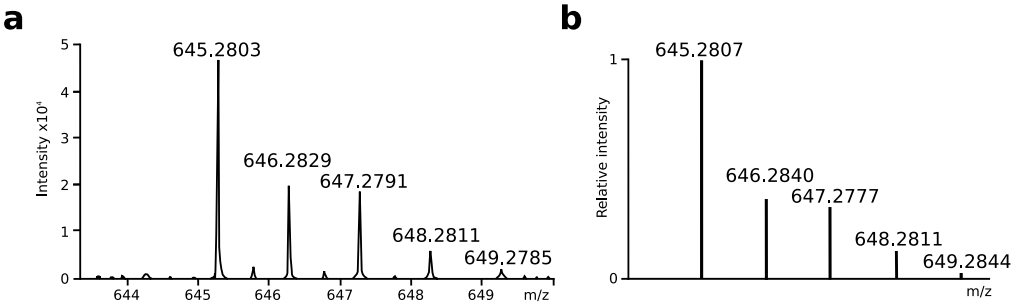

Figure 4.4: a) MS isotopic pattern for ion 645.2803 from the retention interval 16.0-16.2 minutes, b) predicted mass spectrum for $C_{33}H_{47}ClNaO_9{}^+$.

4.4.2 NMR spectroscopic analysis

The structure of vibroxin was solved using ^{1}H-^{1}H COSY, ^{1}H-^{13}C HMBC, ^{1}H-^{13}C HSQC and ^{1}H-^{1}H NOESY correlation spectra (Fig. 4.5). The ^{1}H and ^{13}C assignments are listed in Table. 4.1. Quaternary carbon chemical shifts were calculated from HMBC resonances. A characteristic doublet at 1.02 ppm suggests the presence of an *iso*-propanyl group at the terminus distal to the cyclohexane moiety.

A $^3J_{H,H}$ coupling constant of 15.1 Hz for H-18 and H-19 is consistent with an *E*-configured double bond between positions C-18 and C-19. Position C-15

exhibits couplings to C-14 and C-16, and has a chemical shift consistent with a hydroxyl group at this position. The minor vibroxin species, coined *iso*-vibroxin, demonstrated visible light-inducible isomerisation from vibroxin during purification. This species corresponds to the retention interval 16.5-16.6 minutes. As a consequence, light exposure was minimised during all procedures after purification of vibroxin.

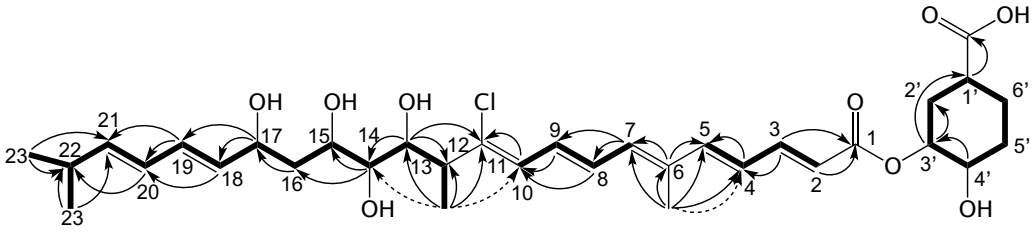

Figure 4.5: The structure of vibroxin showing correlations from NMR experiments: ^{1}H-^{1}H COSY (bold), ^{1}H-^{13}C HMBC (arrows), ^{1}H-^{1}H NOESY (dashed arrows).

4.4.3 Base hydrolysis of vibroxin

The cyclohexane moiety at the terminus of vibroxin contains three stereocenters. Purified vibroxin was treated with a potassium hydroxide to hydrolyse the ester at the junction between the polyketide and the dihydroxy-cyclohexane carboxylic acid (DHCCA) to investigate the stereochemistry of this derivative (Fig. 4.6a). LC-HR/MS analysis was performed on hydrolysed vibroxin, hydrolysed enacyloxin, and authentic standards (1′*S*,3′*R*,4′*S*)-DHCCA and (1′*R*,3′*R*,4′*S*)-DHCCA. The DHCCA products were identified from extracted ion chromatograms using m/z=183.0628, which corresponds to the sodiated adduct (Fig. 4.6b). The retention time of the DHCCAs from hydrolysed vibroxin, hydrolysed enacyloxin and the (1′*S*,3′*R*,4′*S*)-DHCCA standard are consistent with each other, whilst the (1′*R*,3′*R*,4′*S*)-DHCCA standard possesses a later retention time (the small impurity with a retention time of 7.2 minutes arises from another DHCA diastereometer). This result provides support that the relative stereochemistry of the cyclohexane moiety in vibroxin is the same as in enacyloxin, although the

Table 4.1: NMR assignment for vibroxin (MeOD, 500 MHz, 298 K).

Position	δ_H**/ppm** **(no. of protons, multiplicity, J/Hz)**	δ_C**/ppm**
1′-COOH		178.1
1′	2.52 (1H, m)	38.5
2′	2.17 (1H, m) 1.74 (1H, m)	31.5
3′	5.22 (1H, m)	72.2
4′	3.73 (1H, m)	69.6
5′	1.82 (2H, m)	28.3
6′	2.05 (1H, m) 1.57 (1H, m)	26.8
1		167.4
2	6.03 (1H, d, 15.0)	120.2
3	7.45 (1H, dd, 15.0, 11.0)	145.1
4	6.55 (1H, dd, 15.0, 11.0)	125.9
5	6.78 (1H, br d 15.0)	145.3
6		135.8
6-Me	1.97 (3H, br s)	11.2
7	6.44 (1H, br d, 10.0)	135.6
8	6.77 (1H, m)	130.1
9	6.77 (1H, m)	130.1
10	6.44 (1H, br d, 9.5)	126.4
11		140.7
12	2.88 (1H, dq, 9.5, 7.0)	46.0
12-Me	1.13 (3H, d, 7.0)	14.8
13	3.96 (1H, br, d, 9.5)	70.8
14	3.37 (1H, m)	72.8
15	3.94 (1H, m)	68.2
16	2.05 (1H, dd, 17.0, 4.5) 1.57 (1H, m)	41.1
17	4.41 (1H, m)	68.8
18	5.67 (1H, dd, 15.0, 6.5)	134.2
19	6.21 (1H, dd, 15.0, 10.0)	129.7
20	6.03 (1H, dd, 15.0, 10.0)	126.9
21	5.65 (1H, dd, 15.0, 6.5)	141.1
22	2.33 (1H, sept, 7.0, 6.5)	31.3
23	1.02 (6H, d, 7.0)	21.4

absolute configuration has not been elucidated to date.

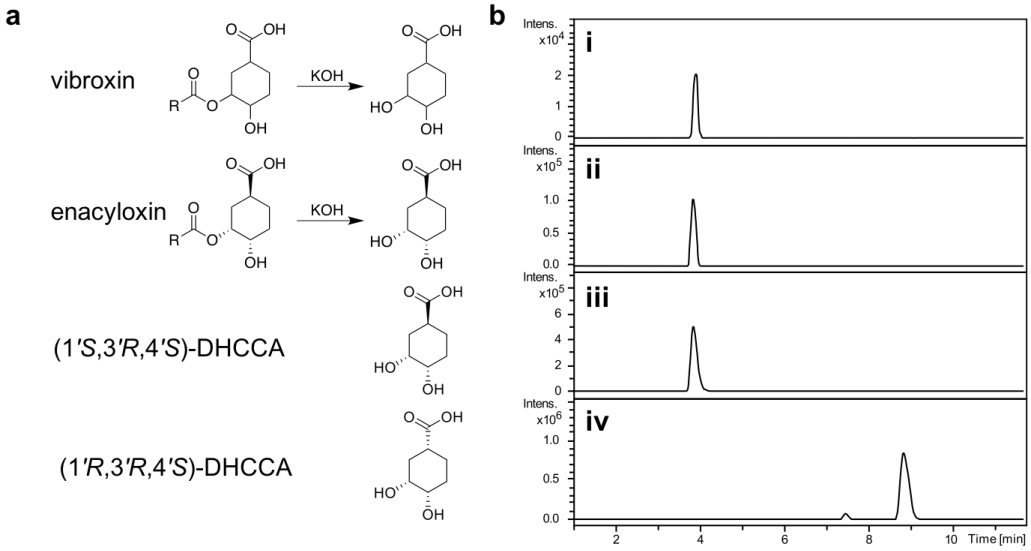

Figure 4.6: a) Base hydrolysis reaction for vibroxin and enacyloxin, showing the sterochemistry of DHCCA standards. b) Extracted ion chromatograms from LC-HR/MS analysis for DHCCA using m/z=183.0628 corresponding to the sodiated adduct. i) hydrolysed vibroxin, ii) hydrolysed enacyloxin IIa, iii) (1′*S*,3′*R*,4′*S*)-DHCCA, iv) (1′*R*,3′*R*,4′*S*)-DHCCA.

4.5 Vibroxin gene cluster

4.5.1 Gene annotations

The vibroxin biosynthetic gene cluster contains seven large PKS genes (*vbxD-vbxJ*) encoding modular components of polyketide synthesis, of which *vbxJ* is a hybrid *cis*-AT and *trans*-AT PKS gene based on phylogenetic analysis. The domains proximal to the C-terminus of VbxJ constitute the *trans*-AT-like region, where a non-elongating KS domain is present (KS^0). This domain shall hereby be referred to as an ACP *trans*-acylase enzyme (TA) which more accurately describes its function. A hybrid PKS-NRPS gene was identified by domain components encoded by *vbxQ*. A type II thioesterase is encoded by *vbxC*.

An NRPS-like condensation domain was identified in VbxO, where a COM_{2N} domain is located at the N-terminus. A putative COM_{2C} motif is proposed to

be present at the C-terminus of VbxQ within the sequence DEYEEGFIV (*cf.* SEYEEGVIR in the homologous Bamb_5917). Genes *vbxP*, *vbxN*, *vbxM*, *vbxL* and *vbxK* are proposed to encode enzymes involved in 3′,4′-dihydroxycyclohexane carboxylic acid (DHCCA) biosynthesis by homology to genes from the enacyloxin biosynthetic gene cluster. Tailoring enzymes include a putative non-heme iron dependent oxygenase, *vbxB*, and a putative $FADH_2$-dependent chlorinase, *vbxA*. There are two regulatory genes, *vbxR* and *vbxS*. At the cluster boundary, *vbxT* is predicted to encode a TIM-barrel efflux pump.

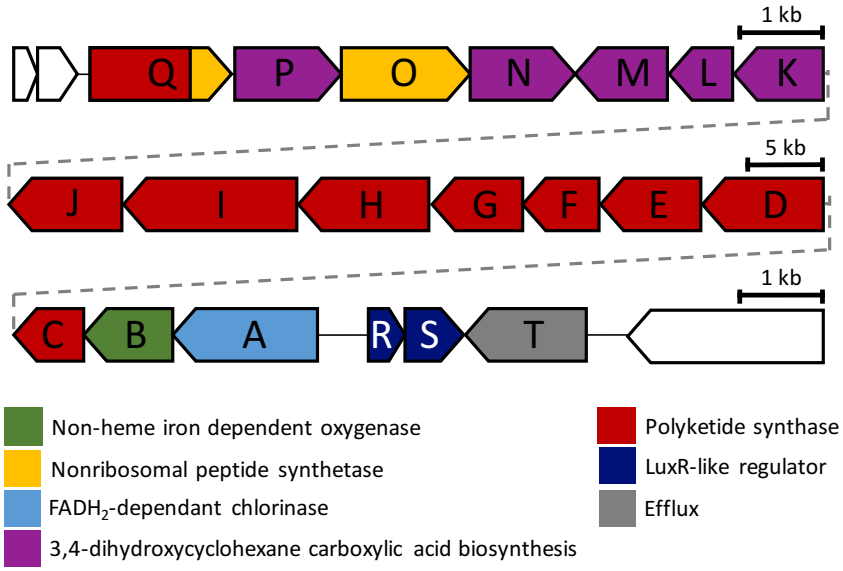

Figure 4.7: Vibroxin biosynthetic gene cluster.

4.5.2 Comparison with enacyloxin biosynthetic gene cluster

In the enacyloxin biosynthetic gene cluster, there are three genes that encode tailoring enzymes (*bamb_5930-bamb_5932*), and a LuxR-like regulation gene (*bamb_5910*), which are not present within the vibroxin biosynthetic gene cluster (Fig. 4.8). All of these genes are located at the periphery of the enacyloxin biosynthetic gene cluster. Furthermore, regulatory gene *vbxR* is located at the opposite end of the gene cluster relative to the location of the homologous gene *bamb_5911* within the enacyloxin biosynthetic gene cluster. The genes *vbxL-vbxQ* (homologous to

bamb_5917-bamb_5912), appear to have undergone an inversion event (or multiple equivalent processes), whereby *bamb_5917* has become truncated compared with *vbxQ* (Fig. 4.8). Rigorous inspection of the domains encoded by gene *vbxQ* reveals four structural components: a region with a relatively low homology to the C-terminal part of a KS domain (presumed inactive), a KS-to-AT linker domain (or a so-called *trans*-AT docking domain), a PCP domain, and a COM_{2C} motif. The KS-like region is absent in Bamb_5917.

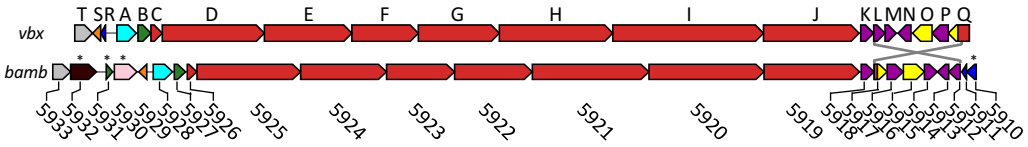

Figure 4.8: Comparison of the vibroxin biosynthetic gene cluster (*vbx*) and the enacyloxin biosynthetic gene cluster (*bamb*). Asterisks denote genes that have no corresponding homologous gene within the vibroxin biosynthetic gene cluster.

The PKS gene that is most different to its enacyloxin gene homologue is *vbxI*, appearing 31% longer in sequence than *bamb_5920*. This is, in part, due to the presence of an additional KR and ACP domain. Despite the absence of this KR domain in Bamb_5920, a reduction at the β-position of the ACP-bound thioester intermediate is still predicted to occur at this stage during enacyloxin biosynthesis. Within VbxI, an additional KS-AT linker domain exists, as well as a region with low homology to an AT domain (not identified using AT HMMs) between the N-terminal KS and the subsequent DH domain (Fig. 4.9). This linker, and the low homology AT region, appear to have been excised in Bamb_5920, as the KS and DH appear concatenated. Typically, malonylation of a functional ACP domain is associated with the presence of either, an AT domain within the same module, or a *trans*-AT docking domain. Within Bamb_5920, neither of these features can be identified within the N-terminal module, and there is no space for any unidentified domains downstream of the KS or upstream of the DH domain. Yet, given the structure of enacyloxin, this KS is presumed to possess elongation activity. Whilst AT-like fragments are found in VbxE, VbxI and VbxJ, they are

assumed to arise from evolutionary excision, and predicted to be catalytically inactive. Presumably, the AT activity of domains in these modules was replaced by another AT at some evolutionary time point, leaving little selective pressure to maintain *cis*-acting ATs at such positions.

The vibroxin and enacyloxin biosynthetic assembly lines consist of both *trans*-AT and *cis*-AT PKS components (Fig. 4.9). The differences between these types of PKS components can be determined by phylogeny of KS and AT domains. *Trans*-AT PKSs typically employ an AT domain acting in *trans*, and *cis*-AT PKSs use a downstream AT domain within the same multi-enzymatic protein, but this does not always hold true.[161] Three motifs can be identified downstream of a *trans*-AT KS: two motifs corresponding to a KS-AT linker denoted KAn and KAc, and a C-terminal KS motif denoted KSc (n.b. shorter motifs are more challenging to identify with HMMs and therefore often give false negative results, effectively failing to identify them). Together these three motifs form a *trans*-AT docking domain (although the exact function of this domain is inconclusive) with a region of low conservation between the KS-AT linker motifs. This region of low sequence conservation has also been proposed to interact with other *trans*-AT docking domains.[162] Structural representations of the differences between *trans*-AT and *cis*-AT PKS domains are shown in Fig. 4.9. The only (phylogenetic) *trans*-AT part of each biosynthetic cluster is the C-terminal region of VbxJ and Bamb_5919. All other KS domains belong to the *cis*-AT clade, even those without adjacent AT domains. Downstream of some of these KS domains, AT fragments can be identified using AT HMMs (labelled with asterisks in Fig. 4.9). These AT fragments are presumed to be non-functional in terms of malonylation because they lack the catalytic amino acid residues within conserved motifs. Instead, these AT-homologous fragments may result from partial AT excision, as a transition towards the domain architecture exhibited in *trans*-AT PKSs, despite a different phylogenetic origin based on the definition employed using KS clades. Whilst some AT fragments are found in vibroxin PKS genes and not within enacyloxin

PKS genes, there is often unannotated space to accommodate fragments. These regions may have decreased AT-homology, preventing their identification as AT fragments using HMMs, further supporting the proposal of evolutionary AT loss.

4.5.3 Proposed biosynthesis of vibroxin

A proposed model for vibroxin biosynthesis is outlined in Fig. 4.10. The loading module within VbxD consists of an 'acyl-recruitment domain',[163] an MT domain and a GNAT domain, followed by an ACP domain. Acyl-recruitment domains are typically found directly upstream of carbon-MT domains within PKS and NRPS enzymes (*i.e.* *cis*-acting carbon-MTs). MT domains act on malonyl-CP, amino-acid-CP or other biosynthetic extender units. The acyl-recruitment domain is considered a prefix to the MT domain and is referred to as dMT due to its dimerisation property observed in these types of structures.[164] The GNAT domain is predicted to load the ACP domain, within the loading module, with malonate. This event is presumably followed by di-methylation of malonyl-ACP by the MT, and subsequent decarboxylation. Another possibility is that the di-methylation occurs after decarboxylation, using acetyl-ACP as the substrate. Interestingly, although the loading module in the enacyloxin biosynthetic assembly line is similar in domain architecture, only a single methylation occurs.

MT and GNAT domains within loading modules have also been described in the biosynthetic assembly lines of saxitoxin,[165] gephyronic acid[166] and apratoxin,[167] which methylate once, twice, and three times, respectively. The loading module from apratoxin contains two distinct MT domains which perform three methylations. This example demonstrates that decarboxylation occurs, at least, before the third methylation takes place. Given this information, and the observation that the active site for malonylation and decarboxylation are proximal within a GNAT, it appears most credible that decarboxylation occurs before methylation within loading modules.

Two *cis*-AT domains are found within the vibroxin gene cluster in VbxD and

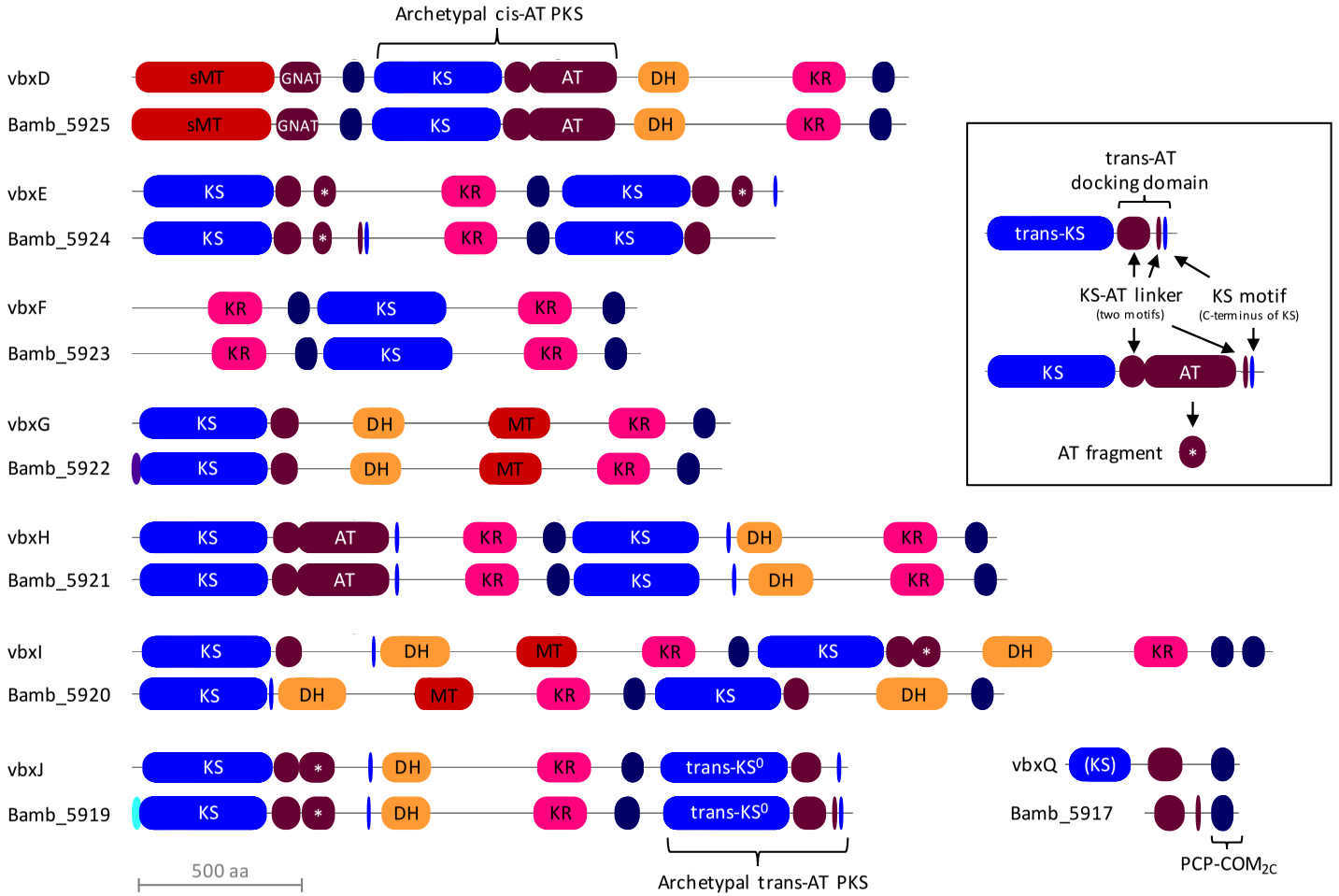

Figure 4.9: Comparison of domain annotations for enacyloxin and vibroxin PKS proteins, displaying domain excision (*e.g.* Bamb_5920), truncation (*e.g.* Bamb_5917), and AT fragments (labelled with asterisks.

VbxH. These are predicted to possess substrate specificity for malonyl-CoA based on NRPS/PKS substrate predictor[14] and Minowa[168] analysis. Each PKS module within the vibroxin biosynthetic pathway is predicted to require malonyl-CoA for chain extension but there are no other AT domains found within the gene cluster. Therefore, either the AT domain from VbxD, VbxH, or an AT from outside of the cluster (or any combination thereof) is predicted to load malonate onto the ACP domains within modules where *cis*-AT domains are not observed.

The stereospecificity of the KR domains was predicted based upon 'fingerprint' rules[16] regarding properties of five amino acids within a particular loop and motif near the catalytic active site (Table 4.2). There are six DH domains within the biosynthetic gene cluster but seven dehydrations are predicted to occur. The DH-less modules within VbxE and VbxH could potentially utilise DH domains from the adjacent modules. Alternatively, modules containing DH domains could act iteratively, while DH-less modules may be skipped. However, the possibility of module skipping is not consistent with stereochemical predictions. The DH domain within VbxG is predicted not to dehydrate the *S*-configured β-hydroxyl of the enzyme-bound intermediate within this module (similar to Bamb_5922).

The C-terminal *trans*-acylase (TA) domain from VbxJ contains the conserved cysteine residue for *trans*-acylase activity but lacks the histidine residue required for chain elongation, forming an AANGA sequence instead of the typical HGTGT active motif. The KS-like fragment from VbxQ is predicted to be catalytically inactive, since it lacks both motifs for activity. Each carrier protein possesses the conserved serine and is therefore presumed to be catalytically active, including the tandem ACP domains in VbxI.

Chain release is accomplished by a NRPS-like C domain with a COM_{2N} domain, VbxO, and involves intermolecular condensation to a shikimate-derived product. The proposed biosynthetic pathway for the shikimate-derived substrate involves VbxL, VbxM, VbxN and VbxP, but it remains unclear which of these enzymes act before or after the condensation reaction. Two tailoring enzymes are

Table 4.2: KR stereochemistry prediction using diagnostic amino acids that were identified from sequence alignments with KR domains of known stereospecificity.

gene	fingerprint regions						KR assignment	predicted stereochemistry
	1	2	3	4	5	6		
VbxD_KR1	VED	L	Q	Y	T	N	B1	R
VbxE_KR1	LAD	V	Q	Y	S	N	B1	R
VbxF_KR1	SQQ	W	L	Y	M	N	A1	S
VbxF_KR2	LEQ	L	Q	Y	F	N	B1	R
VbxG_KR1	GGD	W	K	Y	F	N	A2	$(2S, 3S)$
VbxH_KR1	LSD	L	Q	Y	S	N	B1	R
VbxH_KR2	LSD	L	Q	Y	A	N	B1	R
VbxI_KR1	LDD	M	Q	H	A	N	B1	$(2R, 3R)$
VbxI_KR2	SAN	G	Q	Y	V	N	A1	S
VbxJ_KR1	LAD	L	Q	Y	A	N	B1	R

proposed to be catalytically active; either during PKS assembly, or after chain release. VbxA is a $FADH_2$-dependent chlorinase that is proposed to chlorinate C-11, and VbxB is a non-heme iron-dependent oxygenase that is proposed to hydroxylate C-14.

4.6 Vibroxin loading module

4.6.1 HMM for starter MT

One remarkable difference between vibroxin and enacyloxin is the methyl incorporation event in the loading module of each biosynthetic pathway. In both pathways, the loading modules are poorly annotated with current methods. HMMs for MT domains that reliably identify other MTs within PKSs and NRPSs, are ineffective at identifying the MT domain in VbxD (and Bamb_5925). Similarly, other characterised examples of PKSs utilising an MT within a loading module cannot be identified using the typical MT HMMs. Therefore, the amino acid sequences corresponding to the dimerisation element (dMT) and the associated MT upstream of the GNAT domain were used to build a novel HMM for N-terminal starter MT domains (sMT). This was subsequently used to identify and annotate examples of sMT-encoding genes within enacyloxin, saxitoxin, gephy-

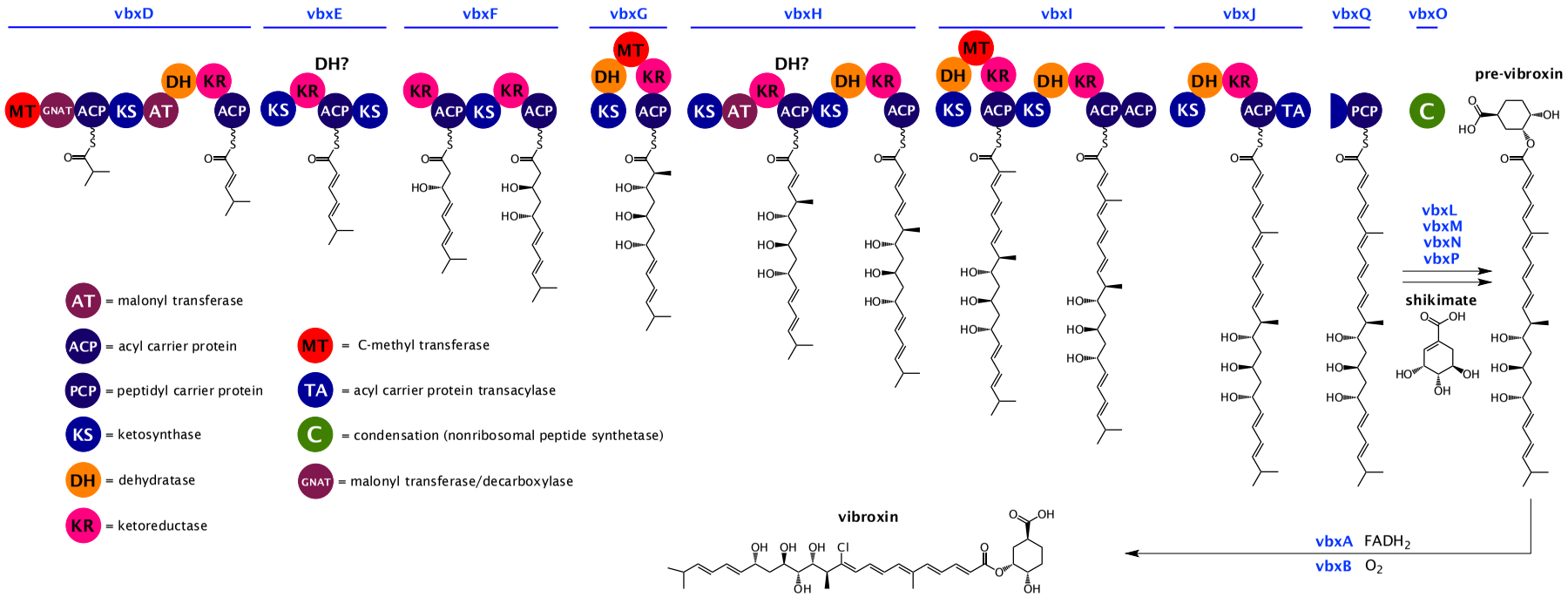

Figure 4.10: Proposed vibroxin biosynthetic pathway.

ronic acid and vibroxin gene clusters, as well as cryptic biosynthetic pathways from a diverse range of organisms (Table. 4.3). These results exemplify the wider occurrence of MT domains within PKS loading modules.

Table 4.3: Example of the diversity of domain organisation within degenerate examples of sMT-containing proteins.

ID	**Domains in gene**
WP_061546530.1	sMT-GNAT-ACP-Amino_t
WP_061581925.1	sMT-GNAT-ACP-KS-KR-ACP-KS-DH-KS-KR-ACP-KS-DH-ACP-KS-ACP
ABI91470.1	sMT-GNAT-ACP-KS-AT-DH-KR-ACP
ABK51302.1	sMT-GNAT-MT-ACP-KS-DH-ACP-KS-ACP-GNAT-TE-KS
EEP93156.1	sMT-GNAT-KS-DH-KR-ACP
ADY62524.1	sMT-GNAT-ACP
CCF16272.1	sMT-ACP-KS-DH-KR
WP_011553948.1	sMT-GNAT-ACP-KS-KR-ACP
WP_017307890.1	sMT-GNAT-ACP-KS-AT-ACP
ERO59981.1	sMT-GNAT-ACP-KS-ACP-KS-DH-KR
WP_023131069.1	sMT-GNAT
AHA38199.1	sMT-GNAT-ACP-KS-AT-DH-MT-KR-ACP-KS-DH-MT-KR-ACP
EYF04556.1	sMT-GNAT-ACP-KS-DH-KR-ACP-KS-DH-KR-ACP-C-A-PCP
WP_027413074.1	sMT-ACP-ACP-KS-DH
WP_035684581.1	sMT-GNAT-ACP-KS-DH-KR-MT-ACP-KS-ACP-KS
WP_044619203.1	sMT-GNAT?-MT-ACP-KS-MT-ACP-ACP-KS-DH-ACP-KS-KR-MT-ACP-KS-MT-ACP-KS-MT-ACP-ACP-KS-KR-ACP-KS
WP_049974262.1	sMT-GNAT-ACP-KS-AT
WP_051908584.1	sMT-GNAT-ACP-ACP-KS-MT-ACP-TE
WP_052421794.1	sMT-GNAT-ACP-KS-ACP-ACP
KVD87326.1	sMT-GNAT-ACP-C-A-PCP-?-A-AT
XP_004991702.1	sMT-GNAT-ACP-KS-?-ER-KR-ACP-KS-AT-KR-ACP-KS-KR-ACP-KS-KR-ACP-KS?
KYG81075.1	sMT
WP_060839246.1	sMT-GNAT-ACP-KS-ACP-Cy-Cy
WP_010587417.1	sMT-GNAT-ACP-KS-AT-DH-KR-ACP

Developing an HMM to describe sMT domains provides an efficient technique for the identification of sMT domains from previously characterised biosynthetic pathways (Fig. 4.11). Examples of natural products putatively utilising active sMT domains are shown in Fig. 4.12. The basiliskamide PKS from *Bacillus lat-*

erosporus (isolated from the ear of a feral hog)[169] was identified as containing a full sMT domain within the loading module, yet the final compound suggests that an acetyl loading unit is incorporated. Presumably, the sMT is non-functional within this biosynthetic pathway. However, a closely related analogue, YM-47522 from *Bacillus* sp. YL-03709B (soil isolate),[170] is built from a propionyl starter unit. Unfortunately, the genomic DNA sequence for YM-47522 biosynthetic gene cluster is not currently available. Nevertheless, the presence of the additional methyl group in YM-47522 is most likely derived from a functional form of the sMT domain within its loading module. Whilst a nonfunctional sMT domain is still present in the basiliskamide PKS, this may be an early step along the evolutionary pathway to the full excision of this domain. If such processes are genuine, uncovering biosynthetic pathways that display different stages of evolutionary loss of sMT can act as 'missing links' to support this hypothesis. Whilst the reverse process is possible, gaining sMT domains or sMT functionality poses a significantly greater entropic expense.

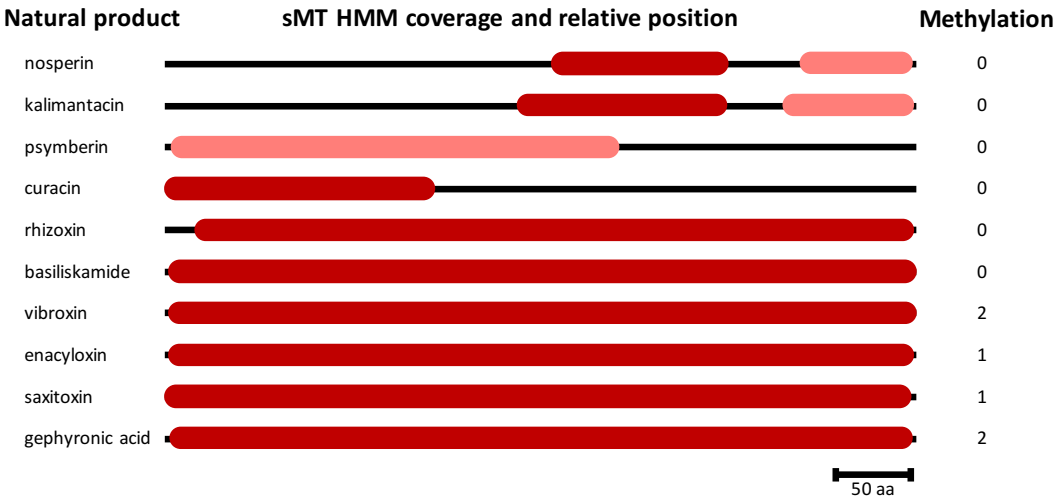

Figure 4.11: Examples of biosynthetic pathways with characterised natural products, displaying sMT coverage, and number of methylations from sMT.

The rhizoxin PKS[171] contains an sMT domain within the loading module of its biosynthetic pathway, yet no methylation activity is predicted to occur at this stage. It is possible that a rhizoxin analogue exists resulting from a biochemically active sMT domain, comparable to YM-47522 and basiliskamide.

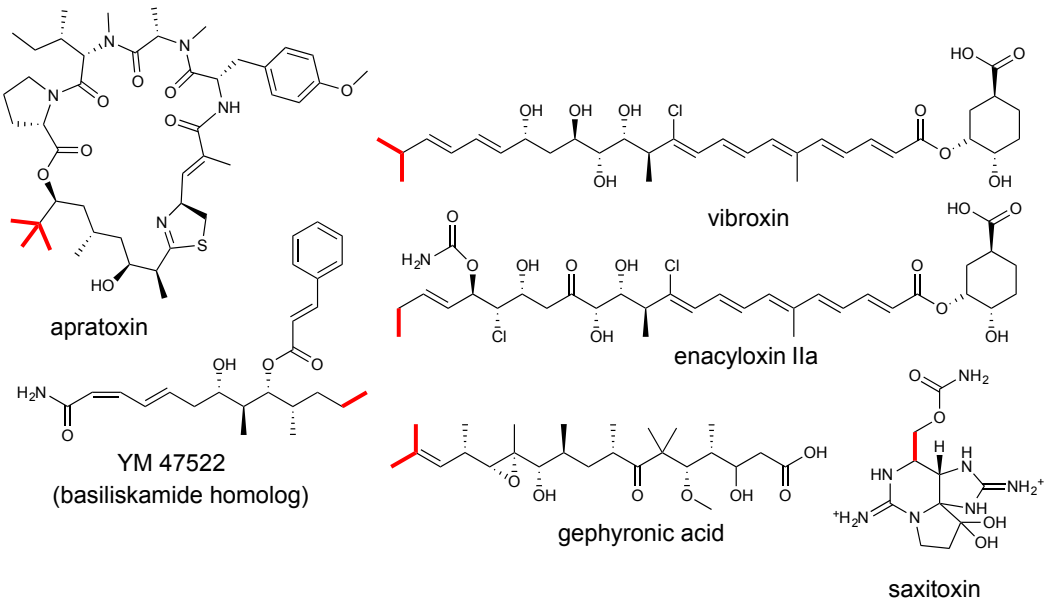

Figure 4.12: Natural products utilising sMT domains. Methylations predicted to take place from MT domains within loading modules are displayed in red.

The sMT HMM also identified a slightly smaller sMT domain, with a relatively low scoring homology, in the psymberin PKS. This is also presumed to be inactive due to no methylation observed at the corresponding position in the molecular structure of psymberin.

There are two characterised biosynthetic clusters which provide examples of sMT excision during the course of biosynthetic evolution. The kalimantacin[172] and nosperin[173] PKSs both contain similar fragmented sMT domains within their loading modules. Their N-termini contain a high scoring sMT fragment domain, followed by a low scoring fragment of sMT. In both cases, the N-terminal region of the sMT domain is not identified, and the C-terminal fragment exhibits very low sequence conservation. No methylations take place during the biosynthesis of these compounds. An example of a more truncated sMT domain can be found within the curacin PKS, where the sMT domain predominantly consists of the dMT subdomain and a minor fragment thereafter. This is also deduced to be nonfunctional from the structure of curacin.

The sMT HMM has assisted with the identification of methyltransferase activity within loading modules of different PKSs, and, as a result, offers a common

explanation for propionyl and isobutyryl starter units found within PKSs with a GNAT loading strategy. As with other PKS enzymes, sMT domains can appear functional yet may be inactive, and inspection of motifs around putative SAM-binding regions will offer more insight. Care must also be taken when identifying sMT fragments which are likely to be non-functional. However, these offer a valuable perspective on the evolution of starter units within biosynthetic pathways, and the potential for historic or unidentified analogues.

4.6.2 Feeding of (^{13}C-Me)-L-methionine

The sMT domain that was identified in VbxD, from the vibroxin biosynthetic pathway, is proposed to perform two methylations within the loading module. This dimethylation activity was probed by feeding (^{13}C-Me)-L-methionine to *V. rhizosphaerae*, to evaluate the number of ^{13}C-methyl groups incorporated into vibroxin (Fig. 4.13a). The labelled methionine is metabolised by *V. rhizosphaerae* to form *S*-adenosyl-methionine (SAM), the cofactor for sMT domains. Subsequently, ^{13}C-methyl groups are incorporated during vibroxin biosynthesis for each SAM-dependent methylation. Three MT domains are present in the vibroxin biosynthetic pathway, within VbxG, VbxI, and VbxD, which is proposed to act twice. Therefore, the incorporation of four ^{13}C atoms is predicted during vibroxin biosynthesis, when feeding *V. rhizosphaerae* with (^{13}C-Me)-L-methionine.

Ethyl extracts were performed from *V. rhizosphaerae* fed with natural abundance methionine (Fig. 4.13bi) and (^{13}C-Me)-L-methionine (Fig. 4.13bii). Incorporation of the isotope precursor was observed at 90% for all of the proposed SAM-derived methyl groups within vibroxin. An estimated 73% of vibroxin molecules are shown to contain four ^{13}C atoms, 20% contained three ^{13}C atoms, and less than 7% contained fewer than three ^{13}C atoms. Together, this ensemble describes the isotopic MS distribution observed in Fig. 4.13bii. The incorporation of four labelled methyl groups supports the proposed dimethylation of the starter unit by the sMT domain within VbxD.

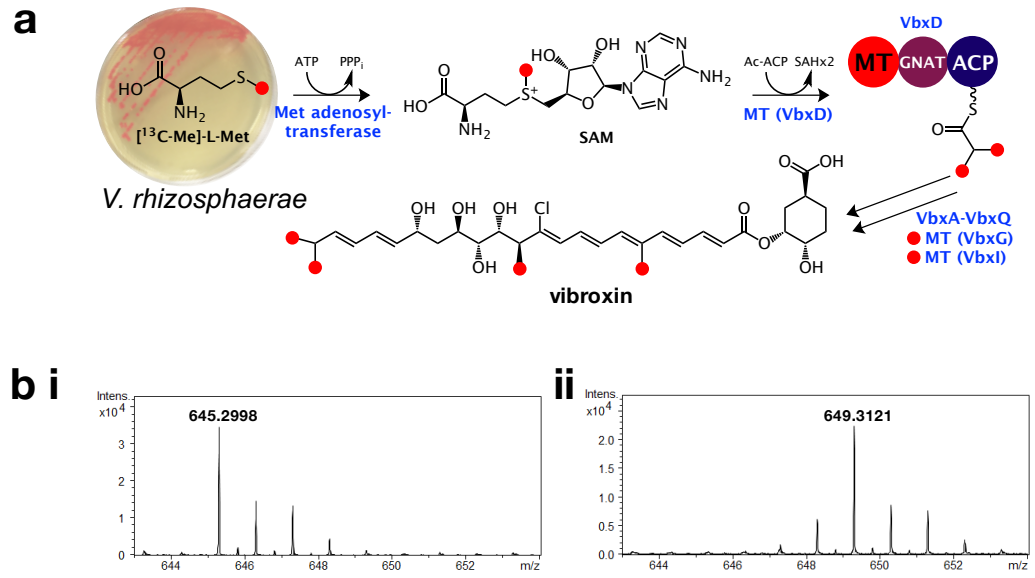

Figure 4.13: a) Schematic of isotopic incorporation of ^{13}C from (^{13}C-Me)-L-methionine into vibroxin, where the sites of incorporation are shown in red circles. b) Isotopic MS distribution for vibroxin from *V. rhizosphaerae* fed with i) natural abundance methionine, or ii) (^{13}C-Me)-L-methionine.

4.7 Antimicrobial activity

Vibroxin possesses a similar molecular structure to enacyloxin IIa, a well-known antibiotic that is active against *A. baumannii* and inhibits elongation factor Tu (EF-Tu), an essential component of the bacterial protein synthesis machinery. *A. baumannii* is an opportunistic human pathogen that is highly resistant to most antibiotics in clinical use and causes hospital-acquired infections in people with compromised immune systems.[3] Current antibiotic treatments for *A. baumannii* are progressively becoming ineffective as resistance develops. The resistance rates for antibiotics that are no longer viable as treatment options are listed in Table 4.4.

Table 4.4: Resistance rates of *A. baumannii* to antibiotics.[174]

Antibiotic	Resistance rate (%)
ciprofloxacin	32-100
cefepime	91-100
piperacillin-tazobactam	90-92
amikacin	24-94
gentamicin	18-85

The current solution is to use colistin in combination with rifampicin, sulbactam and carbapenems. However, colistin is nephrotoxic and resistance to it is reported world-wide (although currently only at relatively low levels (2%) in *A. baumannii*).[174] The development of novel antibiotics with potent activity against *A. baumannii*, such as enacyloxin analogues, is necessary to combat nosocomial infections. To verify whether vibroxin has an analogous antimicrobial spectrum of activity to enacyloxin IIa, a series of antimicrobial assays were designed with *V. rhizosphaerae* and purified vibroxin.

4.7.1 Whole cell activity

Activity of *V. rhizosphaerae* was tested using overlay experiments by monitoring the size of inhibition zones arising from growth of a spotted colony of *V. rhizosphaerae* on a lawn of pathogenic bacteria. *V. rhizosphaerae* was shown to exhibit broad spectrum activity against a range of organisms including Gram-positive and Gram-negative bacteria (Fig. 4.14). The most potent activity was observed against *B. multivorans* and *A. baumannii*. This whole cell activity assay tests all compounds produced by *V. rhizosphaerae*, including the bioactive prodigiosin. To assess the potency of vibroxin alone, assays were also performed with purified vibroxin.

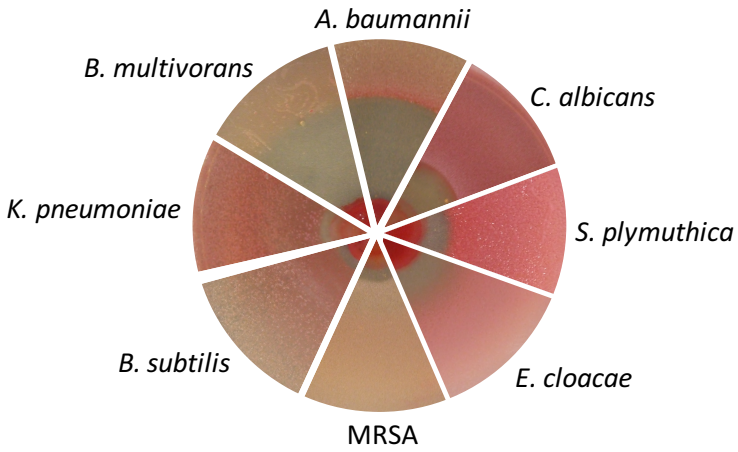

Figure 4.14: Overlay antimicrobial assays with *V. rhizosphaerae* and pathogenic bacteria.

4.7.2 Vibroxin MIC

The quantitative determination of antimicrobial activity is necessary to evaluate antibiotic potency across different pathogenic strains. One method of clinical significance is the minimum inhibitory concentration (MIC) assay using the Clinical Laboratory and Standards Institute (CLSI) broth micro-dilution method. The lowest concentration that inhibits bacterial growth is visually determined for vibroxin following an overnight incubation. The MIC for vibroxin was determined against six representative pathogens with growing multi-drug resistant virulence, *i.e.* the 'ESKAPE' panel (Table 4.5). Vibroxin was shown to demonstrate potent activity against *A. baumannii* with a MIC of 2 μg/mL (*cf.* 2 μg/mL for enacyloxin IIa). The growth of all other tested pathogens was not inhibited by vibroxin below the maximum concentration tested. Whilst the MIC value relates to growth inhibition, the minimal bactericidal concentration (MBC) represents the lowest concentration of a bioactive compound required to kill a particular pathogen. The MBC was determined to be 8 μg/mL for vibroxin against *A. baumannii* which lies on the boundary of what is considered bactericidal or bacteriostatic.

Table 4.5: MIC and MBC values for vibroxin determined against the ESKAPE panel of pathogens.

Organism	**MIC (μg/mL)**	**MBC (μg/mL)**
Enterococcus faecium	>32	>32
Staphylococcus aureus	>32	>32
Klebsiella pneumoniae	>32	>32
Acinetobacter baumannii	2	8
Pseudomonas aeruginosa	>32	>32
Enterobacter cloacae	>32	>32

4.8 Conclusion and further work

Genome mining of *V. rhizosphaerae* led to the targeted discovery of the antibiotic vibroxin, revealing the value of rhizosphere-associated *Vibrio* as a source of natural products. The vibroxin biosynthetic gene cluster reveals insights into the

evolution of the homologous enacyloxin biosynthetic gene cluster. In particular, how genetic inversion events and excisions have left relics in terms of domain fragments. A number of differences between the enacyloxin and vibroxin biosynthetic pathways have not yet been fully explained. Whilst VbxJ contains an additional KR domain, compared to Bamb_5920, both biosynthetic pathways display KR function at this stage. The origin of the KR enzymatic function in enacyloxin biosynthesis therefore remains unknown. A dehydration event is predicted to take place in the second module of the vibroxin PKS, yet no DH domain is present within VbxE. The underlying mechanism of this difference between vibroxin and enacyloxin is still unclear.

The development of an HMM to identify sMT domains assists with the annotation of any PKS gene utilising this type of domain. This HMM can provide valuable insight when predicting core molecular structures of putative polyketides, and has uncovered indications of evolutionary development of loading modules within PKSs. The basis for methylation control by sMT domains has not been elucidated in terms of predicting the number of methylations of a starter unit, but provides an opportunity for engineering starter units within biosynthetic pathways utilising a GNAT-based loading strategy.

The potent antimicrobial activity of vibroxin against the opportunistic pathogen *A. baumannii* highlights the value of targeted drug discovery using a genomics-driven approach. Vibroxin is potentially more exploitable for semi-synthesis or total chemical synthesis due to a simplified molecular structure compared with enacyloxin, and is chemically more stable. Novel molecular features or fragments could be utilised as a reference on which to design more potent, stable vibroxin analogues via biosynthesis or chemical synthesis. The stereochemistry of the hydroxyls in the polyketide chain, and the proximal methyl (C-12 to C-17), is proposed based upon KR stereospecificity, but experimentally determined stereochemistry is required, which may be achieved from NMR spectroscopy of vibroxin derivatives or comparison with chemically synthesized standards.

Chapter 5

Materials and Methods

5.1 Materials

5.1.1 Plasmids

Constructs for the expression of domains from the enaclyoxin biosynthetic pathway were provided by Dr P Sydor (University of Warwick), and those encoding watasemycin NRPS domains were provided by S Zhou (University of Warwick), within the pET151 plasmid; these are listed in Table 5.1. The plasmid containing *sfp* was acquired from Manuela Tosin (University of Warwick) in pET28b.

Table 5.1: Plasmids for protein production.

Protein domain names	**Gene name**	**Length (bp)**
PCP-COM_{2C}	*bamb_5917*	405
ACP	*bamb_5919*	660
COM_{2N}-C	*bamb_5915*	1638
KS^0	*bamb_5919*	2156
PCP-COM_{2C}	*sven_0512*	453
Sfp	*sfp B. subtilis*	-

5.1.2 Microbial Strains

High transformation efficiency, and plasmid propagation, was achieved using Top10 *Escherichia coli* (Invitrogen). For protein production, *E. coli* strains

BL21 (DE3) (Invitrogen) and C43 (DE3) (Invitrogen) were used. *Vibrio rhizosphaerae* MSSRF3 (DSM 18581) was used for vibroxin production. Bioactivity assays using *V. rhizosphaerae* or vibroxin were performed using the following organisms:

- *Bacillus subtilis*
- Methicillin-resistant *Staphylococcus aureus* DSM 21979
- *Staphylococcus aureus* NCTC 12981
- *Escherichia coli* SY397
- *Serratia plymuthica* RVH1
- *Pseudomonas aeruginosa* DSM 29239
- *Burkholderia multivorans* ATCC 17616
- *Klebsiella pneumoniae* DSM 26371
- *Enterobacter cloacae* DSM 16690
- *Acinetobacter baumannii* DSM 25645
- *Candida albicans* SC 5314
- *Enterococcus faecium* DSM 25390

5.1.3 Culture media

Lysogeny broth (LB) (Fisher), tryptic soy broth (TSB) (Fisher) and basal salt medium (BSM),[175] were made following the suppliers recommended dilutions. Glycerol or glucose was used as a carbon source for BSM media to a final concentration of 4 g/L.

M9 minimal media

5 × M9 salt solution

- 32 g $Na_2HPO_4.7H_2O$
- 6 g KH_2PO_4
- 2.5 g NaCl
- Add H_2O up to 500 mL and autoclave

M9 minimal media

- 200 mL 5 × M9 salt solution
- 2 mL of 1 M $MgSO_4$
- 100 μL of 1 M $CaCl_2$
- Add H_2O up to 1 L and autoclave
- 2 g of D-^{13}C-glucose (or 4 g of D-glucose) in 20 mL H_2O with 0.2 μm filtration
- 20 mL of 100 × BME vitamins (Sigma Aldrich)
- 0.5 g of $^{15}NH_4Cl$ in 2.5 mL H_2O with 0.2 μm filtration
- Ampicillin (final concentration 50 μg/mL)

5.2 Molecular Biology

5.2.1 Site-directed mutagenesis

For single, double and truncation mutations, the Q5 site-directed mutagenesis kit (NEB) was used. Primers were designed using the associated NEBaseChanger

tool, where truncation mutants were achieved through substitution mutants using the TGA stop codon. Annealing temperatures, PCR protocols, Kinase-Ligase-DpnI (KLD) reactions, and transformations were conducted according to suppliers guidelines. Typically, three colonies were selected for plasmid isolation using the GeneJET Plasmid Miniprep Kit (Thermo Scientific). DNA concentration was measured by absorbance at 280 nm. Resulting plasmids were confirmed by DNA sequencing. The primers used for site directed mutagenesis were: PCP_δ9_f = GGCGCCTTCGGGCGACCAGGACGATGCATGAGAGTACGAAGAAGGCGTGATCC, PCP_δ9_r = CCGCGGAAGCCCGCTGGTCCTGCTACGTACTCTCATGCTTCTTCCGCACTAGG.

5.2.2 Sequencing

Sanger sequencing (GATC-biotech) was used to determine successful mutations of genes within the pET151 plasmid. Samples were prepared according to the company's requirements. Sequencing primers were T7 forward primer and pET-RP reverse primer.

5.2.3 Preparation of chemically competent *E. coli*

A glycerol stock of *E. coli* (*e.g.* Top10, BL21*, C43) was used to inoculate 10 mL of LB (Lysogeny broth) and grown overnight at 37 °C at 180 rpm. 1 mL of preculture was used to innoculate 100 mL of LB. This was grown at 37 °C, 180 rpm until $OD_{600\ nm}$=0.6. Cells were recovered by centrifugation at 2057 *g* for 7 minutes at 4 °C. The pellet was resuspended by gentle mixing in 20 mL of chemically competent *E. coli* washing buffer (80 mM MgCl, 20 mM $CaCl_2$). Cells were then centrifuged as previously described. The supernatant was decanted and the pellet was resuspended in 1 mL of cold 0.1 M $CaCl_2$. For each transformation 50 μL of cells were used. For long term storage cells were flash frozen and stored at -80 °C.

5.2.4 Transformation of chemically competent *E. coli*

For each transformation, 1 μL of the DNA construct (100 ng/μL) was gently mixed with 50 μL of chemically competent *E. coli* on ice. The mixture was incubated on ice for 30 minutes, transferred to a 42 °C water bath for 90 seconds, then left to rest on ice for 2 minutes. 200 μL of pre-warmed LB or super optimal broth with catabolite repression (SOC) was added to the cells and they were incubated at 37 °C, 180 rpm for 1 hour. Transformants were selected by plating the culture on LB agar plates with a relevant antibiotic. Plates were incubated at 37 °C overnight.

5.3 Protein production

5.3.1 Recombinant protein production

A colony from a successful transformation was isolated from a plate to inoculate 5 mL of LB and a relevant antibiotic (at final concentrations of 100 μg/mL ampicillin or 50 μg/mL kanamycin). The 5 mL culture was grown overnight at 37 °C, 180 rpm. The overnight culture was used to inoculate a flask of LB (typically 1 L) supplemented with an antibiotic selection marker. The culture was grown at 37 °C at 180 rpm until $OD_{600\ nm}$=0.6, then cooled to 4 °C. A sterile solution of isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 0.5 mM, and the culture was left at 15 °C, 180 rpm overnight.

5.3.2 Recombinant protein production - isotope labelling

A 1 L culture of cells was grown in LB supplemented with an antibiotic until $OD_{600\ nm}$=0.6. This culture was inoculated from an overnight culture as previously described. Cells were harvested by centrifugation (6080 g, 20 minutes, 4 °C), washed once with 200 mL of M9 sterile salt solution and centrifuged (6080 g, 20 minutes, 4 °C) thereafter. The pellet was resuspended in 1 L of M9 minimal

media with ^{15}N labelled $^{15}NH_4Cl$ as the sole nitrogen source, D-^{13}C-glucose as the only carbon source (or natural abundance glucose), supplemented with an antibiotic and IPTG (0.5 mM). The culture was incubated overnight at 20 °C, 180 rpm.

5.3.3 Purification of recombinant histidine-tagged proteins

Cells from the overnight culture (natural abundance or isotopically labelled) were centrifuged at 6080 *g* for 20 minutes. The pellet was resuspended in 10 mL of wash buffer (20 mM Tris-HCl pH 8, 100 mM NaCl, 10% glycerol, 20 mM imidazole) and phenylmethylsulfonyl fluoride (PMSF) per 1 L culture. Cell lysis was performed using a cell disruptor in one shot mode at 20 kpsi. The lysate was centrifuged at 38720 *g* for 20 minutes at 4 °C. The supernatant was filtered (0.22 μm) and purified using an Äkta purifier and a 1 mL (or 5 mL) HiTrap HP affinity column (Nickel Sepharose High Performance, GE) at 4 °C. The column was equilibrated with washing buffer. Unbound proteins were removed using washing buffer, and His-tagged proteins were eluted with a gradient application of elution buffer (20 mM Tris-HCl pH 8, 100 mM NaCl, 10% glycerol, 300 mM imidazole). An additional purification step of gel filtration chromatography was carried out when necessary using either a Superdex 200 10/300 or Superdex 75 16/600 columns, with a gel filtration buffer (20 mM Tris-HCl pH 8, 100 mM NaCl). Fractions were collected and concentrated using Amicon Ultra filters with a 3,000 10,000 or 30,000 kDa molecular weight cutoff membrane (Millipore) in a concentration buffer (20 mM Tris-HCl pH 8, 100 mM NaCl, 10% glycerol). Any subsequent exchanging of buffer was performed using PD-10 desalting columns (GE) using the gravity method or HiTrap desalting columns (GE) with an Äkta purifier. Proteins were analysed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE),[176] and were stored at -80 °C. The protein standards ladder used for SDS-PAGE at with the following molecular weights in kDa: 10 (green), 15, 25 (red), 35, 55, 70 (red), 100, 130 and 250. Protein concentration

was calculation by absorbance at 280 nm using predicted extiction coefficients.[177]

5.4 Protein NMR spectroscopy

5.4.1 Sample preparation

Concentrated proteins were exchanged from the concentration buffer to 50 mM potassium phosphate, 150 mM NaCl pH 6.5 (supplemented with 50 mM Arg, 50 mM Glu for assignment sample) via a PD-10 desalting column (GE). After buffer exchange, the proteins were concentrated as previously described. For locking, 5% D_2O was added, and 40 μM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) was included.

5.4.2 PCP assignment

NMR experiments were acquired at 288 K on Bruker Avance II 700 MHz equipped with 5 mM triple-resonance $^{1}H\{^{13}C/^{15}N\}$, pulse-field gradient carbon optimised cyroprobe. PCP_{17}-COM_{2C} backbone assignment was carried out using the suite of standard Bruker 3D pulse programs (HN(ca)CO, HN(co)CA, CBCA(co)NH). Side-chain resonances were assigned using HCCH-TOCSY/COSY, HBHA(co)NH, H(CCCO)NH and verified using 3D ^{1}H- and ^{13}C-filtered NOESY. Assignment was performed with 500 μM [U-^{13}C,^{15}N]PCP-COM_{2C} in 50 mM phosphate, 50 mM arginine, 50 mM glutamine, 50 mM NaCl, 5% D_2O, 2 mM TCEP, 40 μM DSS at 288 K. Data was processed in Topspin and exported to NMRFAM-Sparky[178] for further analysis. Resonances were picked automatically and manually edited. Predicted assignments were performed using PINE[179] and were confirmed or edited accordingly.

5.4.3 Protein interactions by NMR

For each concentration point during a protein-protein interaction experiment a new sample was made (from the same protein stock) to circumvent dilution effects

and bias during signal acquisition and data processing. Signal intensities were calculated in NMRFAM-Sparky for backbone amides. For PCP and C domain interaction experiments, each sample contained 250 μM of labelled carrier protein, with varying amounts of unlabelled condensation domain in 50 mM potassium phosphate, 150 mM NaCl pH 6.5, 5% D_2O and 40 μM DSS.

5.5 Biochemical assays

5.5.1 Carrier protein acetylation

Apo-carrier proteins (*apo*-CPs) were posttranslationally modified to *holo*-CPs (or acetyl-CPs) using a phosphopantetheinyl transferase (PPTase) Sfp from *B. subtilis* and a coenzyme A (or acetyl-CoA). The reaction was performed using 200 μM CP, 10 μM Sfp, 0.5 mM acetyl-CoA (or CoA for *holo*-CP), 12.5 mM $MgCl_2$, 2 mM TCEP, 100 mM Tris pH 7.5. The reaction was incubated at 30 °C for 90 minutes and monitored by matrix-assisted laser desorption/ionization time-of-flight (MALDI-ToF) or liquid chromatography–mass spectrometry (LC-MS).

5.5.2 Condensation assay

The COM_{2N}-C catalysed condensation of acetyl-PCP-COM_{2C} (or acetyl-PCP) and shikimate, was carried out with 200 μM acetyl-PCP-COM_{2C} incubated with 20 μM COM_{2N}-C and 1.5 mM shikimate overnight at 30 °C. Reactions were stopped by adding two volume equivalents of MeOH to the reaction. Centrifugation and filtration was performed before analysis by MaXis Impact to detect small molecules. The LC elution profile is outlined in Table 5.2 at 0.2 μL/min flow rate using the same column as for natural product extracts.

Table 5.2: LC-MS elution profile for condensation assay product detection.

Time (minutes)	H_2O % (0.1% FA)	Acetonitrile % (0.1% FA)
0	95	5
5.3	95	5
17.3	0	100
22.6	0	100
25.3	95	5
34	95	5

5.6 Vibroxin from *V. rhizosphaerae*

5.6.1 Natural product extraction

Vibrio rhizosphaerae DSM 18581 was acquired from Leibniz Institute DSMZ (German Collection of Microorganisms and Cell Cultures). Growth of *V. rhizosphaerae* was performed in TSB2S (tryptic soy broth supplemented with 2% NaCl) or BSM2S (basal salt medium supplement with 2% NaCl), in liquid and on solid media. For vibroxin production, *V. rhizosphaerae* was grown on solid BSM2S media and incubated at 30 °C for 65 hours. *V. rhizosphaerae* was removed from BSM2S-agar plates and the BSM2S-agar was cut into small pieces. Ethyl acetate was added to cover the BSM2S agar and left to stand for 1 hour before filtering the extract, and concentrating *in vacuo*. The extract was solubilised in acetonitrile for detection by LC-HR/MS. A ZORBAX Eclipse Plus C18 column 1.8 μm, 2.1 $\times$ 100 mm (Agilent) was used following the elution profile from Table 5.3 with a 0.2 μL flow rate. The absorbance was measured at 360 nm corresponding to the unsaturated/diene molecular constituents. Data was internally calibrated using sodium formate and molecular formulae were predicted using DataAnalysis.

5.6.2 Bioactivity using whole cell

A fresh culture of *V. rhizosphaerae*, grown on TSB2S-agar, was isolated with a loop and resuspended in 4 mL of 0.9% NaCl. From this solution, 5 μL was

Table 5.3: LC-MS elution profile for natural product extracts.

Time (minutes)	**H_2O % (0.1% FA)**	**Acetonitrile % (0.1% FA)**
0	70	30
5.3	70	30
17.3	0	100
22.6	0	100
25.3	70	30
34	70	30

spotted in the centre of each TSB2S-agar plate and incubated at 30 °C for 65 hours. Spotted *V. rhizosphaerae* cultures were killed using chloroform vapour. Glycerol stock cultures of *Escherichia coli* SY327, *Burkholderia multivorans* and *Staphylococcus aureus* (add ESKAPE) were used to inoculate 5 mL of TSB media and incubated overnight at 30 °C. From each culture, 15 μL was taken and diluted in 15 mL of LB-soft-agar supplemented with 2 mg of 2,3,5-triphenyltetrazolium chloride (TTC). For each plate, 7.5 mL of this mixture was spread evenly to create a lawn for overnight incubation at 30 °C.

5.6.3 Vibroxin isolation and purification

Large scale solid BMS-agar cultures were performed as previously described for natural product extraction using eighty 15 mL plates. Extraction with ethyl acetate, filtering, and concentration was repeated twice. During and after extraction, light contact was minimised to deter isomerisation of the natural product. The extract was filtered and purified by analytical HPLC with using a ZORBAX-SB C18 column (21.2 × 100 mm, 5μm) following the elution profile from Table 5.4 monitoring the absorbance at 360 nm. Fractions from multiple runs were pooled, concentrated *in vacuo*, and subsequently lyophilized.

5.6.4 Vibroxin characterisation by NMR

Purified vibroxin was dissolved in 150 μL d_4-MeOH whilst minimising light exposure. 1D ^{1}H, 2D ^{1}H-^{1}H COSY, ^{1}H-^{13}C HSQC, ^{1}H-^{13}C HMBC and ^{1}H-^{1}H NOESY

Table 5.4: HPLC elution profile for vibroxin purification.

Time (minutes)	H_2O % (0.1% FA)	Methanol % (0.1% FA)
0	30	70
2	30	70
5	15	85
30	0	100
32	0	100
36	30	70
40	30	70

experiments were recorded using standard procedures at 288 K on a 500 MHz Bruker Avance III HD spectrometer with a DCH cryoprobe. NOESY mixing times were 30 and 100 ms. The obtained spectra are reproduced in the Appendix Fig. 7.15-7.19.

5.6.5 Feeding experiments

Feeding experiments were performed for *V. rhizosphaerae* grown on BMS2S-glycerol-agar (basal salt medium with glycerol as the primary carbon source, supplemented with 2% NaCl) containing 10 mM (^{13}C-methyl)-L-methionine. Each culture was grown on 10 mL of media. A subset of *V. rhizosphaerae* cultures were subjected to feeding of a saturated (^{13}C-methyl)-L-methionine solution, distributed by spotting 5 μL upon the culture after 18 and 42 hours of incubation at 30 °C. After three days, ethyl extracts were performed from solid media using 20 mL of ethyl acetate for each 10 mL culture before concentration in vacuo. Extracts were solubilised in acetonitrile/water 70/30 for analysis by LC-HR/MS. The LC-MS protocol was the same as for other *V. rhizosphaerae* extracts as described in Section 5.6.1.

5.6.6 M.I.C measurement

The ESKAPE panel were grown overnight in Müller-Hinton (MH) media at 30 °C overnight. Each organism was diluted to a final concentration of 5 $\times$ 10^5 colony-

forming units/μL using McFarlands turbidity standards.

5.6.7 M.B.C. determination

The minimum bactericidal concentration (MBC) was determined for vibroxin by subculturing from the MIC assay to LB-agar plates without vibroxin.

5.7 Computational

5.7.1 Hidden Markov models

Hidden markov models (HMMs) for a number of PKS and NRPS domains were acquired from antiSMASH 3.0[180], and others were acquired from Pfam. Novel HMMs were built from manually curated sequences aligned using clustalw2 and developed using HMMER 3.1b2[181]. Searches using HMMs were also carried out using this software.

5.7.2 PKS/NRPS sequence databases

Sequences used for terminal motif identification were scraped from ClusterMine 360 from which a subset of clusters were used to form the basis of a database of PKS/NPRS sequences. Individual clusters of interest were added manually thereafter. As a test data set, sequences were downloaded from an atlas of PKS and NRPS sequences used for testing novel HMMs.

5.7.3 Peptide binding simulation

CABS-Dock[182] was used for peptide-protein interactions for COM_{2C} with COM_{2N}. The C-terminal COM_{2C} domain was evaluated with the isolated COM_{2N} domain as the peptides binding target. The CABS-dock procedure utilizes Replica Exchange Monte Carlo dynamics with 10 replicas uniformly spread on the temperature scale using randomised initial peptide structures.[182]

Chapter 6

Conclusions

6.1 Summary and outcomes

6.1.1 Objective I review

A biophysical basis for chain release in the enacyloxin PKS-NRPS has been described in the form of COM domains, across the PCP and C domain protein boundary. A protein complex model of PCP-COM_{2C} and COM_{2N}-C provides a representation of the intermolecular interaction. This presents a picture of the accessibility of PCP, to receive substrates from the previous domain, and to donate substrates to the C domain, in a recursive procedure, whilst maintaining complex formation with COM domains.

The type 2 COM domain interaction, observed in enacyloxin PKS-NRPS, was extended to many other characterised NRPS and hybrid PKS-NRPS biosynthetic pathways. This was possible due to the development of a HMM to describe COM_{2N} domains, and the curation of a PKS/NRPS database. Investigating characterised and cryptic biosynthetic pathways revealed a degree of variation in the enzymatic domains found adjacently to COM domains. The corresponding COM_{2C} interaction partners displayed conservation in an EExxL motif, which was consistent with interaction sites on PCP-COM_{2C} (within Bamb_5917 from enacyloxin PKS) using protein NMR spectroscopy. Subsequently, biochemical

assays showed that the COM_{2C} domain was necessary for efficient catalysis, and the non-native PCP-COM_{2C} (within Sven_5012) from watasemycin NRPS could facilitate catalysis with the COM_{2N}-C didomain (within Bamb_5915) from enacyloxin PKS-NRPS. This demonstrates the inherent promiscuity for this particular pair of COM domains.

The abundance of type 2 COM domains lead to the critical evaluation of a tool for identifying type 1 COM domains and docking domains. Investigations of terminal amino acid sequences from PKS and NRPS proteins led to new HMMs, to identify previously unidentified features from characterised biosynthetic pathways. The use of HMMs, to describe docking domains, also led to the identification of docking domains that were not located at protein termini. An observation from a collection of N-terminal DH domains led to the proposal of a novel form of intermolecular interaction. A C-terminal interaction peptide was proposed to interact directly with the DH domain. This domain was described using a HMM, which appeared to identify DH domains, where an ACP-bound (2*E*,4*Z*)-configured diene produced by the same module.

6.1.2 Objective II review

The genomics-driven search for enacyloxin analogues led to the identification of a homologous biosynthetic gene cluster within *V. rhizosphaerae*. The production of the enacyloxin analogue, vibroxin, was optimised, extracted, purified and structurally characterised. A narrow spectrum of antimicrobial activity was observed, with potent antimicrobial activity against the multi-drug resistant pathogen *A. baumannii*.

The core molecular structure of vibroxin exhibited some features that were not predicted from the biosynthetic gene cluster. Namely, the dimethylation of the starter unit, and a dehydration event corresponding to the second extension module. Together, this formed a proposed biosynthetic pathway for the antibiotic vibroxin. An HMM was built to describe sMT domains within loading modules.

This facilitated their annotation within PKS proteins. Subsequently, fragments of sMT domains were identified within characterised biosynthetic pathways. This uncovered pieces of PKS evolution where enzmye activity is lost, and domains are partially excised. Similarly, fragments of AT domains were also identified within vibroxin PKS-NRPS and enacyloxin PKS-NRPS.

6.2 Context and outlook

The widespread use of docking domains and COM domains, to facilitate biosynthesis across protein boundaries, provides new opportunities to exploit these features to engineer biosynthetic pathways. Not only are type 2 COM domains involved in the biosynthesis of enacyloxin and vibroxin, but they present a common form of intermolecular interactions within biosynthetic pathways. These intermolecular interactions direct the biosynthesis of antibiotics, antifungal agents, insecticides, protease inhibitors, biosurfactants and cannabinomimetics from a diverse genera of producing organisms. Cryptic biosynthetic clusters indicate more diverse biochemistry across COM-junctions. Thus, providing opportunities for targeted intermediate tailoring during thiol-template biosynthesis, and chain release mechanisms, represented by thioesterases, thioester-reductases, halogenases and methyltransferases.

The identification of putative C-terminal docking peptides across a set of KS and DH domain boundaries, would benefit from *in vitro* biophysical and biochemical analyses from an isolated system. Similarly, the mechanism of putative (2E,4Z)-configuring DH domains from characterised biosynthetic pathways could be probed experimentally. Furthermore, a demonstration of the predictive efficacy of the (2E,4Z)-DH HMM, would be to identify this domain within a cryptic biosynthetic gene cluster, and investigate the configuration of a corresponding diene.

The discovery of antibiotic vibroxin is a testimony to genomics-driven approaches for bioactive natural product discovery. It also highlights the untapped

potential of natural products from rhizosphere-associated *Vibrio*. Whilst a predicted core molecular structure assisted with the identification of vibroxin, experimental characterisation revealed the limitations of current predictive tools, and our understanding of modular PKSs. The identification of sMT domains within cryptic biosynthetic clusters will increase this predictive efficacy. The identification of inactive sMT domains, sMT fragments, AT fragments, and KS fragments demonstrates consequences of blind evolution. It is convenient to regard combinatorial biosynthetic pathways as collections of modular building blocks, susceptible to rearrangement. However, it is a misconception to interpret nature's biosynthetic pathways as rational designs. As a consequence, domain fragments, inactive domains (or partially active domains *e.g.* KS^0) and unusual domain architectures should be viewed with appreciation of their evolutionary development.

Bibliography

(1) Enright, M. C.; Robinson, D. A.; Randle, G.; Feil, E. J.; Grundmann, H.; Spratt, B. G. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7687–7692.

(2) Livermore, D. M. *Clin. Infect. Dis.* **2002**, *34*, 634–640.

(3) Dijkshoorn, L.; Nemec, A.; Seifert, H. *Nat. Rev. Microbiol.* **2007**, *5*, 939–951.

(4) Zhang, R.; Eggleston, K.; Rotimi, V.; Zeckhauser, R. J. *Global Health* **2006**, *2*, 1.

(5) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2007**, *70*, 461–477.

(6) Maier, T.; Jenni, S.; Ban, N. *Science* **2006**, *311*, 1258–1262.

(7) Jenni, S.; Leibundgut, M.; Boehringer, D.; Frick, C.; Mikolásek, B.; Ban, N. *Science* **2007**, *316*, 254–261.

(8) Ciccarelli, L.; Connell, S. R.; Enderle, M.; Mills, D. J.; Vonck, J.; Grininger, M. *Structure* **2013**, *21*, 1251–1257.

(9) Leibundgut, M.; Jenni, S.; Frick, C.; Ban, N. *Science* **2007**, *316*, 288–290.

(10) Khosla, C.; Kapur, S.; Cane, D. E. *Curr. Opin. Chem. Biol.* **2009**, *13*, 135–143.

(11) Yu, D.; Xu, F.; Zeng, J.; Zhan, J. *IUBMB life* **2012**, *64*, 285–295.

(12) Hertweck, C.; Luzhetskyy, A.; Rebets, Y.; Bechthold, A. *Nat. Prod. Rep.* **2007**, *24*, 162–190.

(13) Wagner, D. T.; Stevens, D. C.; Mehaffey, M. R.; Manion, H. R.; Taylor, R. E.; Brodbelt, J. S.; Keatinge-Clay, A. T. *Chem. Commun. (Camb.)* **2016**, *52*, 8822–8825.

(14) Khayatt, B. I.; Overmars, L.; Siezen, R. J.; Francke, C. *PloS one* **2013**, *8*, e62136.

(15) Stachelhaus, T; Mootz, H. D.; Marahiel, M. A. *Chem. Biol.* **1999**, *6*, 493–505.

(16) Keatinge-Clay, A. T. *Chem. Biol.* **2007**, *14*, 898–908.

(17) Kwan, D. H.; Sun, Y.; Schulz, F.; Hong, H.; Popovic, B.; Sim-Stark, J. C. C.; Haydock, S. F.; Leadlay, P. F. *Chem. Biol.* **2008**, *15*, 1231–1240.

(18) Whicher, J. R.; Smaga, S. S.; Hansen, D. A.; Brown, W. C.; Gerwick, W. H.; Sherman, D. H.; Smith, J. L. *Chem. Biol.* **2013**, *20*, 1340–1351.

(19) Broadhurst, R. W.; Nietlispach, D.; Wheatcroft, M. P.; Leadlay, P. F.; Weissman, K. J. *Chem. Biol.* **2003**, *10*, 723–731.

(20) Buchholz, T. J.; Geders, T. W.; Bartley, F. E.; Reynolds, K. A.; Smith, J. L.; Sherman, D. H. *ACS Chem. Biol.* **2009**, *4*, 41–52.

(21) Dorival, J.; Annaval, T.; Risser, F.; Collin, S.; Roblin, P.; Jacob, C.; Gruez, A.; Chagot, B.; Weissman, K. J. *J. Acc. Chem. Res.* **2016**, *138*, 4155–4167.

(22) Zeng, J.; Wagner, D. T.; Zhang, Z.; Moretto, L.; Addison, J. D.; Keatinge-Clay, A. T. *ACS Chem. Biol.* **2016**, *11*, 2466–2474.

(23) Richter, C. D.; Nietlispach, D; Broadhurst, R. W. *Nat. Chem. Biol.* **2008**, *4*, 75–81.

(24) Tanovic, A.; Samel, S. A.; Essen, L.-O.; Marahiel, M. A. *Science* **2008**, *321*, 659–663.

(25) Crosby, J.; Crump, M. P. *Nat. Prod. Rep.* **2012**, *29*, 1111–1137.

(26) Castoe, T. A.; Stephens, T.; Noonan, B. P.; Calestani, C. *Gene* **2007**, *392*, 47–58.

(27) Jenke-Kodama, H; Sandmann, A; Müller, R *Mol. Biol. Evol.* **2005**, *22*, 2027–2039.

(28) Liu, X.; Cheng, Y.-Q. *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 275–284.

(29) Coenye, T.; Vandamme, P. *Environ. Microbiol. Rep.* **2003**, *5*, 719–729.

(30) Parke, J. L.; Gurian-Sherman, D *Annu. Rev. Phytopathol.* **2001**, *39*, 225–258.

(31) Chain, P. S. G. et al. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 15280–15287.

(32) O'Sullivan, L. A.; Weightman, A. J.; Jones, T. H.; Marchbank, A. M.; Tiedje, J. M.; Mahenthiralingam, E. *Environ. Microbiol. Rep.* **2007**, *9*, 1017–1034.

(33) Mahenthiralingam, E.; Urban, T. A.; Goldberg, J. B. *Nat. Rev. Microbiol.* **2005**, *3*, 144–156.

(34) Mahenthiralingam, E.; Song, L.; Sass, A.; White, J.; Wilmot, C.; Marchbank, A.; Boaisha, O.; Paine, J.; Knight, D.; Challis, G. L. *Chem. Biol.* **2011**, *18*, 665–677.

(35) Masschelein, J.; Challis, G. L. *unpublished.*

(36) Sydor, P. K.; Challis, G. L. *unpublished.*

(37) Hahn, M.; Stachelhaus, T. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 15585–15590.

(38) Chen, W.-H.; Li, K.; Guntaka, N. S.; Bruner, S. D. *ACS Chem. Biol.* **2016**, *11*, 2293–2303.

(39) Li, Y.; Weissman, K. J.; Müller, R. *Chembiochem : a European journal of chemical biology* **2010**, *11*, 1069–1075.

(40) Pancsa, R.; Fuxreiter, M. *IUBMB life* **2012**, *64*, 513–520.

(41) Sharma, R.; Raduly, Z.; Miskei, M.; Fuxreiter, M. *FEBS letters* **2015**, *589*, 2533–2542.

(42) Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee, S. Y.; Fischbach, M. A.; Müller, R.; Wohlleben, W.; Breitling, R.; Takano, E.; Medema, M. H. *Nucleic Acids Res.* **2015**, *43*, W237–43.

(43) Medema, M. H. et al. *Nat. Chem. Biol.* **2015**, *11*, 625–631.

(44) Bailey, T. L.; Elkan, C *Proc. Int. Conf. Intell Syst. Mol. Biol.* **1994**, *2*, 28–36.

(45) Punta, M; Qureshi, M; Sangrador-Vegas, A *Nucleic Acids Res.* **2016**, *44*, D279–D285.

(46) Letunic, I.; Doerks, T.; Bork, P. *Nucleic Acids Res.* **2015**, *43*, D257–60.

(47) Angiuoli, S. V.; Matalka, M *BMC Bioinformatics* **2011**, *12*, 1.

(48) Haft, D. H.; Selengut, J. D.; White, O *Nucleic Acids Res.* **2003**, *31*, 371–373.

(49) Finn, R. D.; Clements, J; Eddy, S. R. *Nucleic Acids Res.* **2011**, *39*, W29–W37.

(50) Wang, H.; Fewer, D. P.; Holm, L.; Rouhiainen, L.; Sivonen, K. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 9259–9264.

(51) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. *Genome Res.* **2004**, *14*, 1188–1190.

(52) Conway, K. R.; Boddy, C. N. *Nucleic Acids Res.* **2013**, *41*, D402–7.

(53) Ishida, K.; Welker, M.; Christiansen, G.; Cadel-Six, S.; Bouchier, C.; Dittmann, E.; Hertweck, C.; Tandeau de Marsac, N. *J Appl. Environ. Microbiol.* **2009**, *75*, 2017–2026.

(54) Ishida, K; Okita, Y; Matsuda, H; Okino, T; Murakami, M *Tetrahedron* **1999**, *55*, 10971–10988.

(55) Buntin, K.; Rachid, S.; Scharfe, M.; Blöcker, H.; Weissman, K. J.; Müller, R. *Angew. Chem. Int. Ed. Engl.* **2008**, *47*, 4595–4599.

(56) Kunze, B.; Jansen, R.; Höfle, G.; Reichenbach, H. *J. Antibiot.* **2004**, *57*, 151–155.

(57) Rouhiainen, L; Paulin, L; Suomalainen, S; Hyytiäinen, H; Buikema, W; Haselkorn, R; Sivonen, K *Mol. Microbiol.* **2000**, *37*, 156–167.

(58) Namikoshi, M; Rinehart, K. L. *J. Ind. Microbiol. Biotechnol.* **1996**, *17*, 373–384.

(59) Christiansen, G.; Philmus, B.; Hemscheidt, T.; Kurmayer, R. *J. Bacteriol.* **2011**, *193*, 3822–3831.

(60) Spoof, L.; Błaszczyk, A.; Meriluoto, J.; Cegłowska, M.; Mazur-Marzec, H. *Mar. Drugs* **2016**, *14*, 8.

(61) Schwarzer, D; Finking, R; Marahiel, M. A. *Nat. Prod. Rep.* **2003**, *20*, 275–287.

(62) Johnson, B. A.; Anker, H; Meleney, F. L. *Science* **1945**, *102*, 376–377.

(63) Chang, Z.; Flatt, P.; Gerwick, W. H.; Nguyen, V.-A.; Willis, C. L.; Sherman, D. H. *Gene* **2002**, *296*, 235–247.

(64) Orjala, J; Gerwick, W. H. *J. Nat. Prod.* **1996**, *59*, 427–430.

(65) Du, L; Sánchez, C; Chen, M; Edwards, D. J.; Shen, B *Chem. Biol.* **2000**, *7*, 623–642.

(66) Blum, R. H.; Carter, S. K.; Agre, K *Cancer* **1973**, *31*, 903–914.

(67) Rachid, S.; Krug, D.; Kunze, B.; Kochems, I.; Scharfe, M.; Zabriskie, T. M.; Blöcker, H.; Müller, R. *Chem. Biol.* **2006**, *13*, 667–681.

(68) Menhofer, M. H.; Kubisch, R.; Schreiner, L.; Zorn, M.; Foerster, F.; Mueller, R.; Raedler, J. O.; Wagner, E.; Vollmar, A. M.; Zahler, S. *PloS one* **2014**, *9*, e112542.

(69) Zhao, B.; Moody, S. C.; Hider, R. C.; Lei, L.; Kelly, S. L.; Waterman, M. R.; Lamb, D. C. *Int. J. Mol. Sci.* **2012**, *13*, 8500–8513.

(70) Kleigrewe, K.; Almaliti, J.; Tian, I. Y.; Kinnel, R. B.; Korobeynikov, A.; Monroe, E. A.; Duggan, B. M.; Di Marzo, V.; Sherman, D. H.; Dorrestein, P. C.; Gerwick, L.; Gerwick, W. H. *J. Nat. Prod.* **2015**, *78*, 1671–1682.

(71) Jones, A. C.; Monroe, E. A.; Eisman, E. B.; Gerwick, L.; Sherman, D. H.; Gerwick, W. H. *Nat. Prod. Rep.* **2010**, *27*, 1048–1065.

(72) Smith, C. D.; Zhang, X; Mooberry, S. L.; Patterson, G. *Cancer Res.* **1994**, *54*, 3779–3784.

(73) Rounge, T. B.; Rohrlack, T *BMC Microbiol.* **2008**, *8*, 1.

(74) Feng, Z.; Qi, J.; Tsuge, T.; Oba, Y.; Kobayashi, T.; Suzuki, Y.; Sakagami, Y.; Ojika, M. *Biosci. Biotechnol. Biochem.* **2005**, *69*, 1372–1380.

(75) Ojika, M; Suzuki, Y; Tsukamoto, A; Sakagami, Y; Fudou, R; Yoshimura, T; Yamanaka, S *J. Antibiot.* **1998**, *51*, 275–281.

(76) Xu, Y; Kersten, R. D.; Nam, S. J.; Lu, L *J. Am. Chem. Soc.* **2012**, *134*, 8625–8632.

(77) Jiang, T. L.; Liu, R. H.; Salmon, S. E. *Cancer Chemother. Pharmacol.* **1983**, *11*, 1–4.

(78) Meiser, P.; Weissman, K. J.; Bode, H. B.; Krug, D.; Dickschat, J. S.; Sandmann, A.; Müller, R. *Chem. Biol.* **2008**, *15*, 771–781.

(79) Meiser, P.; Bode, H. B.; Müller, R. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 19128–19133.

(80) Mahenthiralingam, E.; Song, L.; Sass, A.; White, J.; Wilmot, C.; Marchbank, A.; Boaisha, O.; Paine, J.; Knight, D.; Challis, G. L. *Chem. Biol.* **2011**, *18*, 665–677.

(81) Vallet-Gely, I.; Novikov, A.; Augusto, L.; Liehl, P.; Bolbach, G.; Péchy-Tarr, M.; Cosson, P.; Keel, C.; Caroff, M.; Lemaitre, B. *J Appl. Environ. Microbiol.* **2010**, *76*, 910–921.

(82) Stevens, D. C.; Hari, T. P. A.; Boddy, C. N. *Nat. Prod. Rep.* **2013**, *30*, 1391–1411.

(83) Watkins, E. B.; Chittiboyina, A. G.; Jung, J. C. *Curr. Pharm. Des.* **2005**, *11*, 1615–1653.

(84) Heather, Z.; Holden, M. T. G.; Steward, K. F.; Parkhill, J.; Song, L.; Challis, G. L.; Robinson, C.; Davis-Poynter, N.; Waller, A. S. *Mol. Microbiol.* **2008**, *70*, 1274–1292.

(85) Ramaswamy, A. V.; Sorrels, C. M.; Gerwick, W. H. *J. Nat. Prod.* **2007**, *70*, 1977–1986.

(86) Marquez, B. L.; Watts, K. S.; Yokochi, A.; Roberts, M. A.; Verdier-Pinard, P.; Jimenez, J. I.; Hamel, E.; Scheuer, P. J.; Gerwick, W. H. *J. Nat. Prod.* **2002**, *65*, 866–871.

(87) Kopp, M.; Irschik, H.; Gemperlein, K.; Buntin, K.; Meiser, P.; Weissman, K. J.; Bode, H. B.; Müller, R. *Mol. Biosyst.* **2011**, *7*, 1549–1563.

(88) Bode, H. B.; Irschik, H.; Wenzel, S. C.; Reichenbach, H.; Müller, R.; Höfle, G. *J. Nat. Prod.* **2003**, *66*, 1203–1206.

(89) Kleigrewe, K; Gerwick, L; Sherman, D. H. *Nat. Prod. Rep.* **2016**, *33*, 348–364.

(90) Choi, H.; Mevers, E.; Byrum, T.; Valeriote, F. A.; Gerwick, W. H. *European J. Org. Chem.* **2012**, *2012*, 5141–5150.

(91) De Bruijn, I; de Kock, M. J. D.; de Waard, P; van Beek, T. A.; Raaijmakers, J. M. *J. Bacteriol.* **2008**, *190*, 2777–2789.

(92) Weinig, S; Hecht, H. J.; Mahmud, T; Müller, R *Chem. Biol.* **2003**, *10*, 939–952.

(93) Sasse, F; Böhlendorf, B; Herrmann, M; Kunze, B; Forche, E; Steinmetz, H; Höfle, G; Reichenbach, H; Hermann, M *J. Antibiot.* **1999**, *52*, 721–729.

(94) Rounge, T. B.; Rohrlack, T.; Nederbragt, A. J.; Kristensen, T.; Jakobsen, K. S. *BMC genomics* **2009**, *10*, 396.

(95) Okino, T; Matsuda, H; Murakami, M; Yamaguchi, K *Tetrahedron Lett.* **1993**, *34*, 501–504.

(96) Nishizawa, A.; Arshad, A. B.; Nishizawa, T.; Asayama, M.; Fujii, K.; Nakano, T.; Harada, K.-i.; Shirai, M. *J. Gen. Appl. Microbiol.* **2007**, *53*, 17–27.

(97) Hoffmann, T.; Müller, S.; Nadmid, S.; Garcia, R.; Müller, R. *J. Acc. Chem. Res.* **2013**, *135*, 16904–16911.

(98) Bewley, C. A.; Debitus, C; Faulkner, D. J. *J. Acc. Chem. Res.* **1994**, *116*, 7631–7636.

(99) Silakowski, B; Nordsiek, G; Kunze, B; Blöcker, H *Chem. Biol.* **2001**, *8*, 59–69.

(100) Gerth, K; Jansen, R; Reifenstahl, G; Höfle, G; Irschik, H; Kunze, B; Reichenbach, H; Thierbach, G *J. Antibiot.* **1983**, *36*, 1150–1156.

(101) Perlova, O; Fu, J; Kuhlmann, S; Krug, D *J Appl. Environ. Microbiol.* **2006**, *72*, 7485–7494.

(102) Gerth, K; Irschik, H; Reichenbach, H *J. Antibiot.* **1980**, *33*, 1474–1479.

(103) Krastel, P; Roggo, S; Schirle, M *Angew. Chem.* **2015**, *54*, 10149–10154.

(104) Hoffmann, H.; Kogler, H.; Heyse, W.; Matter, H.; Caspers, M.; Schummer, D.; Klemke-Jahn, C.; Bauer, A.; Penarier, G.; Debussche, L.; Brönstrup, M. *Angew. Chem. Int. Ed. Engl.* **2015**, *54*, 10145–10148.

(105) Becker, J. E.; Moore, R. E.; Moore, B. S. *Gene* **2004**, *325*, 35–42.

(106) Herfindal, L.; Myhren, L.; Kleppe, R.; Krakstad, C.; Selheim, F.; Jokela, J.; Sivonen, K.; Døskeland, S. O. *Mol. Pharm.* **2011**, *8*, 360–367.

(107) Liaimer, A.; Helfrich, E. J. N.; Hinrichs, K.; Guljamow, A.; Ishida, K.; Hertweck, C.; Dittmann, E. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 1862–1867.

(108) Ma, Z.; Geudens, N.; Kieu, N. P.; Sinnaeve, D.; Ongena, M.; Martins, J. C.; Höfte, M. *Front. Microbiol.* **2016**, *7*, 382.

(109) Jang, J. Y.; Yang, S. Y.; Kim, Y. C.; Lee, C. W.; Park, M. S.; Kim, J. C.; Kim, I. S. *J. Agric. Food Chem.* **2013**, *61*, 6786–6791.

(110) Zachow, C.; Jahanshah, G.; de Bruijn, I.; Song, C.; Ianni, F.; Pataj, Z.; Gerhardt, H.; Pianet, I.; Lämmerhofer, M.; Berg, G.; Gross, H.; Raaijmakers, J. M. *Mol. Plant Microbe Interact.* **2015**, *28*, 800–810.

(111) Hrouzek, P.; Kuzma, M.; Černý, J.; Novák, P.; Fišer, R.; Šimek, P.; Lukešová, A.; Kopecký, J. *Chem. Res. Toxicol.* **2012**, *25*, 1203–1211.

(112) Donadio, S; Monciardini, P; Sosio, M *Nat. Prod. Rep.* **2007**, *24*, 1073–109.

(113) Gupte P Kulkarni B Ganguli, M *Appl. Microbiol. Biotechnol.* **2002**, *58*, 46–57.

(114) Cheng, Y. Q.; Yang, M; Matter, A. M. *J Appl. Environ. Microbiol.* **2007**, *73*, 3460–3469.

(115) VanderMolen, K. M.; McCulloch, W; Pearce, C. J. *J. Antibiot.* **2011**, *64*, 525–31.

(116) D'aes, J.; Kieu, N. P.; Léclère, V.; Tokarski, C.; Olorunleke, F. E.; De Maeyer, K.; Jacques, P.; Höfte, M.; Ongena, M. *Environ. Microbiol. Rep.* **2014**, *16*, 2282–2300.

(117) Hua, G.; Höfte, M *Plant Soil* **2015**, *388*, 243–253.

(118) Burch, A. Y.; Shimada, B. K.; Mullin, S. *J. Bacteriol.* **2012**, *194*, 1287–1298.

(119) Burch, A. Y.; Zeisler, V.; Yokota, K.; Schreiber, L.; Lindow, S. E. *Environ. Microbiol. Rep.* **2014**, *16*, 2086–2098.

(120) Scholz-Schroeder, B. K.; Soule, J. D. *Mol. Plant Microbe Interact.* **2003**, *16*, 271–280.

(121) Buber, E; Stindl, A; Acan, N. L.; Kocagoz, T; Zocher, R *Nat. Prod. Lett.* **2002**, *16*, 419–423.

(122) Galm, U; Wendt-Pienkowski, E; Wang, L *J. Nat. Prod.* **2011**, *74*, 526–536.

(123) Nicaise, C; Ajani, J; Goudeau, P; Rozencweig, M; Levin, B; Krakoff, I *Cancer Chemother. Pharmacol.* **1990**, *26*, 221–222.

(124) Scherlach, K.; Lackner, G.; Graupner, K.; Pidot, S.; Bretschneider, T.; Hertweck, C. *Chembiochem* **2013**, *14*, 2439–2443.

(125) Nutkins, J. C.; Mortishire-Smith, R. J. *J. Acc. Chem. Res.* **1991**, *113*, 2621–2627.

(126) Lamidi, O. F.; Sani, M.; Lazzari, P.; Zanda, M.; Fleming, I. N. *J. Cancer Res. Clin. Oncol.* **2015**, *141*, 1575–1583.

(127) De Bruijn, I; de Kock, M.; Yang, M *Mol. Microbiol.* **2007**, *63*, 417–28.

(128) Neu, T. R.; Härtner, T; Poralla, K *Appl. Microbiol. Biotechnol.* **1990**, *32*, 518–520.

(129) Sasaki, O.; Igarashi, Y.; Saito, N.; Furumai, T. *J. Antibiot.* **2002**, *55*, 249–255.

(130) Rokni-Zadeh, H.; Li, W.; Sanchez-Rodriguez, A.; Sinnaeve, D.; Rozenski, J.; Martins, J. C.; De Mot, R. *J Appl. Environ. Microbiol.* **2012**, *78*, 4826–4834.

(131) Li, W.; Rokni-Zadeh, H.; De Vleeschouwer, M.; Ghequire, M. G. K.; Sinnaeve, D.; Xie, G.-L.; Rozenski, J.; Madder, A.; Martins, J. C.; De Mot, R. *PloS one* **2013**, *8*, e62946.

(132) Galm, U; Wendt-Pienkowski, E; Wang, L *Mol Biosyst.* **2009**, *5*, 77–90.

(133) Argoudelis, A. D.; Bergy, M. E.; Pyke, T. R. *J. Antibiot.* **1971**, *24*, 543–557.

(134) Edwards, D. J.; Marquez, B. L.; Nogle, L. M.; McPhail, K.; Goeger, D. E.; Roberts, M. A.; Gerwick, W. H. *Chem. Biol.* **2004**, *11*, 817 –833.

(135) Nazi, I.; Koteva, K. P.; Wright, G. D. *Anal. Biochem.* **2004**, *324*, 100–105.

(136) Medema, M. H.; Fischbach, M. A. *Nat. Chem. Biol.* **2015**, *11*, 639–648.

(137) Chaudhuri, K.; Chatterjee, S. N. *Springer Berlin Heidelberg* **2009**, 13–31.

(138) Yeung, P. S. M.; Boor, K. J. *Foodborne Pathog. Dis.* **2004**, *1*, 74–88.

(139) Qadri, F.; Chowdhury, N. R.; Takeda, Y.; Nair, G. B. *Springer US Boston, MA* **2005**, 277–295.

(140) Lee, K. H.; Ruby, E. G. *J Appl. Environ. Microbiol.* **1994**, *60*, 1565–1571.

(141) Lee, M.-J.; Jeong, D.-Y.; Kim, W.-S.; Kim, H.-D.; Kim, C.-H.; Park, W.-W.; Park, Y.-H.; Kim, K.-S.; Kim, H.-M.; Kim, D.-S. *J Appl. Environ. Microbiol.* **2000**, *66*, 1698–1701.

(142) Rameshkumar, N; Fukui, Y; Sawabe, T; Nair, S *Int. J. Syst. Evol. Microbiol.* **2008**, *58*, 1608–1615.

(143) Rameshkumar, N; Gomez-Gil, B; Sproer, C.; Lang, E.; Dinesh Kumar, N; Krishnamurthi, S; Nair, S.; Roque, A *Syst. Appl. Microbiol.* **2011**, *34*, 487–493.

(144) Rameshkumar, N; Sproer, C.; Lang, E.; Nair, S. *FEMS Microbiol. Lett.* **2010**, *307*, 35–40.

(145) Kumar, N. R.; Nair, S *Int. J. Syst. Evol. Microbiol.* **2007**, *57*, 2241–2246.

(146) Mansson, M.; Gram, L.; Larsen, T. O. *Mar. Drugs* **2011**, *9*, 1440–1468.

(147) Harwood, C. S. *Curr. Microbiol.* **1978**, *1*, 233–238.

(148) Shieh, W. Y. *Int. J. Syst. Evol. Microbiol.* **2003**, *53*, 479–484.

(149) D'Aoust, J. Y.; Gerber, N. N. *J. Bacteriol.* **1974**, *118*, 756–757.

(150) Gerber, N. N. *Tetrahedron Lett.* **1983**, *24*, 2797–2798.

(151) Lattasch, H.; Thomson, R. H. *Tetrahedron Lett.* **1983**, *24*, 2701–2704.

(152) Pandey, R; Chander, R; Sainis, K *Curr. Pharm. Des.* **2009**, *15*, 732–741.

(153) Arellano, M. L.; Borthakur, G.; Berger, M.; Luer, J.; Raza, A. *Clin. Lymphoma Myeloma Leuk.* **2014**, *14*, 534–539.

(154) Fürstner, A. *Angew. Chem. Int. Ed. Engl.* **2003**, *42*, 3582–3603.

(155) Chen, X.; Schauder, S.; Potier, N.; Van Dorsselaer, A.; Pelczer, I.; Bassler, B. L.; Hughson, F. M. *Nature* **2002**, *415*, 545–549.

(156) Hamann, M. T. *Curr. Opin. Mol. Ther.* **2004**, *6*, 657–665.

(157) Noguchi, T; Hwang, D. F.; Arakawa, O; Sugita, H; Deguchi, Y; Shida, Y; Hashimoto, K *Mar. Biol.* **1987**, *94*, 625–630.

(158) Lee, M.-J.; Jeong, D.-Y.; Kim, W.-S.; Kim, H.-D.; Kim, C.-H.; Park, W.-W.; Park, Y.-H.; Kim, K.-S.; Kim, H.-M.; Kim, D.-S. *J Appl. Environ. Microbiol.* **2000**, *66*, 1698–1701.

(159) Hagen, N. A.; du Souich, P.; Lapointe, B.; Ong-Lam, M.; Dubuc, B.; Walde, D.; Love, R.; Ngoc, A. H.; Canadian Tetrodotoxin Study Group *J. Pain Symptom Manage.* **2008**, *35*, 420–429.

(160) Ross, C.; Opel, V.; Scherlach, K.; Hertweck, C. *Mycoses* **2014**, *57*, 48–55.

(161) Musiol, E. M.; Härtner, T.; Kulik, A.; Moldenhauer, J.; Piel, J.; Wohlleben, W.; Weber, T. *Chem. Biol.* **2011**, *18*, 438–444.

(162) Gay, D. C.; Wagner, D. T.; Meinke, J. L.; Zogzas, C. E.; Gay, G. R.; Keatinge-Clay, A. T. *J. Struct. Biol.* **2016**, *193*, 196–205.

(163) Gu, L.; Geders, T. W.; Wang, B.; Gerwick, W. H.; Håkansson, K.; Smith, J. L.; Sherman, D. H. *Science* **2007**, *318*, 970–974.

(164) Chen, S.-C.; Huang, C.-H.; Lai, S.-J.; Liu, J.-S.; Fu, P.-K.; Tseng, S.-T.; Yang, C. S.; Lai, M.-C.; Ko, T.-P.; Chen, Y. *Sci. Rep.* **2015**, *5*, 10100.

(165) Mihali, T. K.; Carmichael, W. W.; Neilan, B. A. *PloS one* **2011**, *6*, e14657.

(166) Young, J.; Stevens, D. C.; Carmichael, R.; Tan, J.; Rachid, S.; Boddy, C. N.; Müller, R.; Taylor, R. E. *J. Nat. Prod.* **2013**, *76*, 2269–2276.

(167) Grindberg, R. V.; Ishoey, T.; Brinza, D.; Esquenazi, E.; Coates, R. C.; Liu, W.-t.; Gerwick, L.; Dorrestein, P. C.; Pevzner, P.; Lasken, R.; Gerwick, W. H. *PloS one* **2011**, *6*, e18565.

(168) Minowa, Y.; Araki, M.; Kanehisa, M. *J. Mol. Biol.* **2007**, *368*, 1500–1517.

(169) Theodore, C. M.; Stamps, B. W.; King, J. B.; Price, L. S. L.; Powell, D. R.; Stevenson, B. S.; Cichewicz, R. H. *PloS one* **2014**, *9*, e90124.

(170) Shibazaki, M.; Sugawara, T.; Nagai, K.; Shimizu, Y.; Yamaguchi, H.; Suzuki, K. *J. Antibiot.* **1996**, *49*, 340–344.

(171) Partida Martinez, L. P.; Hertweck, C. *ChemBioChem* **2007**, *8*, 41–45.

(172) Mattheus, W.; Gao, L.-J.; Herdewijn, P.; Landuyt, B.; Verhaegen, J.; Masschelein, J.; Volckaert, G.; Lavigne, R. *Chem. Biol.* **2010**, *17*, 149–159.

(173) Kampa, A.; Gagunashvili, A. N.; Gulder, T. A. M.; Morinaka, B. I.; Daolio, C.; Godejohann, M.; Miao, V. P. W.; Piel, J.; Andrésson, Ó. S. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, E3129–37.

(174) Necati Hakyemez, I.; Kucukbayrak, A.; Tas, T.; Burcu Yikilgan, A.; Akkaya, A.; Yasayacak, A.; Akdeniz, H. *Pak. J. Med. Sci.* **2013**, *29*, 1245–1248.

(175) Hareland, W. A.; Crawford, R. L.; Chapman, P. J.; Dagley, S *J. Bacteriol.* **1975**, *121*, 272–285.

(176) Sambrook, J; Russell, D. W., *Molecular cloning*; Cold Spring Habour Laboratory Press: 2001; Vol. 1-3.

(177) Gasteiger, E; Hoogland, C; Gattiker, A; Duvaud, S, *Protein identification and analysis tools on the ExPASy server.* Humana Press: 2005.

(178) Lee, W.; Tonelli, M.; Markley, J. L. *Bioinformatics* **2015**, *31*, 1325–1327.

(179) Lee, W.; Westler, W. M.; Bahrami, A.; Eghbalnia, H. R.; Markley, J. L. *Bioinformatics* **2009**, *25*, 2085–2087.

(180) Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee, S. Y.; Fischbach, M. A.; Müller, R.; Wohlleben, W., et al. *Nucleic Acids Res.* **2015**, *43*, W237–W243.

(181) Finn, R. D.; Clements, J.; Eddy, S. R. *Nucleic Acids Res.* **2011**, gkr367.

(182) Kurcinski, M.; Jamroz, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. *Nucleic Acids Res.* **2015**, *43*, W419–24.

Chapter 7

Appendix

7.1 Sequences of PKS and NRPS domains

>Bamb_5919_ACP_aa
MHHHHHHGKPIPNPLLGLDSTENLYFQGIDPFTLAELVEAEAGATGASG
APRVERRAGGTAGAALLAGLASLDAAARAARLKRHLEAAIRKLLNRADT
LDDRASMFDLGLDSLLSIDLRMQLEKDLACSLSTTVLHDHPTIEALAGFLA
ERVGAPPAGTVRAGAAGGAGAGTGAPAGATGAAAAHAVSSASPVPAG
AASAAASAASAAAAAGAPSRATFA

>Bamb_5919_ACP_dna
ATGCATCATCACCATCACCATGGTAAGCCTATCCCTAACCCTCTCCT
CGGTCTCGATTCTACGGAAAACCTGTATTTTCAGGGAATTGATCCCT
TCACCCTGGCCGAGCTGGTCGAGGCCGAGGCAGGCGCGACGGGCGC
GAGCGGAGCACCGCGCGTCGAGCGCCGCGCCGGCGGCACGGCCGGC
GCGGCACTGCTGGCGGGCCTCGCGAGCCTCGATGCGGCGGCGCGCG
CGGCGCGCCTGAAGCGCCATCTCGAGGCCGCGATCCGCAAGCTGCTC
AACCGCGCCGATACGCTCGACGATCGCGCCAGCATGTTCGATCTCGG
TCTCGATTCGCTGCTCAGTATCGACCTGCGCATGCAGCTCGAGAAGG
ACCTGGCCTGCAGCCTCTCGACCACGGTGCTGCACGACCATCCGACC
ATCGAGGCGCTGGCGGGCTTCCTGGCCGAACGCGTGGGTGCGCCGC
CGGCGGGGACGGTTCGCGCAGGGGCCGCGGGCGGTGCCGGTGCAGG
CACCGGCGCGCCTGCCGGCGCCACTGGGGCCGCGGCTGCGCATGCC
GTATCGTCGGCCTCGCCCGTGCCGGCCGGGGCCGCGTCGGCCGCTG
CATCCGCTGCATCCGCTGCAGCGGCAGCCGGCGCCCCGTCGCGCGCC
ACGTTCGCG

>Bamb_5919_KS^0_aa
MHHHHHHGKPIPNPLLGLDSTENLYFQGIDPFTPSRATFAAEPRRAGGA
ALPPGAGPDDIAIIGVSGRYPGAADLGAFWDNLRDGHDAITPIPPERWN
HDAYFDRQRNVPGKSYSAWGGFIEDVDAFDPAFFSISPRMSAYLDPKER
LFLETVWNLLEEAGETRERMQQAYGAQVGVFVGAMYQLYGACAADEG
ERVATALSSYNAIAHRTSYFFNLRGPSIALDTMCSSSLTAVHYACRSLLDG
DCALAIAGGVNLSLHPRKYVGLSQAQIVGSHADSRSFSDGDGYLPAEGVG
AVLLKPLARALADDDRILAVIKASSVNHGGRATGYYAPNANAQVDLMEA
SFRKAGVSPESIDYIEAAANGTSLGDAVELRALARVFDGTARDGRRVPIG

TVKSNIGHPEAASGIAQLTKVILQMQHETLVPSIKTEPVNPNLDLAHTPFR
LLSRQAAWPSDPARPRRATVSSFGASGANAHLIVEAFETVEAEPAPAVA
QAAPPAEIVVLSARTPAQLREVARRLLAWLATRQAAGRAESAVPLAERG
RACSLANLAHTLQIGREAMDCRLALLADSLDTLGDGLRRFLGESAGAAE
PAIYHGNVQDQLEMRNLLAGAAGDAMAQTLVAERNLEGLMLHWVQGG
NVPWAALREGRPARRLVLPTYPFERERYWLSGASDAAGRGAGEPQVPA
EPAEAASEPSVVDGRA

>Bamb_5919_KS^0_dna
ATGCATCATCACCATCACCATGGTAAGCCTATCCCTAACCCTCTCCT
CGGTCTCGATTCTACGGAAAACCTGTATTTTCAGGGAATTGATCCCT
TCACCCCGTCGCGCGCCACGTTCGCGGCCGAGCCGCGTCGTGCCGGC
GGCGCGGCGCTGCCGCCCGGCGCCGGCCCCGACGACATCGCCATCAT
CGGCGTATCGGGCCGCTACCCCGGCGCGGCCGACCTCGGCGCGTTCT
GGGACAACCTGCGCGACGGCCACGACGCGATCACCCCGATCCCGCCC
GAGCGCTGGAACCACGACGCCTACTTCGACCGGCAGCGCAACGTGCC
CGGCAAGAGCTACAGCGCCTGGGGCGGCTTCATCGAGGACGTCGAC
GCCTTCGACCCGGCCTTCTTCAGCATCTCGCCGCGGATGTCGGCCTA
CCTCGATCCGAAGGAGCGGCTGTTCCTCGAGACGGTCTGGAACCTGC
TGGAGGAGGCGGGCGAGACGCGCGAGCGCATGCAGCAGGCCTATGG
CGCGCAGGTGGGCGTGTTCGTCGGCGCGATGTACCAGCTCTATGGC
GCCTGCGCGGCCGACGAGGGCGAGCGCGTGGCCACCGCGCTGTCCT
CCTACAACGCGATCGCGCATCGCACCTCGTACTTCTTCAACCTGCGC
GGGCCGAGCATCGCGCTCGACACAATGTGCTCGTCGTCGCTGACGG
CGGTCCACTACGCCTGCCGCAGCCTGCTCGACGGCGACTGCGCGCTG
GCCATCGCGGGCGGCGTGAACCTGTCGCTGCATCCGCGCAAGTACG
TCGGGCTGAGCCAGGCGCAGATCGTCGGCAGCCATGCCGACAGCCG
CAGCTTCAGCGACGGCGACGGCTACCTGCCGGCCGAGGGCGTGGGC
GCCGTGCTGCTCAAGCCGCTGGCCCGCGCGCTGGCCGACGACGACC
GGATCCTGGCGGTGATCAAAGCCTCCTCGGTCAACCACGGCGGCCG
CGCGACCGGCTACTACGCGCCGAACGCGAACGCCCAGGTCGACCTGA
TGGAGGCCAGCTTCCGCAAGGCCGGCGTGTCGCCCGAGTCGATCGA
CTACATCGAGGCCGCCGCCAACGGCACCAGCCTCGGCGACGCGGTCG
AGCTGCGCGCGCTGGCGCGCGTGTTCGACGGCACCGCGCGCGACGG
CCGGCGCGTGCCGATCGGCACTGTGAAGTCGAACATCGGCCATCCCG
AGGCGGCCTCGGGCATCGCGCAACTGACCAAGGTGATCCTGCAGAT
GCAGCACGAGACGCTGGTGCCCTCGATCAAGACCGAGCCCGTCAACC
CCAACCTCGACCTGGCCCACACGCCGTTCCGCCTGCTCTCGCGGCAG
GCGGCCTGGCCGTCCGATCCGGCGCGGCCGCGGCGCGCCACGGTCA
GCTCGTTCGGCGCGAGCGGCGCGAACGCGCACCTGATCGTCGAGGC
CTTCGAGACGGTCGAGGCGGAGCCCGCGCCAGCCGTCGCGCAAGCG
GCGCCGCCGGCCGAGATCGTGGTGCTGTCGGCGCGCACGCCGGCTC
AGTTGCGCGAGGTGGCGCGGCGCCTGCTGGCCTGGCTCGCCACGCG
GCAGGCGGCGGGCAGGGCGGAATCGGCGGTGCCGCTGGCCGAGCGC
GGCCGCGCCTGCTCGCTCGCGAATCTCGCGCACACGCTGCAGATCGG
GCGCGAGGCGATGGACTGCCGGCTCGCGCTGCTGGCCGACAGCCTC
GACACGCTTGGCGATGGCCTGCGGCGTTTCCTCGGCGAATCGGCCG
GCGCGGCCGAGCCCGCGATCTACCACGGCAACGTGCAGGACCAGCT

CGAGATGCGCAACCTGCTGGCGGGCGCCGCCGGCGACGCGATGGCG
CAGACCCTGGTGGCCGAACGCAACCTGGAGGGGCTGATGCTGCACT
GGGTCCAGGGCGGCAACGTGCCCTGGGCCGCCCTGCGCGAGGGCCG
GCCGGCGCGCCGCCTGGTGCTGCCGACCTATCCGTTCGAGCGCGAG
CGCTACTGGCTGTCCGGCGCGAGCGACGCCGCGGGCCGCGGCGCGG
GCGAGCCGCAGGTCCCTGCCGAGCCGGCCGAGGCGGCCAGCGAACC
CAGTGTCGTCGATGGGCGGGCCTGA

>Bamb_5917_PCP-COM$_{2C}$_aa
MHHHHHHGKPIPNPLLGLDSTENLYFQGIDPFTGAAAGVSAAGIEPDLTA
IWQALFALPAVGRHQDFFALGGDSQLGLRMLAQLRERHGVDLPLRCLYE
APTVARLAETIVRLAAPAPSGDQDDASEYEEGVIR

>Bamb_5917_PCP-COM$_{2C}$_dna
ATGCATCATCACCATCACCATGGTAAGCCTATCCCTAACCCTCTCCT
CGGTCTCGATTCTACGGAAAACCTGTATTTTCAGGGAATTGATCCCT
TCACCGGCGCCGCCGCGGGCGTCTCGGCGGCCGGCATCGAGCCTGA
CCTGACGGCGATCTGGCAGGCGCTGTTCGCCCTGCCGGCGGTTGGC
CGCCACCAGGATTTCTTCGCGCTCGGCGGCGATTCGCAGCTGGGCCT
GCGCATGCTGGCCCAGCTGCGTGAGCGCCACGGCGTCGACCTGCCG
CTGCGTTGCCTCTACGAGGCGCCCACCGTCGCCCGCCTGGCCGAGAC
GATCGTGCGGCTCGCGGCGCCGGCGCCTTCGGGCGACCAGGACGAT
GCATCGGAGTACGAAGAAGGCGTGATCCGCTGA

>Bamb_5915_C_aa
MHHHHHHGKPIPNPLLGLDSTENLYFQGIDPFTMTIPALLSELQARGITLS
LADGELSFRAPKGALTPADRATLSARREAIVAYLAAKAARRTDPVTITPS
AELRPSLLQELWWHWYGLPPRQLNQERLPLVKLFPGVTAGRVAEALRA
IVARHHTLRSSFHEEDGRLTVTLNEAAALPIEFVEADGTLPREELEPALK
AQAAEYAARQLPLDGQWLLRARVVSLAPDQSLLLCVFHHIIVDAASLLLIL
AELDARLADPPRALPAAAQFLDYAAWERAWMADPARQPLIDYWARRF
RALPELVGPLTGRSLAWQPGSKVDHRFVIPAAQLRRMQAAATRLQTSLF
SALLSAFGVALARWSGSERVPVRCVGDLRTSPELANLVGYLVCSDVIEIH
APAKADFVSILKASEIESHSAMMLRVPTLMRHPLHRGGSGIEDPRGIAATI
NMFSVRIPGAGAPLDERADPPWPPQLTRSAGEPWPIPLPSIYLRLIDYGH
ALEGSLELNDTLLTAAEQAALIEALFDALDRFLLQAAPAAAPLTTEVL

>Bamb_5915_C_dna
ATGCATCATCACCATCACCATGGTAAGCCTATCCCTAACCCTCTCCT
CGGTCTCGATTCTACGGAAAACCTGTATTTTCAGGGAATTGATCCCT
TCACCATGACGATTCCCGCCTTGCTGTCCGAGCTGCAGGCGCGCGGC
ATCACGCTGTCCCTGGCCGATGGCGAGCTGTCGTTCCGCGCGCCCAA
GGGCGCGTTGACGCCGGCCGATCGCGCGACGCTGTCGGCGAGGCGC
GAAGCGATCGTTGCCTACCTGGCCGCCAAGGCCGCGCGCCGCACCGA
TCCGGTGACGATCACGCCCTCGGCCGAGCTGCGTCCCTCGCTGCTCC
AGGAACTCTGGTGGCACTGGTACGGCCTGCCGCCGCGCCAGCTGAA
CCAGGAGCGCCTGCCGCTCGTCAAGCTGTTCCCGGGCGTGACGGCC
GGGCGCGTGGCCGAGGCGCTGCGCGCCATCGTGGCCCGCCACCACA

CGCTGCGCTCCTCGTTCCATGAAGAGGACGGCCGGCTGACCGTCACG
CTCAACGAAGCCGCGGCGCTGCCGATCGAGTTCGTCGAGGCCGACG
GCACGCTGCCGCGCGAGGAGCTGGAGCCCGCGCTGAAAGCGCAGGC
GGCCGAGTATGCCGCGCGCCAGCTGCCGCTCGACGGGCAGTGGCTG
CTGCGCGCGCGCGTCGTCTCGCTGGCGCCCGACCAGAGCCTGCTGCT
GTGCGTGTTCCATCACATCATCGTCGACGCGGCCTCGCTGCTGCTGA
TCCTCGCCGAGCTCGACGCGCGGCTGGCCGATCCGCCGCGCGCGCTG
CCAGCCGCCGCGCAGTTCCTCGACTACGCGGCCTGGGAGCGGGCCT
GGATGGCCGATCCCGCGCGTCAGCCGTTGATCGACTACTGGGCGCG
GCGCTTTCGCGCGCTGCCCGAGCTGGTCGGCCCGCTCACGGGTCGCT
CGCTGGCCTGGCAGCCGGGCAGCAAGGTCGACCATCGCTTCGTGAT
TCCGGCCGCGCAGCTGCGCCGCATGCAGGCCGCCGCCACGCGCCTGC
AGACCTCGCTCTTCAGCGCGCTGCTGAGCGCCTTCGGCGTGGCGCTG
GCGCGCTGGTCCGGCAGCGAGCGGGTGCCGGTGCGCTGCGTCGGCG
ACCTGCGCACCTCGCCCGAGCTGGCGAACCTCGTCGGTTACCTGGTC
TGCAGCGACGTGATCGAGATCCACGCGCCGGCCAAGGCCGATTTCG
TGTCGATCCTGAAGGCCAGCGAGATCGAATCGCACAGCGCGATGAT
GCTGCGCGTGCCGACCCTGATGCGCCATCCGCTGCACCGGGGCGGC
AGCGGCATCGAGGATCCGCGCGGCATCGCGGCCACCATCAACATGT
TCTCGGTGCGCATCCCCGGCGCCGGCGCGCCGCTCGACGAGCGCGC
CGATCCGCCCTGGCCGCCGCAGCTGACGCGCTCGGCCGGCGAGCCCT
GGCCGATTCCCTTGCCGTCGATCTACCTGCGGCTGATCGATTACGGC
CACGCGCTGGAAGGTTCGCTGGAACTCAACGACACGCTGCTGACGG
CCGCCGAGCAGGCCGCGCTGATCGAGGCGCTGTTCGATGCGCTCGA
CCGCTTCCTGCTGCAGGCCGCGCCGGCGGCGGCGCCCCTCACCACGG
AGGTTTTATGA

7.2 Protein NMR spectroscopy buffer screening

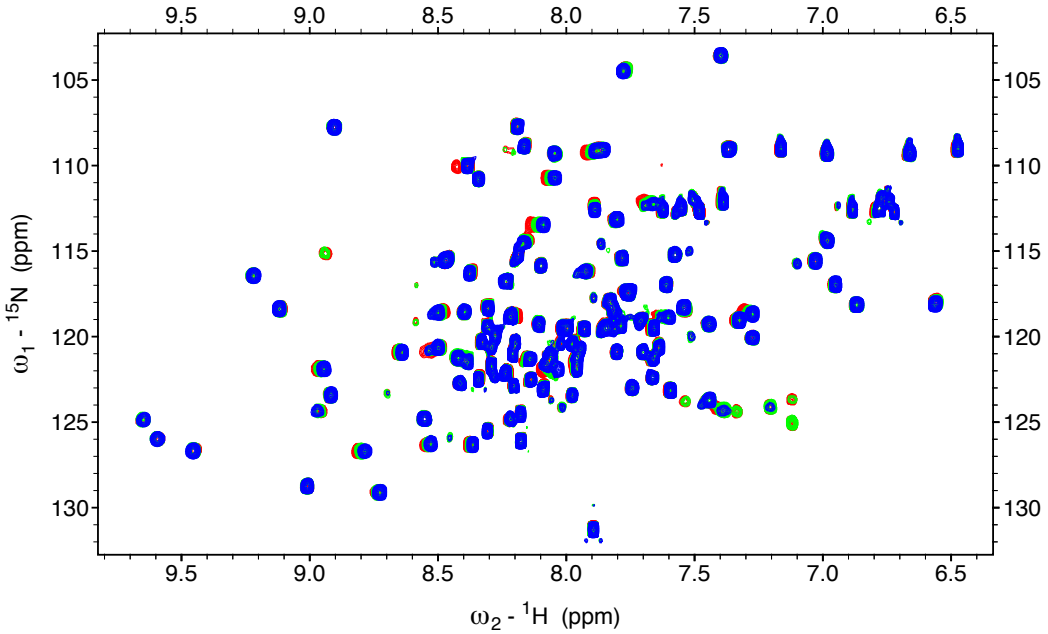

Figure 7.1: HSQC of PCP screening buffer pH from 7.0 (blue) to 5.5 (red).

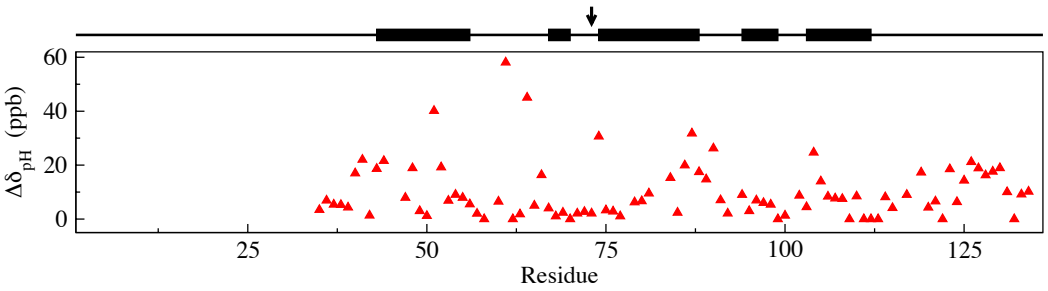

Figure 7.2: Chemical shift perturbations of PCP from pH 6 to 7.

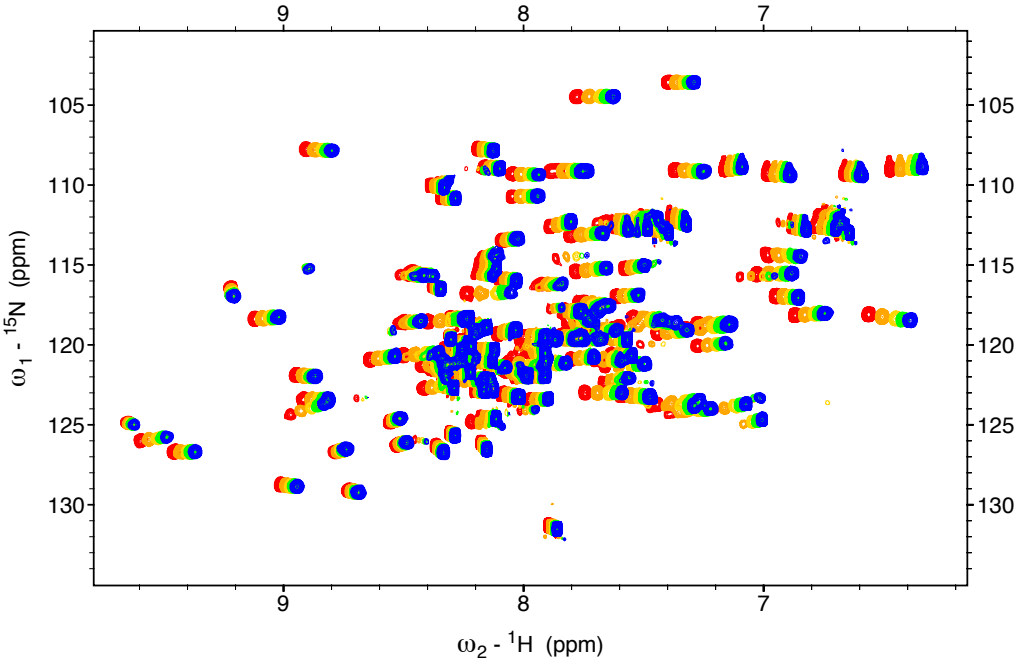

Figure 7.3: HSQC of PCP screening temperature from 10 °C (blue) to 25 °C (red), referenced using DSS.

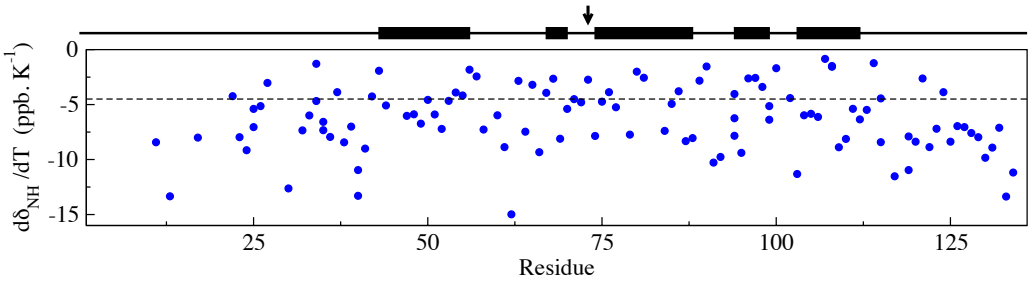

Figure 7.4: Temperature from pH 6 to 7. Values above dashed line are proposed to be involved in H-bonding.

Table 7.1: Chemical shift assignments for PCP-COM$_{2C}$ (in collaboration with Dr Angelo Gallo).

Residue number	Residue type	Atom label	Chemical shift (ppm)
1	GLY	H1	8.067
1	GLY	N	110.713
1	GLY	C	173.714
1	GLY	CA	45.48
1	GLY	HA2	3.913
1	GLY	HA3	3.851
2	ALA	C	177.365
2	ALA	CA	52.714
2	ALA	CB	19.273
2	ALA	H	8.193
2	ALA	HA	4.285
2	ALA	QB	1.345
2	ALA	N	123.662
3	ALA	C	177.497
3	ALA	CA	52.665
3	ALA	CB	19.389
3	ALA	H	8.402
3	ALA	HA	4.272
3	ALA	QB	1.365
3	ALA	N	123.073

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
4	ALA	C	178.179
4	ALA	CA	52.723
4	ALA	CB	19.221
4	ALA	H	8.258
4	ALA	HA	4.264
4	ALA	QB	1.36
4	ALA	N	122.841
5	GLY	C	174.054
5	GLY	CA	45.345
5	GLY	H	8.327
5	GLY	HA2	3.922
5	GLY	HA3	3.881
5	GLY	N	108.095
6	VAL	C	175.977
6	VAL	CA	62.176
6	VAL	CB	32.954
6	VAL	CG1	21.134
6	VAL	CG2	20.753
6	VAL	H	8.049
6	VAL	HA	4.107
6	VAL	HB	2.006
6	VAL	QG1	0.839
6	VAL	QG2	0.884
6	VAL	N	119.898
7	SER	C	174.726
7	SER	CA	57.979
7	SER	CB	64.59
7	SER	H	8.446
7	SER	HA	4.531
7	SER	HB2	3.901
7	SER	HB3	3.97
7	SER	N	120.116
8	ALA	C	178.691
8	ALA	CA	54.975
8	ALA	CB	18.978
8	ALA	H	8.983
8	ALA	HA	3.929
8	ALA	QB	1.35
8	ALA	N	126.735
9	ALA	HA	4.035
9	ALA	QB	1.322
9	ALA	N	119.306
9	ALA	C	179.39
9	ALA	CA	53.997
9	ALA	CB	18.708
9	ALA	H	8.297
10	GLY	C	174.761
10	GLY	CA	45.772
10	GLY	H	7.887
10	GLY	HA2	3.882
10	GLY	HA3	3.921
10	GLY	N	104.837
11	ILE	C	176.623
11	ILE	CA	60.863
11	ILE	CB	39.263
11	ILE	CD1	13.326
11	ILE	CG1	27.074
11	ILE	CG2	17.69
11	ILE	H	7.754
11	ILE	HA	4.171
11	ILE	HB	1.744
11	ILE	QD1	0.633
11	ILE	HG12	1.394
11	ILE	HG13	1.366
11	ILE	QG2	0.803
11	ILE	N	118.79
12	GLU	CA	56.503
12	GLU	CB	30.209
12	GLU	CG	36.411
12	GLU	H	7.66
12	GLU	HB2	1.946
12	GLU	HB3	1.926
12	GLU	HG2	2.267
12	GLU	HG3	2.261
12	GLU	N	114.846
12	GLU	C	175.699
12	GLU	HA	4.196
13	PRO	C	177.145
13	PRO	CA	63.354
13	PRO	CB	32.096
13	PRO	CD	50.925
13	PRO	CG	27.059
13	PRO	HA	4.428
13	PRO	HB2	2.288
13	PRO	HB3	2.257
13	PRO	HD2	3.799
13	PRO	HD3	3.808
13	PRO	HG2	1.893
13	PRO	HG3	1.912
14	ASP	C	177.497
14	ASP	CA	56.05
14	ASP	CB	41.499
14	ASP	H	8.619
14	ASP	HB3	2.805
14	ASP	HB2	2.53

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
14	ASP	N	127.077
14	ASP	HA	4.189
15	LEU	C	177.849
15	LEU	CA	55.161
15	LEU	CB	42.299
15	LEU	CD1	22.304
15	LEU	CD2	22.493
15	LEU	CG	26.091
15	LEU	H	7.804
15	LEU	QD1	0.539
15	LEU	QD2	0.541
15	LEU	HG	0.666
15	LEU	N	121.08
15	LEU	HB2	1.49
15	LEU	HB3	1.506
15	LEU	HA	3.947
16	THR	CG2	21.705
16	THR	H	8.583
16	THR	HA	4.225
16	THR	HB	4.881
16	THR	QG2	1.187
16	THR	N	115.603
16	THR	C	175.685
16	THR	CA	70.166
16	THR	CB	68.405
17	ALA	C	181.418
17	ALA	CA	55.944
17	ALA	CB	18.275
17	ALA	H	7.79
17	ALA	HA	4.151
17	ALA	QB	1.569
17	ALA	N	122.169
18	ILE	C	178.698
18	ILE	CA	65.668
18	ILE	CB	38.964
18	ILE	CD1	13.797
18	ILE	CG1	27.121
18	ILE	CG2	18.681
18	ILE	H	7.748
18	ILE	HA	3.912
18	ILE	HB	2.26
18	ILE	QD1	0.959
18	ILE	HG12	1.563
18	ILE	HG13	1.575
18	ILE	QG2	1.189
18	ILE	N	118.751
19	TRP	C	178.029
19	TRP	CA	62.441
19	TRP	CB	28.781
19	TRP	H	9.166
19	TRP	HA	4.375
19	TRP	HB2	3.439
19	TRP	HB3	3.243
19	TRP	HE1	8.82
19	TRP	N	122.133
19	TRP	NE1	120.909
20	GLN	C	178.179
20	GLN	CA	60.269
20	GLN	CB	28.988
20	GLN	CG	34.524
20	GLN	H	9.232
20	GLN	HA	3.888
20	GLN	HB2	2.208
20	GLN	HB3	2.254
20	GLN	HE21	6.844
20	GLN	HE22	7.129
20	GLN	HG2	2.457
20	GLN	HG3	2.471
20	GLN	N	118.397
20	GLN	NE2	109.445
21	ALA	C	180.416
21	ALA	CA	54.626
21	ALA	CB	18.372
21	ALA	H	7.417
21	ALA	HA	4.257
21	ALA	QB	1.54
21	ALA	N	118.88
22	LEU	C	178.382
22	LEU	CA	57.872
22	LEU	CB	42.78
22	LEU	CD1	25.211
22	LEU	CD2	24.983
22	LEU	CG	27.358
22	LEU	H	8.134
22	LEU	HA	3.895
22	LEU	HB2	1.461
22	LEU	HB3	1.12
22	LEU	QD1	0.601
22	LEU	QD2	0.433
22	LEU	HG	1.524
22	LEU	N	119.625
23	PHE	C	174.225
23	PHE	CA	58.717
23	PHE	CB	40.315
23	PHE	H	7.923

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
23	PHE	HA	4.395
23	PHE	HB2	3.273
23	PHE	HB3	2.494
23	PHE	N	113.301
24	ALA	C	176.118
24	ALA	CA	53.289
24	ALA	CB	16.497
24	ALA	H	7.838
24	ALA	HA	4.047
24	ALA	QB	1.399
24	ALA	N	123.233
25	LEU	C	176.453
25	LEU	CA	51.244
25	LEU	CB	42.153
25	LEU	CD1	23.061
25	LEU	CD2	25.671
25	LEU	CG	26.86
25	LEU	H	8.026
25	LEU	QD1	0.984
25	LEU	QD2	0.874
25	LEU	HG	1.603
25	LEU	N	118.05
25	LEU	HA	5.01
25	LEU	HB3	1.496
25	LEU	HB2	1.524
26	PRO	C	176.12
26	PRO	CA	63.454
26	PRO	CB	32.095
26	PRO	CD	50.962
26	PRO	CG	28.951
26	PRO	HA	4.326
26	PRO	HB2	1.621
26	PRO	HB3	1.608
26	PRO	HD2	3.853
26	PRO	HD3	3.823
26	PRO	HG2	2.135
26	PRO	HG3	2.119
27	ALA	C	175.773
27	ALA	CA	51.39
27	ALA	CB	21.498
27	ALA	H	7.112
27	ALA	HA	4.5
27	ALA	QB	1.208
27	ALA	N	117.162
28	VAL	C	175.431
28	VAL	CA	61.241
28	VAL	CB	35.671
28	VAL	CG1	22.881
28	VAL	CG2	21.512
28	VAL	H	8.252
28	VAL	HA	4.164
28	VAL	HB	1.885
28	VAL	QG1	0.96
28	VAL	QG2	1.213
28	VAL	N	122.003
29	GLY	C	175.944
29	GLY	CA	45.161
29	GLY	H	9.442
29	GLY	HA2	4.144
29	GLY	HA3	3.895
29	GLY	N	117.106
30	ARG	C	173.666
30	ARG	CA	58.076
30	ARG	CB	32.07
30	ARG	CD	43.508
30	ARG	CG	27.107
30	ARG	H	8.563
30	ARG	HA	3.87
30	ARG	HB2	1.514
30	ARG	HB3	1.64
30	ARG	HD2	3.155
30	ARG	HD3	3.126
30	ARG	HG2	1.816
30	ARG	HG3	1.525
30	ARG	N	121.74
31	HIS	C	176.278
31	HIS	CA	54.787
31	HIS	CB	31.145
31	HIS	H	8.277
31	HIS	HA	5.122
31	HIS	HB2	3.336
31	HIS	HB3	2.834
31	HIS	N	113.397
32	GLN	C	174.856
32	GLN	CA	57.42
32	GLN	CB	31.765
32	GLN	CG	36.455
32	GLN	H	7.4
32	GLN	HA	4.282
32	GLN	HB2	2.273
32	GLN	HB3	2.248
32	GLN	HG2	2.275
32	GLN	HG3	2.144
32	GLN	N	120.086
33	ASP	C	175.763

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
33	ASP	CA	54.409
33	ASP	CB	44.501
33	ASP	H	8.745
33	ASP	HA	4.877
33	ASP	HB2	2.836
33	ASP	HB3	2.551
33	ASP	N	126.206
34	PHE	C	175.427
34	PHE	CA	61.873
34	PHE	CB	41.895
34	PHE	H	9.733
34	PHE	HA	4.897
34	PHE	HB2	2.824
34	PHE	HB3	2.887
34	PHE	N	125.94
35	PHE	C	180.49
35	PHE	CA	60.32
35	PHE	CB	39.842
35	PHE	H	8.29
35	PHE	HA	4.181
35	PHE	HB2	3.264
35	PHE	HB3	2.884
35	PHE	N	116.781
36	ALA	C	180.593
36	ALA	CA	55.442
36	ALA	CB	18.328
36	ALA	H	9.196
36	ALA	HA	4.178
36	ALA	QB	1.596
36	ALA	N	129.022
37	LEU	C	175.948
37	LEU	CA	55.831
37	LEU	CB	42.549
37	LEU	CD1	26.164
37	LEU	CD2	20.974
37	LEU	CG	27.415
37	LEU	H	7.769
37	LEU	HA	3.792
37	LEU	HB2	1.476
37	LEU	HB3	1.585
37	LEU	QD1	0.943
37	LEU	QD2	0.498
37	LEU	HG	1.459
37	LEU	N	116.96
38	GLY	C	175.077
38	GLY	CA	44.285
38	GLY	H	7.527
38	GLY	HA2	4.341
38	GLY	HA3	3.641
38	GLY	N	103.697
39	GLY	C	169.934
39	GLY	CA	46.461
39	GLY	H	8.188
39	GLY	HA2	3.582
39	GLY	HA3	3.563
39	GLY	N	109.603
40	ASP	C	176.758
40	ASP	CA	52.462
40	ASP	CB	42.617
40	ASP	H	6.995
40	ASP	HB2	2.865
40	ASP	HB3	3.523
40	ASP	N	118.086
40	ASP	HA	5.092
41	SER	C	175.453
41	SER	CA	62.944
41	SER	CB	62.415
41	SER	H	9.127
41	SER	HA	4.899
41	SER	HB2	4.193
41	SER	HB3	4.112
41	SER	N	115.321
42	GLN	C	179.498
42	GLN	CA	59.51
42	GLN	CB	28.433
42	GLN	CG	34.589
42	GLN	H	7.96
42	GLN	HA	4.07
42	GLN	HB2	2.132
42	GLN	HB3	2.236
42	GLN	HG2	2.459
42	GLN	HG3	2.384
42	GLN	N	120.912
43	LEU	C	179.724
43	LEU	CA	58.323
43	LEU	CB	42.68
43	LEU	CD1	24.62
43	LEU	CD2	25.467
43	LEU	CG	27.163
43	LEU	H	9.064
43	LEU	HA	4.023
43	LEU	HB2	1.708
43	LEU	HB3	1.487
43	LEU	QD1	0.778
43	LEU	QD2	0.269

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
43	LEU	HG	1.57
43	LEU	N	123.543
44	GLY	C	174.746
44	GLY	CA	48.319
44	GLY	H	9.053
44	GLY	N	107.893
44	GLY	HA3	3.143
44	GLY	HA2	3.177
45	LEU	CB	42.606
45	LEU	CA	60.197
45	LEU	C	178.7
45	LEU	HA	3.709
45	LEU	H	8.48
45	LEU	QB	1.647
45	LEU	N	119.16
45	LEU	CG	27.366
45	LEU	CD1	27.211
45	LEU	CD2	26.5
45	LEU	HG	1.567
45	LEU	QD1	0.896
45	LEU	QD2	0.844
46	ARG	C	177.674
46	ARG	CA	56.894
46	ARG	CB	30.279
46	ARG	CD	43.378
46	ARG	CG	27.25
46	ARG	H	8.698
46	ARG	HA	4.203
46	ARG	HB2	1.953
46	ARG	HB3	1.924
46	ARG	HD2	3.154
46	ARG	HD3	3.113
46	ARG	HG2	1.604
46	ARG	HG3	1.563
46	ARG	N	115.649
48	LEU	C	176.974
48	LEU	CA	58.552
48	LEU	CB	42.132
48	LEU	CD1	24.85
48	LEU	CD2	24.17
48	LEU	CG	27.287
48	LEU	HB2	1.504
48	LEU	HB3	1.527
48	LEU	QD1	0.853
48	LEU	QD2	0.786
48	LEU	HG	1.107
48	LEU	HA	3.917
49	ALA	H	7.38
49	ALA	N	118.862
49	ALA	C	175.776
49	ALA	CA	57.76
49	ALA	CB	18.166
49	ALA	HA	4.251
49	ALA	QB	1.551
50	GLN	C	179.56
50	GLN	CA	60.209
50	GLN	CG	34.59
50	GLN	H	8.007
50	GLN	HA	4.087
50	GLN	HB2	1.933
50	GLN	HB3	1.691
50	GLN	HG2	2.384
50	GLN	HG3	2.366
50	GLN	N	119.738
50	GLN	CB	29.54
51	LEU	C	180.421
51	LEU	CA	58.793
51	LEU	CB	42.534
51	LEU	CD1	24.716
51	LEU	CD2	25.244
51	LEU	CG	27.009
51	LEU	H	9.068
51	LEU	QD1	0.773
51	LEU	QD2	0.273
51	LEU	HG	1.596
51	LEU	N	123.409
51	LEU	HA	4.074
51	LEU	HB2	1.929
51	LEU	HB3	1.542
52	ARG	C	178.524
52	ARG	CA	59.162
52	ARG	CB	30.248
52	ARG	CD	43.601
52	ARG	CG	27.091
52	ARG	H	7.82
52	ARG	HA	4.085
52	ARG	HB2	1.901
52	ARG	HB3	1.963
52	ARG	HD2	3.133
52	ARG	HD3	3.143
52	ARG	HG2	0.947
52	ARG	HG3	1.551
52	ARG	N	120.974
53	GLU	C	178.195
53	GLU	CA	59.022

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
53	GLU	CB	30.848
53	GLU	CG	36.08
53	GLU	H	8.57
53	GLU	HB2	1.846
53	GLU	HB3	1.943
53	GLU	HG2	2.229
53	GLU	HG3	2.241
53	GLU	N	121.236
53	GLU	HA	3.867
54	ARG	C	177.66
54	ARG	CA	58.292
54	ARG	CB	29.978
54	ARG	CD	41.562
54	ARG	CG	27.25
54	ARG	H	8.341
54	ARG	HD2	3.002
54	ARG	HD3	3.018
54	ARG	HG2	1.563
54	ARG	HG3	1.602
54	ARG	N	114.593
54	ARG	HA	4.052
54	ARG	HB2	2.045
54	ARG	HB3	2.11
55	HIS	C	175.962
55	HIS	CA	55.997
55	HIS	CB	31.49
55	HIS	H	7.792
55	HIS	HA	4.898
55	HIS	HB2	3.2
55	HIS	HB3	2.893
55	HIS	N	112.431
56	GLY	C	173.542
56	GLY	CA	47.099
56	GLY	H	7.984
56	GLY	HA2	4.057
56	GLY	HA3	3.888
56	GLY	N	109.208
57	VAL	C	173.57
57	VAL	CA	61.476
57	VAL	CB	33.774
57	VAL	CG1	21.115
57	VAL	CG2	20.868
57	VAL	H	6.647
57	VAL	HA	4.059
57	VAL	HB	1.68
57	VAL	QG1	0.825
57	VAL	QG2	0.856
57	VAL	N	118.368
58	ASP	C	174.063
58	ASP	CA	52.871
58	ASP	CB	42.203
58	ASP	H	8.928
58	ASP	HA	4.698
58	ASP	HB2	2.899
58	ASP	HB3	2.252
58	ASP	N	129.077
59	LEU	CA	57.518
59	LEU	CB	42.449
59	LEU	CD1	26.44
59	LEU	CD2	26.37
59	LEU	CG	31.378
59	LEU	H	8.754
59	LEU	QD1	0.788
59	LEU	QD2	0.804
59	LEU	HG	1.453
59	LEU	N	125.028
59	LEU	C	174.232
59	LEU	HA	4.72
59	LEU	QB	1.692
60	PRO	C	177.251
60	PRO	CA	61.423
60	PRO	CB	32.435
60	PRO	CD	50.23
60	PRO	CG	27.628
60	PRO	HA	4.565
60	PRO	HB2	2.272
60	PRO	HB3	2.13
60	PRO	HD2	3.616
60	PRO	HD3	3.795
60	PRO	HG2	2.053
60	PRO	HG3	2.002
61	LEU	C	178.762
61	LEU	CA	58.434
61	LEU	CB	41.648
61	LEU	CD1	21.484
61	LEU	CD2	23.336
61	LEU	CG	27.345
61	LEU	H	8.46
61	LEU	HA	4.045
61	LEU	HB2	1.505
61	LEU	HB3	1.527
61	LEU	QD1	0.832
61	LEU	QD2	0.899
61	LEU	HG	1.463
61	LEU	N	121.866

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
62	ARG	C	176.644
62	ARG	CA	60.973
62	ARG	CB	31.827
62	ARG	CD	43.619
62	ARG	CG	27.25
62	ARG	H	8.625
62	ARG	HA	4.169
62	ARG	HB2	1.766
62	ARG	HB3	1.646
62	ARG	HD2	3.191
62	ARG	HD3	3.231
62	ARG	HG2	1.571
62	ARG	HG3	1.606
62	ARG	N	115.716
63	CYS	C	178.35
63	CYS	CA	63.505
63	CYS	CB	27.895
63	CYS	H	7.904
63	CYS	HA	4.111
63	CYS	HB3	2.718
63	CYS	N	115.28
63	CYS	HB2	2.913
64	LEU	C	176.973
64	LEU	CA	57.422
64	LEU	CB	43.349
64	LEU	CD1	22.751
64	LEU	CD2	26.003
64	LEU	CG	26.505
64	LEU	H	7.59
64	LEU	HA	4.135
64	LEU	HB2	1.001
64	LEU	HB3	1.028
64	LEU	QD1	0.392
64	LEU	QD2	-0.291
64	LEU	HG	1.019
64	LEU	N	119.43
65	TYR	C	178.889
65	TYR	CA	61.578
65	TYR	CB	38.913
65	TYR	H	7.092
65	TYR	HA	4.159
65	TYR	HB2	2.899
65	TYR	HB3	2.995
65	TYR	N	114.368
66	GLU	C	175.791
66	GLU	CA	57.106
66	GLU	CB	30.066
66	GLU	CG	36.294
66	GLU	H	7.727
66	GLU	HA	4.124
66	GLU	HB2	2.116
66	GLU	HB3	1.926
66	GLU	HG2	2.296
66	GLU	HG3	2.29
66	GLU	N	115.187
67	ALA	CA	50.403
67	ALA	CB	18.955
67	ALA	H	7.525
67	ALA	N	123.657
67	ALA	C	172
67	ALA	HA	4.626
67	ALA	QB	1.199
68	PRO	C	176.283
68	PRO	CA	64.699
68	PRO	CB	29.413
68	PRO	CD	50.795
68	PRO	CG	27.002
68	PRO	HA	4.88
68	PRO	HB2	2.238
68	PRO	HB3	2.278
68	PRO	HD2	3.849
68	PRO	HD3	3.747
68	PRO	HG2	1.92
68	PRO	HG3	1.993
69	THR	C	173.541
69	THR	CA	58.885
69	THR	CB	72.072
69	THR	CG2	22.137
69	THR	H	7.49
69	THR	HA	5.223
69	THR	HB	4.941
69	THR	QG2	1.066
69	THR	N	109.086
70	VAL	C	177.642
70	VAL	CA	67.695
70	VAL	CB	31.614
70	VAL	CG1	20.858
70	VAL	CG2	21.466
70	VAL	H	9.863
70	VAL	HA	2.759
70	VAL	HB	1.47
70	VAL	QG1	0.494
70	VAL	QG2	-0.4
70	VAL	N	125.007
71	ALA	C	181.623

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
71	ALA	CA	55.761
71	ALA	CB	18.598
71	ALA	H	8.668
71	ALA	HA	3.829
71	ALA	QB	1.12
71	ALA	N	118.558
72	ARG	C	179.378
72	ARG	CA	61.435
72	ARG	CB	30.038
72	ARG	CD	43.942
72	ARG	CG	27.25
72	ARG	H	8.066
72	ARG	HA	4.1
72	ARG	HB2	1.94
72	ARG	HB3	2.008
72	ARG	HD2	3.152
72	ARG	HD3	3.192
72	ARG	HG2	1.516
72	ARG	HG3	1.535
72	ARG	N	116.205
73	LEU	C	178.903
73	LEU	CA	58.268
73	LEU	CB	41.709
73	LEU	CD1	23.93
73	LEU	CD2	21.918
73	LEU	CG	26.86
73	LEU	H	9.616
73	LEU	HA	4.071
73	LEU	HB2	1.611
73	LEU	HB3	1.644
73	LEU	QD1	0.892
73	LEU	QD2	0.804
73	LEU	HG	1.881
73	LEU	N	126.772
74	ALA	C	178.827
74	ALA	CA	55.473
74	ALA	CB	19.266
74	ALA	H	8.608
74	ALA	HA	3.817
74	ALA	QB	1.41
74	ALA	N	120.544
75	GLU	C	178.884
75	GLU	CA	59.5
75	GLU	CB	29.656
75	GLU	CG	36.398
75	GLU	H	7.133
75	GLU	HA	3.928
75	GLU	HB2	2.113
75	GLU	HB3	2.083
75	GLU	HG2	2.223
75	GLU	HG3	2.203
75	GLU	N	115.585
76	THR	C	175.447
76	THR	CA	67.079
76	THR	CB	68.924
76	THR	CG2	21.505
76	THR	H	7.899
76	THR	HA	3.84
76	THR	HB	4.503
76	THR	QG2	1.229
76	THR	N	117.929
77	ILE	C	177.794
77	ILE	CA	64.451
77	ILE	CB	36.447
77	ILE	CD1	18.27
77	ILE	CG1	27.265
77	ILE	CG2	18.914
77	ILE	H	8.109
77	ILE	HA	3.888
77	ILE	HB	2.051
77	ILE	QD1	1.195
77	ILE	HG12	1.656
77	ILE	HG13	1.518
77	ILE	QG2	1.124
77	ILE	N	121.307
78	VAL	C	178.36
78	VAL	CA	65.69
78	VAL	CB	32.045
78	VAL	CG1	22.909
78	VAL	CG2	21.465
78	VAL	H	7.924
78	VAL	HA	3.606
78	VAL	HB	2.113
78	VAL	QG1	0.968
78	VAL	QG2	0.928
78	VAL	N	118.038
79	ARG	C	178.54
79	ARG	CA	59.253
79	ARG	CB	30.15
79	ARG	CD	43.505
79	ARG	CG	27.129
79	ARG	H	7.873
79	ARG	HA	4.097
79	ARG	QB	1.924
79	ARG	HD2	3.202

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
79	ARG	HD3	3.192
79	ARG	HG2	1.824
79	ARG	HG3	1.854
79	ARG	N	119.615
80	LEU	C	177.83
80	LEU	CA	56.483
80	LEU	CB	39.032
80	LEU	CD1	26.397
80	LEU	CD2	24.552
80	LEU	CG	27.641
80	LEU	H	7.967
80	LEU	HA	4.209
80	LEU	HB2	1.743
80	LEU	HB3	1.938
80	LEU	QD1	0.819
80	LEU	QD2	0.774
80	LEU	HG	1.58
80	LEU	N	118.439
81	ALA	C	176.98
81	ALA	CA	52.667
81	ALA	CB	19.26
81	ALA	H	7.73
81	ALA	HA	4.277
81	ALA	QB	1.416
81	ALA	N	121.635
82	ALA	C	175.079
82	ALA	CA	50.806
82	ALA	CB	18.394
82	ALA	H	7.77
82	ALA	N	123.696
82	ALA	HA	4.556
82	ALA	QB	1.348
83	PRO	C	176.462
83	PRO	CA	62.951
83	PRO	CB	31.847
83	PRO	CD	50.694
83	PRO	CG	27.397
83	PRO	HA	4.431
83	PRO	HB2	1.891
83	PRO	HB3	1.861
83	PRO	HD2	3.799
83	PRO	HD3	3.608
83	PRO	HG2	2.005
83	PRO	HG3	1.99
84	ALA	C	175.597
84	ALA	CA	50.575
84	ALA	CB	18.812
84	ALA	H	8.544
84	ALA	N	125.926
84	ALA	HA	4.559
84	ALA	QB	1.378
85	PRO	C	176.974
85	PRO	CA	62.754
85	PRO	CB	32.097
85	PRO	CD	50.593
85	PRO	CG	27.274
85	PRO	HA	4.427
85	PRO	HB2	1.879
85	PRO	HB3	2.271
85	PRO	HD2	3.615
85	PRO	HD3	3.784
85	PRO	HG2	2.001
85	PRO	HG3	2.047
86	SER	C	175.327
86	SER	CA	58.631
86	SER	CB	64.28
86	SER	H	8.6
86	SER	HA	4.426
86	SER	HB2	3.895
86	SER	HB3	3.914
86	SER	N	116.735
87	GLY	C	173.897
87	GLY	CA	45.512
87	GLY	H	8.53
87	GLY	HA2	3.967
87	GLY	HA3	3.994
87	GLY	N	110.977
88	ASP	C	175.771
88	ASP	CA	54.573
88	ASP	CB	41.269
88	ASP	H	8.279
88	ASP	HA	4.548
88	ASP	HB2	2.696
88	ASP	HB3	2.617
88	ASP	N	122.061
89	GLN	C	175.509
89	GLN	CA	55.722
89	GLN	CB	29.331
89	GLN	CG	33.948
89	GLN	H	8.463
89	GLN	HA	4.305
89	GLN	HB2	1.957
89	GLN	HB3	2.119
89	GLN	HE21	6.562
89	GLN	HE22	7.327

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
89	GLN	HG2	2.338
89	GLN	HG3	2.313
89	GLN	N	120.265
89	GLN	NE2	108.786
90	ASP	C	176.286
90	ASP	CA	54.544
90	ASP	CB	41.368
90	ASP	H	8.364
90	ASP	HA	4.872
90	ASP	HB2	2.676
90	ASP	HB3	2.595
90	ASP	N	120.407
91	ASP	C	176.128
91	ASP	CA	54.559
91	ASP	CB	41.406
91	ASP	H	8.373
91	ASP	HA	4.562
91	ASP	HB2	2.681
91	ASP	HB3	2.583
91	ASP	N	121.152
92	ALA	C	178.02
92	ALA	CA	52.731
92	ALA	CB	19.199
92	ALA	H	8.351
92	ALA	HA	4.275
92	ALA	QB	1.355
92	ALA	N	124.781
93	SER	C	174.737
93	SER	CA	59.205
93	SER	CB	64.086
93	SER	H	8.36
93	SER	HA	4.332
93	SER	HB2	3.804
93	SER	HB3	3.84
93	SER	N	115.307
94	GLU	C	176.126
94	GLU	CA	56.475
94	GLU	CB	30.791
94	GLU	CG	36.417
94	GLU	H	8.407
94	GLU	HA	4.199
94	GLU	HB2	1.836
94	GLU	HB3	1.925
94	GLU	HG2	2.08
94	GLU	HG3	2.148
94	GLU	N	122.316
95	TYR	C	175.604
95	TYR	CA	57.943
95	TYR	CB	39.082
95	TYR	H	8.16
95	TYR	HA	4.517
95	TYR	HB2	2.897
95	TYR	HB3	2.993
95	TYR	N	120.721
96	GLU	C	175.951
96	GLU	CA	56.449
96	GLU	CB	30.519
96	GLU	CG	36.467
96	GLU	H	8.262
96	GLU	HA	4.203
96	GLU	HB2	1.938
96	GLU	HB3	1.854
96	GLU	HG2	2.181
96	GLU	HG3	2.2
96	GLU	N	123.573
97	GLU	C	177.152
97	GLU	CA	57.318
97	GLU	CB	30.223
97	GLU	CG	36.427
97	GLU	H	8.526
97	GLU	HA	4.13
97	GLU	HB2	1.951
97	GLU	HB3	2.006
97	GLU	HG2	2.26
97	GLU	HG3	2.277
97	GLU	N	122.867
98	GLY	C	173.876
98	GLY	CA	45.433
98	GLY	H	8.577
98	GLY	HA2	3.919
98	GLY	HA3	3.886
98	GLY	N	110.376
99	VAL	C	175.777
99	VAL	CA	62.631
99	VAL	CB	32.023
99	VAL	CG1	21.138
99	VAL	CG2	21.138
99	VAL	H	7.824
99	VAL	HA	4.046
99	VAL	HB	1.919
99	VAL	QG1	0.841
99	VAL	QG2	0.843
99	VAL	N	119.904
100	ILE	C	175.248
100	ILE	CA	62.055

Table 7.1: (continued)

Residue number	Residue type	Atom label	Chemical shift (ppm)
100	ILE	CB	38.294
100	ILE	CD1	12.716
100	ILE	CG1	27.574
100	ILE	CG2	17.647
100	ILE	H	8.409
100	ILE	HA	4.047
100	ILE	HB	1.798
100	ILE	QD1	0.794
100	ILE	HG12	1.137
100	ILE	HG13	1.154
100	ILE	QG2	0.86
100	ILE	N	126.7
101	ARG	C	180.817
101	ARG	CA	58.299
101	ARG	CB	31.713
101	ARG	CD	43.679
101	ARG	CG	27.25
101	ARG	H	8.118
101	ARG	HD2	3.209
101	ARG	HD3	3.178
101	ARG	HG2	1.567
101	ARG	HG3	1.51
101	ARG	N	131.766
101	ARG	HB2	1.774
101	ARG	HB3	1.642
101	ARG	HA	4.136

7.3 dd-containing biosynthetic pathways

7.3.1 Natural products from biosynthetic pathways using dd2 domains

Aculeximycin, Akaeolide, Ansamitocin, Apoptolidin, Bafilomycin, Barbamide, BE-14106, Calcimycin, Chalcomycin, Chaxamycin, Chlorothricin, Concanamycin, Coronafacic, Cremimycin, Dihydrochalcomycin, Divergolide, E-837, Ebelactone, Elaiophylin, FK520, Halstoctacosanolide, Herbimycin, Herboxidiene, Hitachimycin, Hygrocin, Incednine, Indanomycin, Kendomycin, Kijanimicin, Lasalocid, Leucanicidin, Lipomycin, Lobophorin, Lorneic, Macbecin, Meilingmycin, Methymycin, ML-449, Mycinamicin, Nanchangmycin, Naphthomycin, Natamycin, Nemadectin, Neoaureothin, Niddamycin, Nystatin, Nystatin-like, Piericidin, Pimaricin, Pladienolide, Polyoxypeptin, Reveromycin, Rifamycin, Rubradirin, Salinomycin, Sanglifehrin, Spinosad, Spirangien, Stambomycin, Streptolydigin, Tautomycin, Tetrocarcin, Tiacumicin, Tirandamycin, Tylactone, Versipelostatin.

7.3.2 Natural products from biosynthetic pathways using dd3 domains

Ambruticin, Anatoxin, Aurafuron, Bacillaene, Chlorizidine, Crocacin, Cylindrospermopsin, Enacyloxin, FK228, Gobichelin, Gulmirecin, JBIR-34, Jerangolid, Microsclerodermins, Nannocystin, Nostophycin, Oxazolomycin, Pellasoren, Pyralomicin, Spiruchostatin, Stigmatellin, Tetronomycin.

Figure 7.5: Word cloud of genera from Atlas PKS/NRPS database that contain COM_{2N} domains, where font size corresponds to the number of occurrence (created using http://tagcrowd.com).

7.4 COM_{2N} example biosynthetic genes

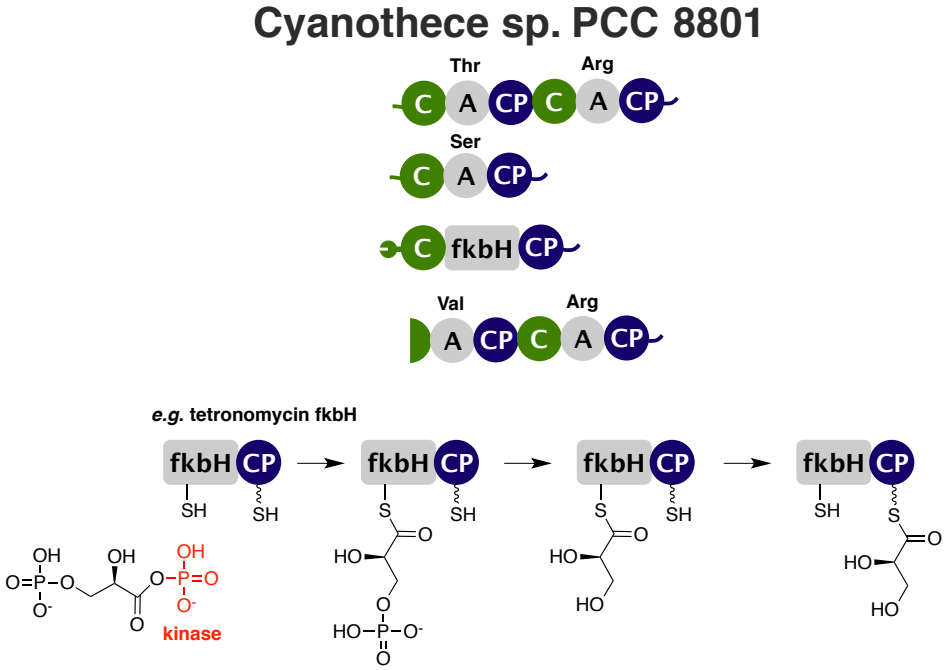

Figure 7.6: Example COM2-boundary: *Cyanothece* (WP_012596902.1).

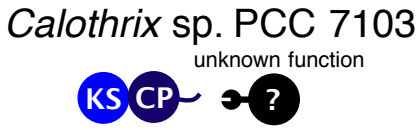

Figure 7.7: Example COM2-boundary: *Calothrix*

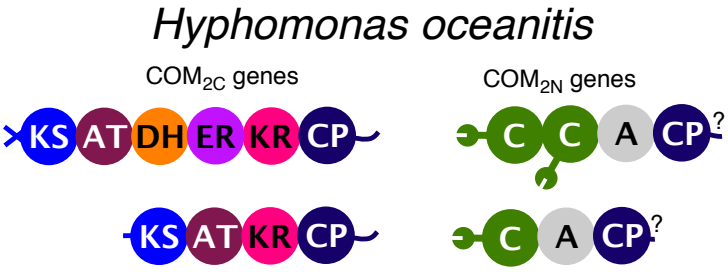

Figure 7.8: Example COM2-boundary: *Hyphomonas* (WP_051624978.1).

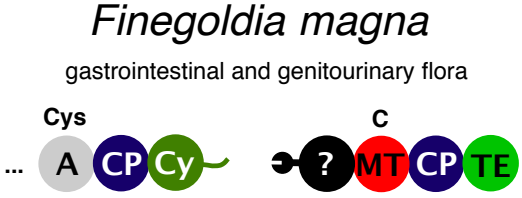

Figure 7.9: Example COM2-boundary: *Finegoldia* (WP_002841792.1).

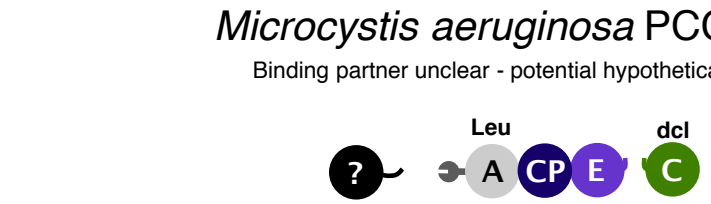

Figure 7.10: Example COM2-boundary: *Microcystis*

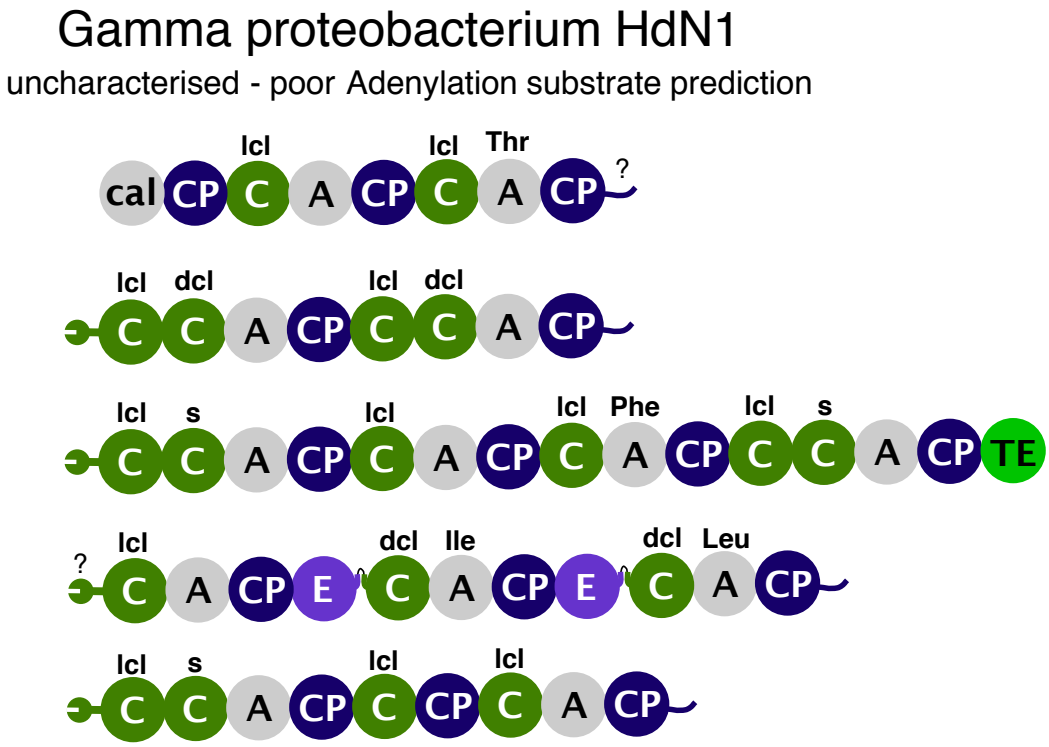

Figure 7.11: Example COM2-boundary: Gamma proteobacterium HdN1 (WP_013261592.1).

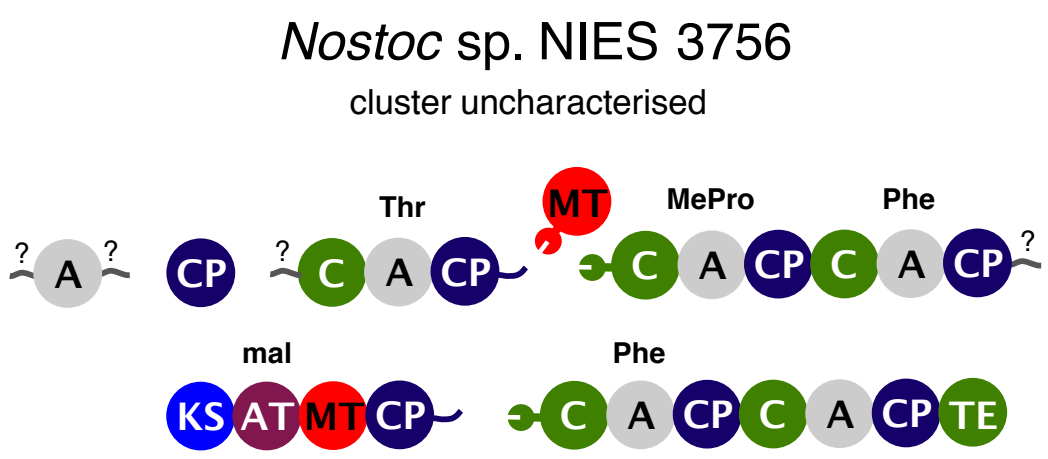

Figure 7.12: Example COM2-boundary: *Nostoc* (WP_067768737.1).

Crocosphaera watsonii

cluster uncharacterised - poor sequencing

Figure 7.13: Example COM2-boundary: *Crocosphaera* (WP_007303634.1).

Bradyrhizobium sp. ARR65

uncharacterised

Figure 7.14: Example COM2-boundary: *Bradyrhizobium* (WP_024513186.1).

7.4.1 COM_{2N} proteins from Atlas PKS/NRPS database

AAF19812.1, AAY37654.1, AAY91421.3, AAY92759.1, ABA21236.1, ABA51274.1, ABA74009.1, ABC29733.1, ABC29735.1, ABE54463.1, ABF86893.1, ABF90173.1, ABF90459.1, ABF92546.1, ABF92546.1, ABG19543.1, ABG52864.1, ABN93222.1, ABN93233.1, ABP55261.1, ABS45963.1, ABW31348.1, ABW31348.1, ABW31350.1, ABW31356.1, ABX04504.1, ABX04731.1, ACB52391.1, ACB75253.1, ACC81024.1, ACC84845.1, ACC85408.1, ACK67008.1, ACK67644.1, ACL21978.1, ACL76665.1, ACR31662.1, ACV01940.1, ACV02548.1, ACV29783.1, ACY61368.1, ADC88013.1, ADO71805.1, ADU21412.1, ADU72106.1, ADV97464.1, AEB42527.1, AEI63747.1, AEO73201.1, AEX42750.1, AEX46387.1, AEX68229.1, AEX72947.1, AEY65152.1, AEY66938.1, AFI69976.1, AFJ01561.1, AFK57214.1, AFN46007.1, AFT70785.1, AFW93815.1, AFY33172.1, AFY80136.1, AFY88055.1, AFY89103.1, AFY89103.1, AFY92098.1, AFZ14010.1, AFZ26071.1, AFZ33016.1, AFZ33307.1, AFZ36194.1, AFZ59099.1, AFZ60772.1, AFZ60782.1, AFZ60782.1, AGC45618.1, AGC46260.1, AGC47284.1, AGC47287.1, AGI79641.1, AGL83959.1, AGL85301.1, AGP37466.1, AGP37474.1, AGP40742.1, AGP88177.1, BAG05480.1, BAG05822.1, BAI74310.1, BAK43323.1, BAK92291.1, CAE15905.1, CAH22534.1, CAH39107.1, CAH39108.1, CAK15504.1, CAN00352.1, CAW25429.1, CAW93975.1, CAY48789.1, CBG67371.1, CBJ11325.1, CBL44580.1, CBL45097.1, CBL45102.1, CBL45103.1, CCC86198.1, CCE74273.1, CCH32769.1, EAU66331.1,
NP_252915.1, NP_486681.1, NP_486682.1, NP_792634.1, NP_962675.1, NP_994709.1, XP_004532053.1, YP_001061877.1, YP_001076246.1, YP_001346261.1, YP_001394897.1, YP_001394901.1, YP_001719506.1, YP_001778907.1, YP_001804456.1, YP_001865727.1, YP_001866574.1, YP_001873844.1, YP_002345832.1, YP_002373799.1, YP_002471866.1, YP_002471870.1, YP_002505322.1, YP_002912099.1, YP_003073775.1, YP_003101411.1, YP_003139384.1, YP_003192079.1, YP_003266191.1, YP_003339432.1, YP_003566894.1, YP_003678979.1, YP_003678979.1, YP_003678979.1, YP_003678979.1, YP_003678979.1, YP_003678979.1, YP_003678979.1, YP_003678979.1, YP_003679695.1, YP_003712686.1,

YP_003712687.1, YP_003844094.1, YP_003954443.1, YP_004664826.1, YP_004664826.1, YP_004668885.1, YP_004668888.1, YP_004668980.1, YP_004712293.1, YP_004801211.1, YP_004974482.1, YP_005263991.1, YP_005503890.1, YP_005519936.1, YP_005761026.1, YP_006269092.1, YP_006277307.1, YP_006278375.1, YP_006480704.1, YP_006657687.1, YP_006659040.1, YP_006876063.1, YP_006979112.1, YP_006996375.1, YP_007051204.1, YP_007059070.1, YP_007066010.1, YP_007066010.1, YP_007097198.1, YP_007097198.1, YP_007121020.1, YP_007121020.1, YP_007121351.1, YP_007130468.1, YP_007137551.1, YP_007159684.1, YP_007359089.1, YP_007361940.1, YP_007361943.1, YP_007375312.1, YP_007396976.1, YP_007512510.1, YP_007786329.1, YP_007786384.1, YP_008134859.1, YP_008174055.1, YP_008174456.1, YP_008198470.1, YP_008327602.1, YP_008500999.1, YP_110608.1, YP_234878.1, YP_235654.1, YP_273985.1, YP_322128.1, YP_325330.1, YP_335852.1, YP_347945.1, YP_371431.1, YP_440021.1, YP_519054.1, YP_608605.1, YP_623238.1, YP_653281.1, YP_777794.1, YP_788894.1, YP_838624.1

7.5 Vibroxin biosynthetic genes

Table 7.2: Proposed functions of genes in the vibroxin biosynthetic gene cluster based on sequence similarity to genes in the cluster directing enacyloxin biosynthesis in *B. ambifaria* AMMD

Gene	Putative function	Homologous gene	Percent identity
vbxA	FAD_dependent_oxidoreductase	*bamb_5928*	68
vbxB	Aspartyl/Asparaginyl_beta-hydroxylase	*bamb_5927*	71
vbxC	Thioesterase	*bamb_5926*	45
vbxD	beta-ketoacyl_synthase	*bamb_5925*	48
vbxE	beta-ketoacyl_synthase	*bamb_5924*	49
vbxF	beta-ketoacyl_synthase	*bamb_5923*	46
vbxG	beta-ketoacyl_synthase	*bamb_5922*	46
vbxH	beta-ketoacyl_synthase	*bamb_5921*	50
vbxI	beta-ketoacyl_synthase	*bamb_5920*	47
vbxJ	beta-ketoacyl_synthase	*bamb_5919*	51
vbxK	NADH:flavin_oxidoreductase/NADH_oxidase	*bamb_5918*	71
vbxL	Xylose_isomerase_domain_protein_TIM_barrel	*bamb_5912*	59
vbxM	Shikimate_dehydrogenase	*bamb_5913*	53
vbxN	Alchohol_dehydrogenase,_zinc-binding_domain	*bamb_5914*	57
vbxO	Condensation_domain_protein	*bamb_5915*	49
vbxP	L-carnitine_dehydratase/bile_acid-inducible	*bamb_5916*	67
vbxQ	Phosphopantetheine-binding_domain	*bamb_5917*	36
vbxR	LuxR_transcriptional_family_regulator	*bamb_5911*	54
vbxS	hypothetical_protein	*bamb_5929*	51
vbxT	MATE_efflux_family_protein	*bamb_5933*	55

7.6 NMR spectra for vibroxin

Spectra of 0.3 mg vibroxin were acquired on 500 MHz NMR spectrometer in d-$_4$MeOH (Fig.7.15-7.19).

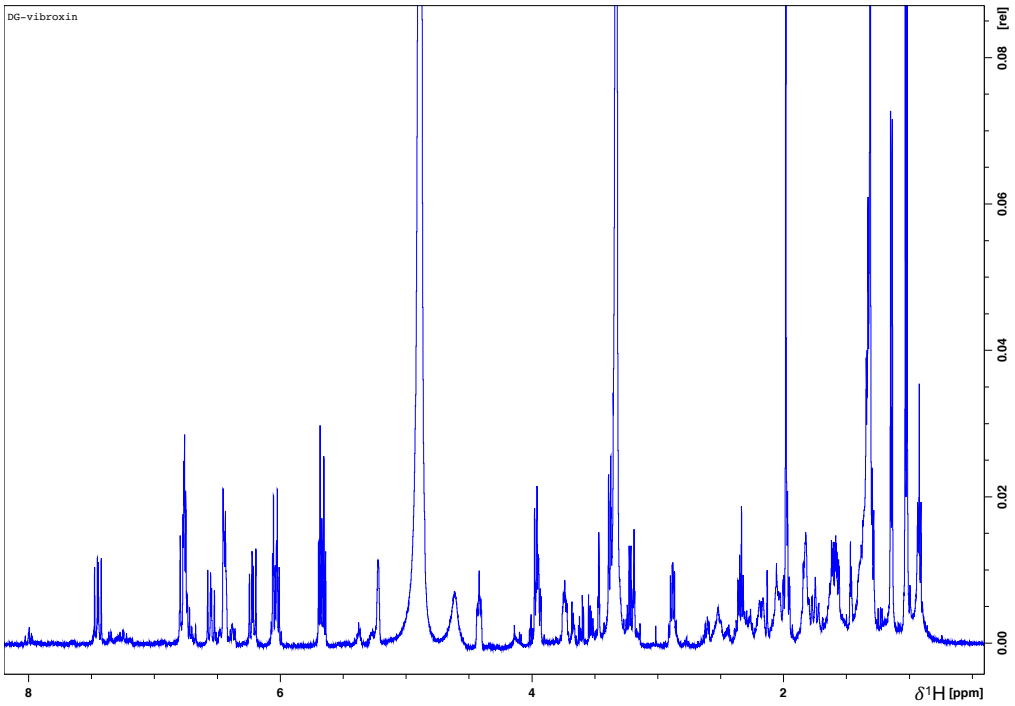

Figure 7.15: ^{1}H spectrum.

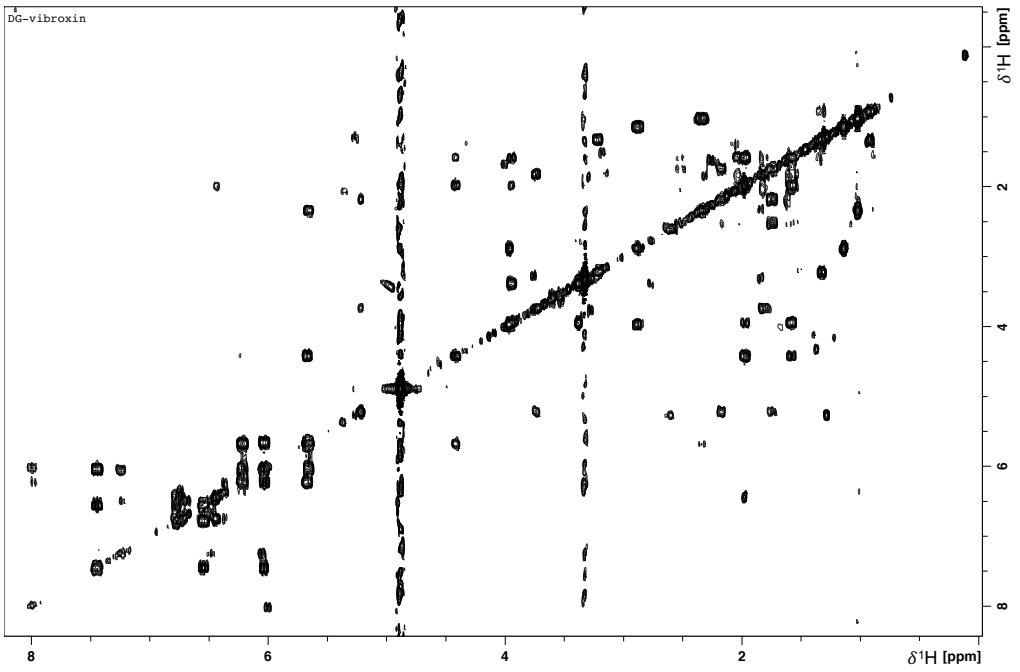

Figure 7.16: COSY spectrum.

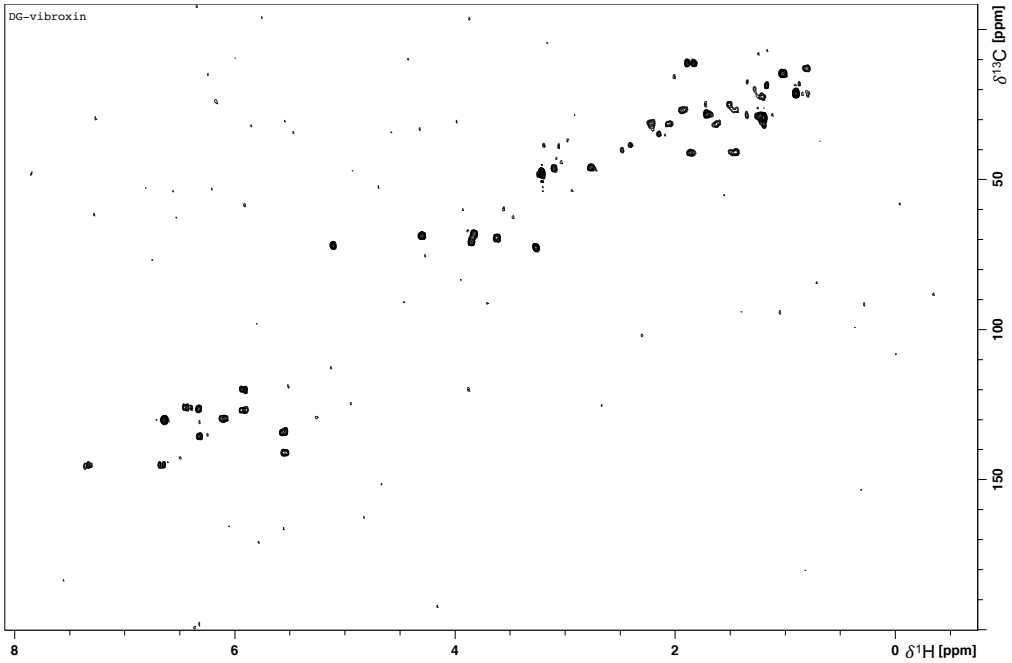

Figure 7.17: ^{1}H-^{13}C HSQC spectrum.

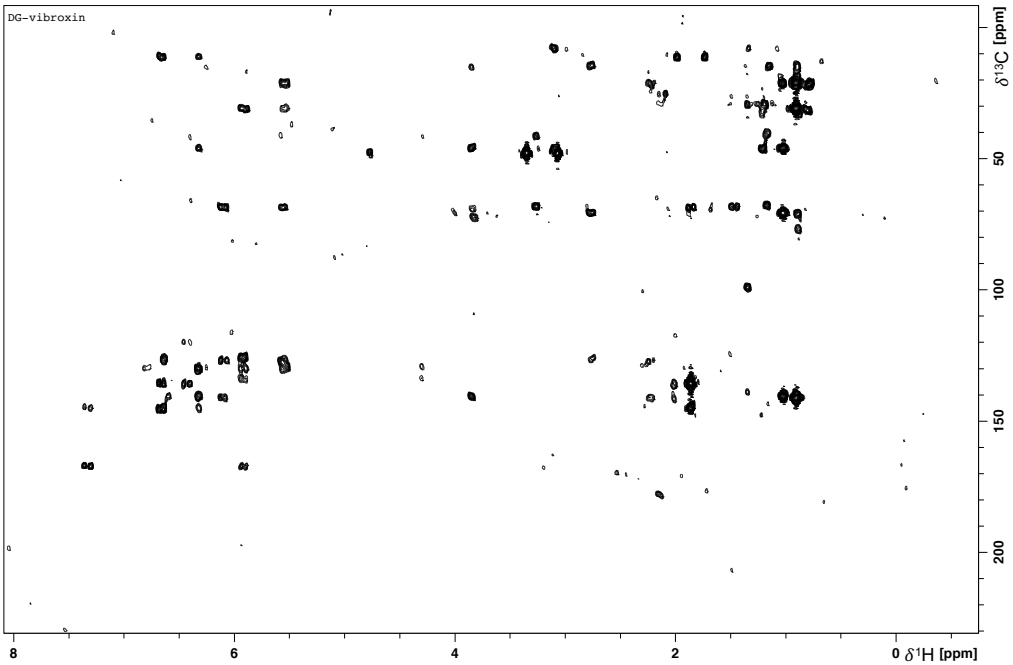

Figure 7.18: ^{1}H-^{13}C HMBC spectrum

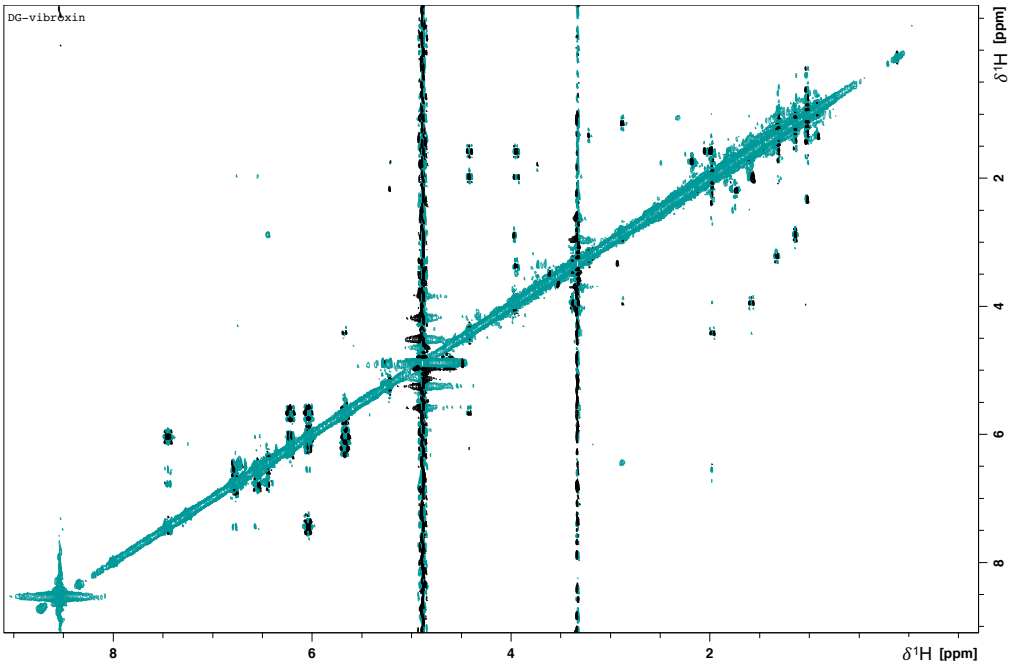

Figure 7.19: NOESY spectrum with 100 ms mixing time.