Master’s thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Fredrik Aunaas Fossheim

Constructing Metabolic Pathways from Identified Biosynthetic Gene Clusters

Master’s thesis in Chemical Engineering and Biotechnology
Supervisor: Eivind Almaas
June 2020

NTNU
Norwegian University of
Science and Technology

Constructing Metabolic Pathways from Identified Biosynthetic Gene Clusters

Fredrik Aunaas Fossheim

June 2020

MASTER THESIS
Department of Biotechnology and Food Science
Norwegian University of Science and Technology

Supervisor 1: Eivind Almaas
Supervisor 2: Snorre Sulheim

Acknowledgement

I would like to thank my supervisors Eivind Almaas and Snorre Sulheim for invaluable input and constructive feedback during work on this thesis. In addition I would like to thank my fathers side of the family (Terje Fossheim et al.) for sparking my interest for nature sciences in general, and my mothers side (Helga Aunaas et al.) for a particular interest in both biotechnology and industrial chemistry, leading me onto the path that I have completed today.

This paper is a master's thesis in systems biology. It was written over the course of 23 weeks, building upon a project report written in the fall of 2019. Some parts of the theory and introduction are copied/heavily influenced by this project report. In these cases, the text in question is put in quotation marks, and followed by ”(Fossheim - Project report, 2019)[1]”

Abstract

We have in this work developed, and implemented, an algorithm that converts information about predicted biosynthetic gene clusters (BGCs) as provided by antiSMASH into metabolic pathways for use in genome-scale metabolic models (GEMs). The accuracy of the algorithm is evaluated through a detailed comparison with experimentally determined pathways for eight BGCs. We report an overall 82% average accuracy for polyketide synthase (PKS) and nonribosomal peptide synthase (NRPS) domains in general, resulting from a 78 % accuracy in substrate specificity for extender units, and 84% accuracy for cofactor-associated reactions. With this algorithm, we have also constructed metabolic pathways for all T1PKS, transAT-PKS and NRPS BGCs that exist in the MIBiG database.

Based on smCOG definitions, we were able to predict the synthesis of the uncommon extender unit methoxymalonyl-ACP. From other smCOG definitions, there was also established a relationship between the number of detected glycosyltransferases in a BGC, and the number of glycosyl groups that took part in the metabolic pathway of the secondary metabolite. Two other tailoring reactions were found to be predictable by the same means. For tailoring reactions that are not included in the constructed metabolic pathways, we attempt to elucidate the consequence of these.

We also discuss the different obstacles one faces when attempting to construct metabolic pathways from BGCs, as well as those of modeling secondary metabolism in general. We end by suggesting that the SubClusterBLAST functionality of antiSMASH is expanded to include additional known tailoring reactions that are found for PKS/NRPS. In addition, we suggest updating the databases used for prediction of NRPS/PKS module specificity so that the predictions that antiSMASH makes - and in turn the metabolic pathways that the algorithm produces - are more true to their real life counterparts.

The project is available from
`https://github.com/FredrikFossheim/MasterThesis`

Sammendrag

Vi har i dette arbeidet utviklet og implementert en algoritme som konverterer informasjon om predikerte biosyntetiske genklustere (BGC'er) som gitt av antiSMASH til metabolske reaksjonsveier for bruk i genomskalamodeller (GEM'er). Nøyaktigheten til algoritmen blir evaluert gjennom en detaljert sammenligning med eksperimentelt bestemte metabolske reaksjonsveier for åtte BGCer. Vi rapporterer 82 % gjennomsnittlig nøyaktighet for PKS - og NRPS-domener generelt, som følge av 78% nøyaktighet i substratspesifisitet for forlenger-enheter, og 84 % nøyaktighet for kofaktorassosierte reaksjoner. Med denne algoritmen har vi også konstruert metabolske veier for alle T1PKS, transAT-PKS og NRPS BGCer som finnes i MIBiG-databasen.

Basert på smCOG-definisjoner, var vi i stand til å forutsi syntese av den uvanlige forlenger-enheten metoksymalonyl-ACP. Fra andre smCOG-definisjoner ble det etablert en sammenheng mellom antall påviste glykosyltransferaser i en BGC, og antallet glykosylgrupper som deltok i den metabolske reaksjonsveien til sekundærmetabolitten. To andre tilleggsreaksjoner kunne også forutses på samme vis. For tilleggsreaksjoner som ikke er inkludert i de konstruerte metabolske reaksjonsveiene, prøver vi å belyse konsekvensen av dette.

Vi diskuterer også de forskjellige hindringene man står overfor når man prøver å konstruere metabolske reaksjonsveier fra BGCer, samt utfordringer man møter ved modellering av sekundærmetabolisme generelt. Vi avslutter med å foreslå at SubClusterBLAST-funksjonaliteten til antiSMASH utvides til å omfatte ytterligere kjente tilleggsreaksjoner som finnes for PKS/NRPS. I tillegg foreslår vi å oppdatere databasene som brukes til prediksjon av NRPS/PKS modulspesifisitet, slik at prediksjonene som antiSMASH gir - og dermed de metabolske reaksjonsveiene som algoritmen produserer - blir mer tro til sine reelle motparter.

Prosjektet er tilgjengelig fra
`https://github.com/FredrikFossheim/MasterThesis`

Table of Contents

Abbreviations

Domain abbreviations	
A	AMP-binding
ACP	Acyl carrier protein
AT	Acyl transferase
B	Branching
C	Condensaton
CAL	Coenzyme A ligase
Cy	Heterocyclisation
DH	Dehydratase
E	Epimerisation
ECH	Enoyl-Coa hydratase/isomerase
ER	Enoyl reductase
F	Formylation
GNAT	GCN5-related N-acetyl transferase
KR	Keto reductase
KS	Keto synthase
MT	Methyltransferase
oMT	Oxygen-methyltransferase
nMT	Nitrogen-methyltransferase
cMT	Carbon-methyltransferase
Ox	Oxidation
TE	Thioesterase

Frequently referenced metabolites	
BHT	Beta-hydroxy-Tyrosine
DHPG	Dihydroxy-Phenyl-Glycine
HPG	Hydroxyphenyl-glycine
SAM	S-adenosyl-L-Methinoine
SAH	S-adenosyl-L-Homocysteine
THF	Tetrahydrofolate

Other abbreviations:	
antiSMASH	antibiotics & Secondary Metabolite Analysis Shell
BGC	Biosynthetic gene cluster
FBA	Flux balance analysis
GEM	Genome-scale metabolic model
MIBiG	Minimum Information about a Biosynthetic Gene cluster
NRPS	Nonribosomal peptide synthase
PKS	Polyketide synthase
smCOG	secondary metabolite Cluster of Orthologous Groups of proteins
T1PKS	Type 1 Polyketide synthase
transAT-PKS	Trans-Acyl Transferase polyketide synthase

Chapter 1

Introduction

Natural products is the major source of new compounds with antitumor, anticancer and antibiotic activities [2–5]. In a time where resistance to antibiotics is a growing problem, the discovery of such new products becomes increasingly important. These natural products are often found as secondary metabolites encoded by biosynthetic gene clusters (BGCs) - collections of genes that through evolution are found co-localized on the genome [6]. In addition to encoding the metabolic pathway of the secondary metabolite, the BGC also contains genes necessary for its regulation and transport. These properties observed for BGCs make transconjugation and expression of BGCs in heterologous host organisms an intriguing subject [1, 7].

However, as a result of improved sequencing technology, sequence databases grow at an exponential rate while protein family databases are growing at a near constant rate [8]. Due to readily available detection of BGCs from this abundance of sequenced genomes, many BGCs have been found that are either not expressed, or are expressed in undetectable amounts. [9, 10]. These "silent" BGCs may still encode the synthesis of secondary metabolites with bioactivities such as those mentioned previously. However, in order to determine a secondary metabolite's therapeutic properties, it must first be produced by the organism on a scale that allows for its isolation and subsequent analysis [11]. One part of this challenge is to ensure that the host organism can provide all the necessary precursor metabolites that are required by the metabolic pathway, and in sufficient amounts [1].

To achieve this, genome-scale metabolic models (GEMs) are promising tools. A GEM is a comprehensive overview of an organism's metabolic repertoire, represented by a set of transport and enzyme-coding genes, the associated metabolic reactions, and their reactants and products[12]" (Fossheim - Project report, 2019) [1]. By leveraging linear programming, they open for suggestions on strain development such as gene deletion, insertion as well as up-or down-regulations of certain genes without the need for data collection through real-world experiments [13, 14]. However, efficient use of these models require that the metabolic pathways catalyzed by the enzymes encoded in the BGCs are known.

Several tools have been developed for the detection and analysis of BGCs [15]. A few of these tools can also predict some details of the function, substrates or products of encoded enzymes and can therefore aid in constructing the metabolic pathway of a BGC, e.g. RODEO can identify Ribosomally synthesised and posttranslationally modified peptides (RiPPs) [16], NRPSpredictor targets Non-ribosomal peptide synthases (NRPS) [17], and transATor works for trans acyl transferase polyketide synthase (transAT PKS) [18]. In addition to these, the antibiotics & Secondary Metabolite Analysis Shell (antiSMASH) has been developed, for the purpose of detecting BGCs in general [19–23]. However, except for well-known BGCs (for which the pathway has been experimentally determined), using GEMs is not readily possible because it is non-trivial to translate all genes identified (by e.g. antiSMASH) into metabolic pathways with well-defined reaction substrates and products. In addition, those that have leveraged GEMs as a means of genetic engineering of these BGCs have found that such work requires non-traditional GEM approaches as secondary metabolism does not agree well with the pseudo-steady state assumptions that are typically made when modeling primary metabolism [24].

Despite all the tools available for their detection and analysis, the term “secondary metabolite” still comprises a wide variety of molecules; to date there have been characterised 52 types of BGC products [23, 25]. The abundance of different BGC products adds to the complexity associated with constructing metabolic pathways, and as each type has their own distinctiveness, a reduction of scope is necessary for the task at hand. Out of these 52, the mechanism behind seven types of BGC is so well-studied that predictions on the pathway can be made with accuracy: Non ribosomal peptide synthase (NRPS), type 1 polyketide synthase (T1PKS), type 2 polyketide synthase (T2PKS), trans acyl transferase polyketide syntase (transAT-PKS), lanthipeptides, lassopeptides and thiopeptides [22]. Out of these seven, NRPS, T1PKS and transAT-PKS represent around 80% of identified clusters [25]. In addition to producing natural products, NRPS and PKS BGCs also exhibit promising targets for use as metabolic pipelines for biosynthesis of designer therapeutics. This, however requires a further insight into the mechanism behind the synthesis of natural products. [26] On the basis of these observations, focus has been directed towards constructing the metabolic pathway of NRPS, T1PKS and transAT-PKS BGCs as closely as possible.

To date 1932 BGC have been characterised and indexed in the Minimum Information of Biosynthetic Gene clusters (MIBiG) database, through crowdsourcing[27]. In addition, the database provides antiSMASH output for each of the experimentally characterised BGCs, allowing for comparison of experimental data to prediction based data. Out of the characterized BGCs in MIBiG, 636 have been found in the genus *Streptomycetes*[27], echoing the assertion that the majority of all currently known antibiotics and other therapeutic compounds are derived from Streptomycetes [28]. To illustrate the scale of this number, the second most frequently annotated species is *Aspergillus* with 88 entries. In addition, *S. coelicolor* is a well studied organism, for which there have been made high quality GEMs[29]. *S. coelicolor* will therefore be used as a reference GEM throughout this work.

Most efforts put into systems biology analysis of BGCs has had the opposite approach than what is presented here - they predict the structure of the secondary metabolite to make a guess on its bioactive properties [17, 18, 23]. The focus of *this* project has been on the ability to more rapidly be able to express a BGC in order to analyze its properties experimentally.

The scope of this project is to develop a tool that can construct the metabolic pathway of NRPS and PKS BGCs using antiSMASH output as the only source of information, laying emphasis on the core structure of the secondary metabolites in question. Rather than predicting other elements of the pathway, focus has been put on elucidating the steps of the pathway that are difficult to predict or cannot be predicted with current methods - and to what degree these affect the metabolic pathway of a BGC when they are disregarded.

Chapter 2

Theory

As for all genes, the genes of BGCs are expressed through the central dogma of biology - transcription and translation. This process involves the transcription of DNA to RNA, and subsequently the translation of RNA to protein. These translated proteins - enzymes - are then responsible for catalysing all reactions necessary to synthesise the secondary metabolite. Defining properties of BGCs are the observed collinearity and subclustering of genes, meaning that there is a correlation between the physical location of the genes in the BGC and their relation to each other. A near complete review of all transAT-PKS by Piel et. al. (2016) found that 46 out of 54 BGCs encoding transAT-PKSs exhibited such collinearity [30]. This also holds true for NRPS and T1PKS [31, 32]. BGCs also vary widely in size, from the smaller RiPPs at around 5 kilo-basepairs (kBP) and up to a few hundred kBP for the larger NRPS and PKS. The relatively gigantic sizes of BGCs, and utility of secondary metabolites can perhaps be best described by the BGC synthesising patellazoles A, B and C, first observed in the tunicate (colloquially reffered to as ”sea squirts”) *Lissoclinum patella*. Within *L. patella* microbiome, the symbiont α-proteobacteria *Endolissoclinum faulkner* was found to be the putative producer of the secondary metabolite. Further studies revealed a highly reduced 1.5 Mbp genome size of *E. faulkner*, with 150 Kbp reserved for encoding the synthesis of the secondary metabolite [30]. By synthesizing the secondary metabolite, *E. faulkner* had found a new home in the sea squirt which benefited from the secondary metabolite. In return, the sea squirt provided nutrients to the organism, allowing the massive reduction in genome size. Still, 10% of *E. faulkner*s total genome size remains dedicated to maintaining this beneficial relationship.

2.1 General structure and function of BGCs

Although this story can highlight the significance of secondary metabolites in organisms, there is a more pragmatic reason for the size of NRPS and PKS BGCs which can be attributed to ”core genes” - the backbone of a BGC. Core genes are one of three types of genes responsible for the synthesis of secondary metabolites, the other being ”tailoring genes” and ”extender unit synthesis genes”. A short description of each type of genes is given in the following list:

- **Core genes** synthesise the core structure of the secondary metabolite - A polyketide (PKS) or polypeptide (NRPS). In addition there are found hybrid PKS and NRPS systems where the core structure contains both ketide and peptide bonds. The core structure is synthesized from various ”extender units” - molecules that can be viewed as building blocks. Extender units are usually malonyl-CoA and its α-substituents, such as methylmalonyl-CoA and ethylmalonyl-CoA for PKS[33], and proteinogenic amino acids for NRPS[17].

- **Extender unit synthesis genes** are responsible for creating more uncommon extender units. While malonyl-CoA and methylmalonyl-CoA are molecules that are otherwise used for fatty acid synthesis in the primary metabolism, other extender units such as methoxymalonyl-ACP and hydroxy-phenyl-glycine (HPG) can also be required for the synthesis of polyketides and nonribosomal peptides, respectively[17, 33]. These are not found in the primary metabolism of most organisms, and must therefore be synthesised for the sole purpose of creating a secondary metabolite.

- **Tailoring genes** have a variety of functions, such as synthesising and adding other molecules to the core structure *after* it has been synthesized by the core genes. They are also in many cases responsible for post synthesis structural modifications to the core structure such as cyclisation reactions, reductions, epoxidations and halogenisations[34–36].

One of the more interesting BGC types are PKS and NRPS. Out of the thousands of polyketides known, 1% of them exhibit bioactivity, which is 5 times more than the average of other natural products [37]. However, the main point of interest lies in the modular structure of the core biosynthetic genes, which allows for reliable prediction of molecular structure of the secondary metabolite [19]. T1PKS, transAT PKS and NRPS all share some common features, and also have their own distinct characteristics that differentiate them from one another. Further in the text, T1PKS will be used as a reference for how BGCs work in general. Then, transAT-PKS and NRPS are presented, and compared to T1PKS. As a visual example to help illustrate the concept behind the three types of genes that are found in their synthesis, the synthesis of the T1PKS secondary metabolite bafilomycin B1 is given in Figure 2.1.

2.2 Core genes in Polyketide synthase

Core genes are the defining property of NRPS and PKS BCGs. These genes collectively translate into mega enzymes consisting of several protein gene products. Further, these mega enzymes produce the core structure from extender units in an assembly line fashion. The reason for the assembly line analogy is due to the modular structure of the mega enzymes - a set of domains that follow specific rules with respect to their sequence. According to the collinearity rule, the number and the order of the modules represents the number and the order of extender units in the final product [39]. Different module structures can be seen in the example given for the bafilomycin BGC in Figure 2.1: BfmA2

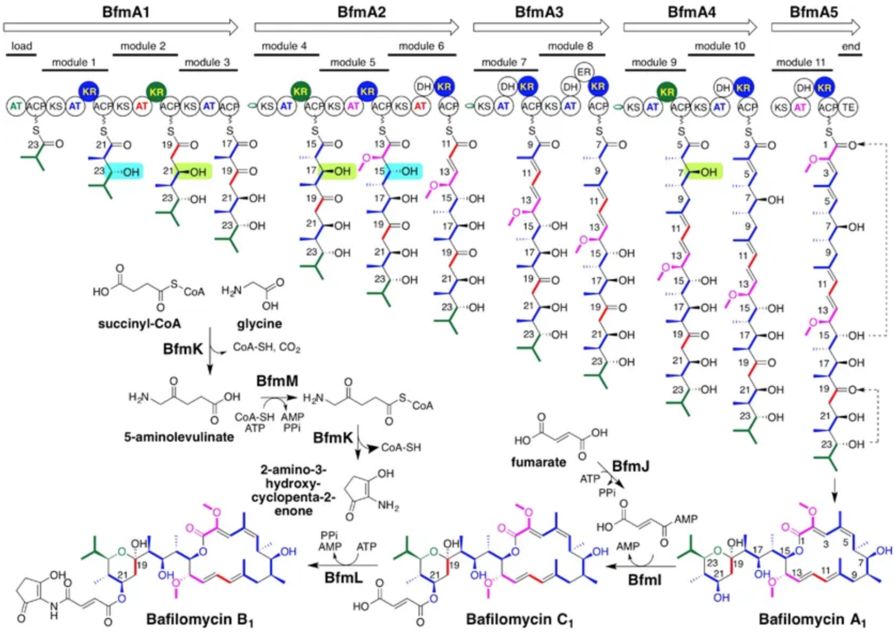


Figure (2.1) ”The biosynthetic pathway of bafilomycin B1 in *Kitasatospora setae* KM-6054. Enzymes BfmA1-BfmA5 are the translated products of the core genes of the cluster and synthesise the core structure (Bafilomycin A1) from various extender units. The molecular origin of each atom in the core structure can be seen from the color coding of the molecular structure of bafilomycin A1. Blue, green, magenta and red represent atoms derived from the extender units methylmalonyl-CoA, isobutyryl-CoA, methoxymalonyl-ACP and malonyl-CoA respectively. Enzymes BfmI, BfmJ, BfmK, BfmL, and BfmM constitute the tailoring genes, producing the end product Bafilomycin B1 from succinyl-CoA, glycine, fumarate and Bafilomycin A1. Extender unit synthesis genes are not explicitly shown in the figure, although they are present in the Bafilomycin BGC. In this case they synthesise the methoxymalonyl-ACP extender unit. Figure was collected from [38].”(Fossheim - Project report, 2019)[1].

encodes a protein product consisting of three modules (module 4, module 5 and module 6). The domain sequence of the three modules can be seen in Table 2.1.

Table (2.1) Domain structure of each module for the protein product encoded by the BfmA2 gene in the Bafilomycin B1 BGC.

Module #	Domain sequence
Module 4	KS-AT-KR-ACP
Module 5	KS-AT-KR-ACP
Module 6	KS-AT-DH-KR-ACP

The function of core genes can best be explained by reference to their hierarchical structure, which at the bottom of the hierarchy is a sequence of domains - sites on a translated core gene that carry the catalytic activity necessary to perform the reactions that are required for synthesising a secondary metabolite. Throughout the history of PKS and NRPS clusters, there have been found a variety of different domains, which each have highly specific functions depending on their type. The most common domains for PKS are given below.

- **Acyltransferase** (AT) - Each step of polyketide synthesis begins with the loading of an extender unit *by* an AT-domain *onto* the Phosphopantetine (Ppant) prosthetic group of the acyl carrier protein (ACP) domain located within the same module [40].

- **Acyl carrier protein** (ACP) - The ACP domain then facilitates transport of both the extender unit and polyketide intermediate between all other domains in the module [40].

- **Keto synthase** (KS) - The KS domain appends an extender unit onto the polyketide intermediate through a klaisen condensation reaction, giving off CO_2 as the condensate and Coenzyme A (CoA) as co-products [40].

- **Keto reductase** (KR) - Reduces the carbonyl group resulting from the reaction catalyzed by the KS to an hydroxyl group. The reaction requires NADPH + H^+ [40].

- **Dehydratase** (DH) Eliminates the β hydroxyl group resulting from the KR domain to form an α - β double CC bond. The reaction releases H_2O [40].

- **Enoyl reductase** (ER) Reduces the α - β double CC bond resulting from the reaction catalyzed by the DH domain to a fully reduced beta methylene group. The reaction requires NADPH + H^+ as a cofactor [40].

- **Methyltransferases** are a group of domains that are differentiated from one another by which type of atom they methylate. Carbon methyltransferases (cMT) methylate the β-C on the polyketide intermediate. Oxygen methyltransferases (oMT) also exist, which act on the hydroxyl group of a reduced keto group on the α carbon of the polyketide intermediate. The last type of methyltransferase are the nitrogen methyltransferases found in NRPS modules. The methyl group is in all these cases provided by (S)-adenosyl-L-methinoine (SAM), leaving its demethylated counterpart - (S)-adenosyl-L-homocysteine (SAH) - after the reaction [40].

- **Thioesterase** (TE) Decouples the polyketide intermediate from ACP. This reaction gives off H_2O [40].

- **Enoyl-CoA hydratase/isomerase** (ECH) either catalyzes the isomerization of a 3E-enoyl-CoA to 2E-enoyl-CoA, or the hydration of a double bond on 2E-enoyl-CoA [41]. For the purposes of this work, all ECH domains are assumed to be isomerases, as the two types cannot be readily differentiated with current methods.

An overview of the reactions that the PKS domains translates to are shown in Table 2.2 [38]. As a consequence of this, the substrates required for this part of the biosynthesis of the secondary metabolite can be obtained by knowing the specificity of AT-domains and observing the presence of e.g reducing and methylating domains.

Table (2.2) Substrates and products associated with each domain in the T1PKS genes. ACP is listed as not having any co-reactants, but requires an Ppant prosthetic group, supplied by Acetyl-CoA. However, this is not a reactant for the production of one polyketide, and is therefore omitted in this list. Table and caption is collected from the project report (Fossheim - Project report, 2019) [1].

Domain	Substrate	Product
AT	-	-
KS	Acyl-CoA	CO2 + CoA
ER	NADPH + H+	NADP+
DH	-	H2O
KR	NADPH + H+	NADP+
MT	SAM	SAH
ACP	-	-
ECH	-	-
TE	H2O	PK

In general, there are 3 types of modules for PKS - one loader module, several extender modules, and finally one terminating module (Table 2.3) [42]. Inconsistent naming in the literature has lead to load modules sometimes being referred to as starter modules, and terminating modules as end modules.

Table (2.3) Possible module compositions for PKS in T1PKS. Domains in brackets are optional and reduce the carbonyl group resulting from the condensation reaction from the KS domain. Table and caption is collected from the project report (Fossheim - Project report, 2019) [1].

Name	domain sequence
Load	AT-ACP-
Extender	-KS-AT-[DH-ER-KR-cMT]-ACP-
End	-KS-AT-[DH-ER-KR-cMT]-ACP-TE

2.2.1 Load module

”The first module in the T1PKS is the load module, which consists of an acyl transferase (AT) domain, and an acyl carrier protein (ACP) domain [43, 44]. Other variations on the loader module do exist, but these are observed for both NRPS and PKS and are presented separately. The AT-domain contains an amino acid signature that is specific for the starter unit. This starter unit can be a wide variety of carboxylates bound to an acyl carrier - usually Coenzyme A (CoA) [45]. When the starter unit is recruited by the AT domain it is transferred to the Ppant prosthetic group on the ACP domain, derived from CoA [42]. The ACP then transfers this starter unit to the KS domain upstream, where the PK synthesis continues through the extender modules. The process is illustrated in Figure 2.2 A.” (Fossheim - Project report, 2019) [1]

Yadav et al. (2003) manually curated the specificity of each AT domain of 321 experimentally determined PKS pathways. The specificity of the first AT domain (i.e. the AT domain of the loader module) of each PKS is given in Table 2.4 [46]. The main observation is that there is large variety in the different starter units, while at the same time, Malonyl-CoA and methylmalonyl-CoA are vastly overrepresented as the starter substrate.

Table (2.4) Substrate specificity of the first AT domain of 321 experimentally determined PKS. CHC-CoA: cyclohexene-l-carboxyl-CoA, Trans-1,2-CPDA: trans-1,2-cyclopentanedicarboxylic acid

# of AT domains	Substrate specificity
193	Malonyl-CoA
104	Methylmalonyl-CoA
5	Ethylmalmalonyl-CoA
4	Isobutyryl-CoA
3	Methoxymalonyl-ACP
3	2-metylbutyryl-CoA
3	Propionyl-CoA
1	CHC-CoA
1	Trans-1,2-CPDA
1	Benzoyl-CoA
1	Acetyl-CoA
1	3-methylbutyryl-CoA
1	Inactive

2.2.2 Extender module

”In a process identical to the mechanism in the load module, an extender unit is loaded onto the AT-domain, which is subsequently bound to the prosthetic Ppant group of ACP. The extender unit is then transferred to the KS-domain neighboring upstream, where it is condensed onto the polyketide intermediate. The process is shown in Figure 2.2. From the KS-domain, the polyketide intermediate is transferred by ACP to all optional domains present in the module (DH, KR, ER, MT, ECH) where the reactions previously described take place. The resulting modifications are illustrated in Figure 2.3. Note that for a DH-domain to be able to perform any operation on the polyketide intermediate, a KR-domain needs to be present in the module, in order to first reduce the carbonyl group into a hydroxyl group. Likewise, the ER-domain is dependent on there being both a KR and DH-domain present. The different configurations and their products are shown in Table 2.5. The table conveys the same information as Figure 2.3. Note also that these domains do not act on the most recently added extender unit, but on the extender unit that was added in the *previous* KS domain, as shown in Figure 2.1. After the reduction, the PK intermediate is transferred onto the next KS domain. This KS domain can either be part of another extender module, or the end/terminating module. The end module operates exactly like the extender module, but in addition contains the TE domain. This domain releases the fully formed polyketide by breaking the thioester bond between ACP and polyketide [42].” (Fossheim - Project report, 2019) [1]

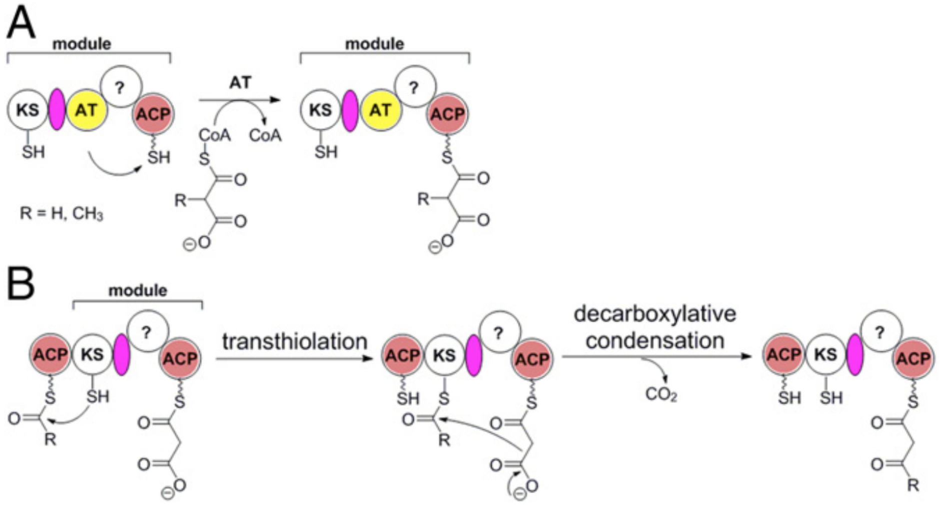


Figure (2.2) ”A) Mechanism of an AT domain recruiting an acyl unit onto Ppant prosthetic group of ACP. B) Mechanism behind an extender unit being appended to a polyketide intermediate. AT and KS domains are responsible for recruiting and appending extender units - Building blocks - to the polyketide. Figure gathered from [47]” (Fossheim - Project report, 2019) [1]

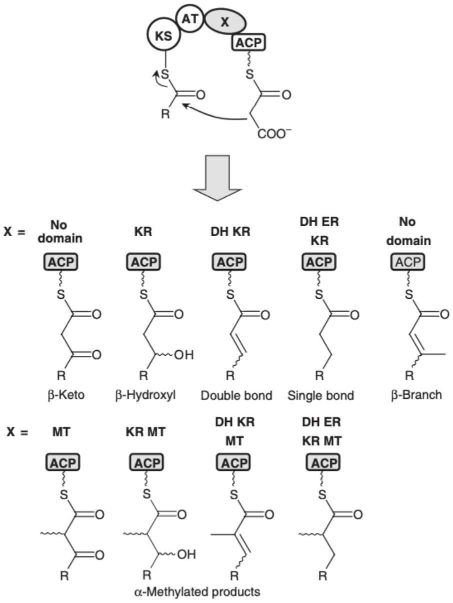


Figure (2.3) How different non-essential domains affect the polyketide structure. Keto reductase (KR) from keto to hydroxyl, Dehydratase (DH) from hydroxyl to double bond, and Enoyl reductase (ER) from double bond to single bond reduce the β-carbon atom. Methyltransferase (MT) methylates the α-carbon atom. Figure gathered from [33]

”Which extender unit is added onto the PK intermediate in each step can be predicted by examining the amino acid sequence of each AT-domain for specific amino acid signatures [19]. The most common extender units are malonyl-CoA and methylmalonyl-CoA [4, 48], while some more uncommon extender units are ethylmalonyl-CoA, and methoxymalonyl-ACP. In rare cases, other extender units are used. [49, 50]. These extender units are described later. For the uncommon extender units, AT domain specificity is usually less specific. For example, the synthesis of the secondary metabolite JBIR-100 can incorporate both methoxymalonyl-ACP and malonyl-CoA as the extender unit in a certain extension step, because both extender units will be accepted by the AT-domain[51].” (Fossheim - project report, 2019) [1]

Table (2.5) ”Possible conformations of tailoring domains in modular T1PKS, and their effect on the PK structure.” Table and caption gathered from (Fossheim - project report, 2019) [1]

Domains	Acts on	Results in	End structure
-	-	carbonyl	=O
KR	carbonyl	Hydroxyl	-OH
DH - KR	Hydroxyl	*Trans* double bond	C=C
DH - ER - KR	*Trans* double bond	Saturated acyl chain	C-C

2.3 Trans-AT PKS core genes

Trans-AT PKS share many similarities with their cis-AT counterparts. Every single type of domain that is observed in Cis-AT PKS can be found in transAT-PKS BGCs (i.e. KS, AT, DH, KR, ER, ACP, ECH and TE domains). However, transAT-PKS' display some features that are unheard of for Cis-AT PKSs. These include: irregularly and frequently placed ACP domains, modules split across two genes (from here on referred to as bridging modules), seemingly unexplainable domain activity and inactive KS domains (KS_0 domains) [30].

2.3.1 Module structure

The main mechanism behind the synthesis of transAT-PKS' is highly similar to that of T1PKS. However, instead of AT-domains being present in every module, the transAT-PKS contain docking sites that free-standing AT domains can attach to and mediate the recruitment of an extender unit. In theory, this would imply that different extender units can be used to synthesize transAT polyketides. However, this is not the observed case - this extender unit is in nearly all cases malonyl-CoA[30]. This rule holds so true that it is common to assume that the extender unit for a transAT-PKS module always is malonyl-CoA [23]. A recent review by Piel et al. (2015) of transAT-PKS found that only 2 (oxazolomycin and kirromycin) out of all the 54 known transAT-PKS incorporated non-malonyl-CoA extender units [30]. For kirromycin and oxazolomycin this activity is observed by the BGC encoding three AT domains, one of which is specific to the unusual extender unit. For kirromycin this extender unit is ethylmalonyl-coa and is one of 16 extender extender units (the rest being malonyl-CoA). For oxazolomycin the extender unit is methoxymalonyl-ACP and is one of the in total 11 extender units (again, the rest being malonyl-CoA) [30].

Although bridging modules are frequently observed in transAT-PKS, one specific type of bridging module is almost always found to be non-extending based on its domain sequence - in this module, the KS-domain is found on the first gene, while the next gene containins a DH domain and an ACP-domain as their first two domains. These ”pseudo-modules” are referred to as Dehydratase-docking (DHD) -modules. Examples of BGCs containing such modules are: Difficidin [52] (BGC0000176), Thailandamide [53] (BGC0000186), Bacillaene [54] (BGC0001089). As for any rule, there are found exeptions to this one as well, such as the difficidin BGC, for which one out of the three total DHD-modules is actually an extending module [52]. The reason that these split modules are often inactive is that there is not enough room between active domains that AT domains can acylate the ACP, causing inactivation of the KS domain [30].

Two other modules are also commonly associated with non-condensing KS_0-domains, which are the O-methyltransferases (oMT) and branching domains. Out of the 14 total modules containing oMT-domains only two have been experimentally determined to be active[30]. Examples of this behaviour can be seen in the BGCs of Misakinolide A, luminaolides, tolytoxin, scytophycins, rhizopodin and thailandamide [30]. One transAT-PKS that illustrates most of the unusual behaviours exhibited by transAT-PKS, is thailandamide (BGC0000186), given in Figure 2.4.

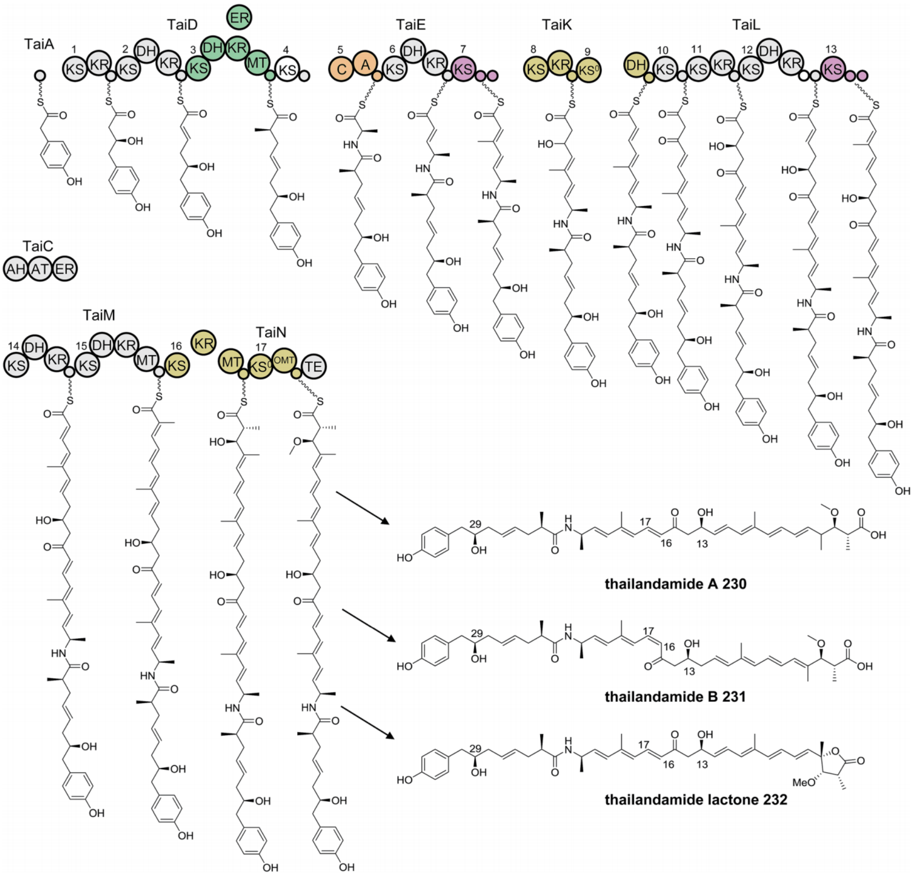


Figure (2.4) The modular structure of transAT-PKS thailandamide. The tandem AT-domains on TaiC are the free-standing enzymes responsible for recruiting malonyl-CoA extender units at each elongation step. An inactive DHD-module can be seen bridging the genes TaiK and TaiL. The KS-DH-oMT-ACP module on TaiN has experimentally been determined to be non-elongating, although the oMT-domain is still known to methylate the hydroxyl group on C3 of the PK intermediate on module 16 (The inactivity of the oMT-containing module can be seen by the fact that there is no chain elongation between modules 16 and 17, as well as by the KS domain noted as a KS0-domain). The bridging module (KS-(KR)-cMT-ACP) between TaiM and TaiN is active, although the domains are located on two different genes. Repeating ACP-domains can be found throughout the cluster, on modules 7, 12 and 13. A free-standing ER domain can be observed acting as part of module 3, on TaiD, and similarly a KR-domain between modules 16 and 17. There can for this BGC not be found any loader module - The starter unit is donated by a free-standing ACP. Figure collected from [30].

2.4 PKS extender units

The mechanism behind polyketide chain extension is the same for both transAT-PKS and their cis-AT counterparts (T1PKS) - condensation of an acyl-unit onto the polyketide, giving off CO_2. However, while transAT-PKS typically only incorporate malonyl-CoA extender units, T1PKS are known to incorporate a wide range of extender units[33]. The purpose of this section is to account for the metabolic origin (i.e. which precursor substrates they are synthesised from) of the various extender units that are used by PKSs, as well as their overall prevalence in the pool of known PKS products.

2.4.1 Malonyl-CoA

Malonyl-CoA is the most frequently used extender unit in polyketide synthesis [23, 33, 46, 48]. There have currently been found two different pathways for synthesis of malonyl-coa in organisms. One that converts malonate directly to Malonyl-CoA, and another pathway that utilizes acetyl-coa and carboxylated biotin in a carboxylation reaction [33]. In *S. coelicolor*, only the latter of the two pathways is found[29]. The synthesis of malonyl-CoA is necessary for all living organisms, as it is a precursor for (among others) the phospholipids that constitute the cell wall [55].

2.4.2 (2S)-Methylmalonyl-CoA

After malonyl-CoA, methylmalonyl-CoA is the second most prevalent extender unit in PK synthesis[23, 33, 46, 48]. Like for malonyl-CoA, this is related to the metabolite's prevalence in fatty acid synthesis, making it an ideal substrate for other reactions because of it's availability [33]. Unlike malonyl-CoA, methylmalonyl-coa can be synthesized through several pathways, as depicted in Figure 2.5.

In *S. coelicolor*, (2R)-methylmalonyl-CoA is synthesized from succinyl-CoA, and epimerized into (2S)-methylmalonyl-CoA. It can also be synthesized from (S)-Methylmalonate semialdehyde, propionyl-CoA and methylmalonate[29]. However, methylmalonyl-CoA does not necessarily exist as a metabolite in all organisms. For instance, E. coli does not produce this extender unit [33].

2.4.3 (2S)-Ethylmalonyl-CoA

Ethylmalonyl-CoA is one of the two non-malonyl-CoA extender units that have been found to participate in one instance of transAT-PKS in the Kirromycin pathway. The BGCs of elaiophylin (BGC0000053), tylosin (BGC0000166), spiramycin (BGC0002033), concanamycin A (BGC0000040), and indanomycin (not annotated in MIBiG) all code for crotonyl-CoA carboxylases that themselves produce ethylmalonyl-coa [33]. Note that neither ethylmalonyl-CoA nor methylmalonyl-CoA share a common pathway with malonyl-CoA.

In *S. coelicolor*, only the pathway through crotonyl-coa is found, crotonyl-CoA being synthesized from butyryl-CoA, 3-hydroxybutyryl-CoA or glutaryl-CoA. Incidentally,

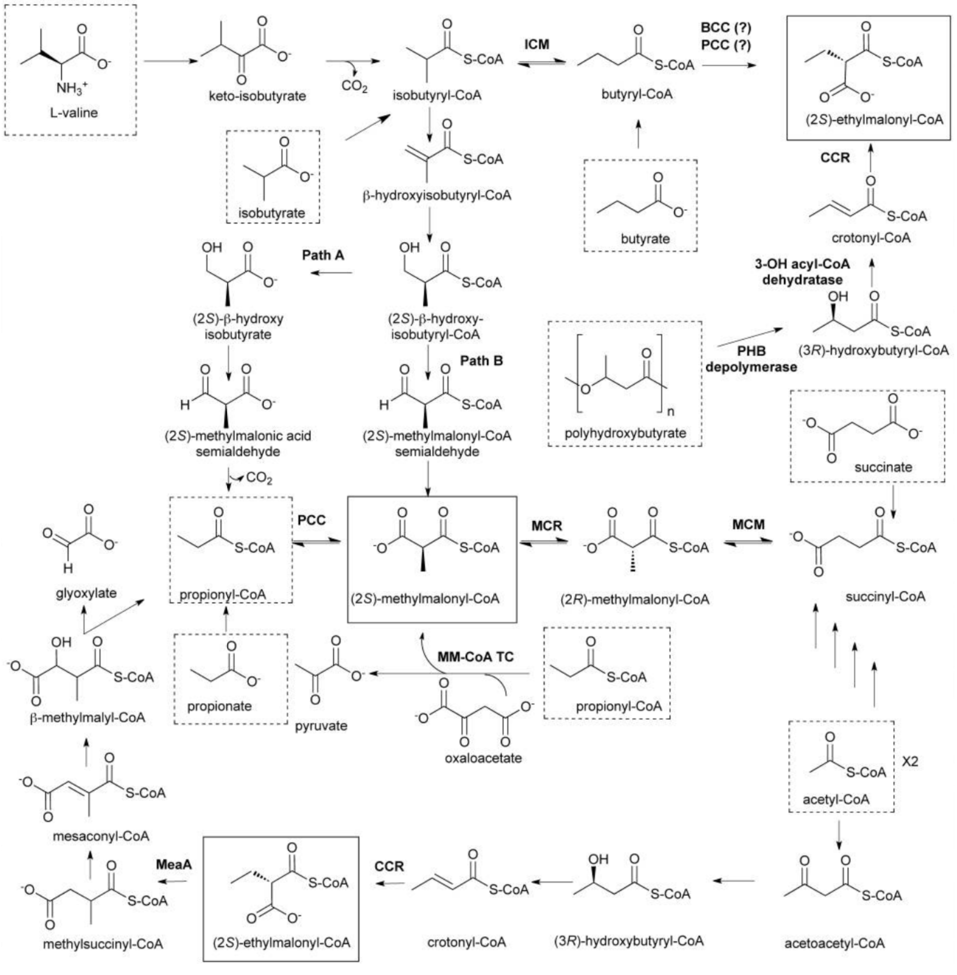


Figure (2.5) Biosynthesis of methylmalonyl-CoA and ethylmalonyl-CoA shown through different pathways. Ethylmalonyl-CoA can be found in the top right and bottom left of the Figure. Methylmalonyl-CoA can be found near the center of the figure. Both are highlighted by solid lined squares. Figure collected from [33].

the synthesis of ethylmalonyl-CoA is the only substrate that uses crotonyl-CoA as a precursor substrate. Ethylmalonyl-CoA is further used in synthesis of germicidin and (2S)-methylsuccinyl-CoA [29].

2.5 Proprietary extender units

From here on down, all extender units are solely used for secondary metabolism. There has been found no evidence that any of the remaining 9 metabolites partake in any form of primary metabolism. The usage of these extender units are often specific to a few closely related BGCs [5, 33].

2.5.1 ACP-bound extender units

The ACP-bound extender units (2R)-methoxymalonyl-ACP, (2R)-hydroxymalonyl-ACP and (2S)-aminomalonyl-ACP are all precursors that are unique to polyketide synthesis. The genes encoding their synthesis are found in close proximity to the BGC, either flanking on either side, or being located between core genes. The synthesis of these extender units can often be suggested by the presence of an FkbH-like protein - FkbH being a protein with no other known relation to that of synthesizing these uncommon extender units [33].

2.5.2 (2R)-Methoxymalonyl-ACP and (2R)-Hydroxymalonyl-ACP

The synthesis of methoxymalonyl-ACP was first characterised in the BGC encoding FR-900520 (BGC0000994), found in *S. hygroscopicus subspecies ascomyceticus* [33]. The names of the enzymes responsible for the synthesis of the extender unit (FkbG, FkbH, FkbI, FkbJ and FkbK) have later been used to characterise the function of other proteins, leading to them being colloquially known as FkbH proteins [23]. In methoxymalonyl-ACP synthesis in FR-900520, an FkbH protein covalently tethers 1,3-biphosphoglycerate to the Ppant group of holo-FkbJ (FkbJ is an ACP. The suffix ”holo” indicates that there is bound a Ppant arm to the ACP), forming glyceryl-FkbJ. Subsequent modifications by FkbG, FkbI, FkbJ and FkbK yield the completed extender unit (Figure 2.6 A). Additional co-factors of this synthesis are NADPH, FAD and S-adenosyl-L-methinoine (SAM). Another mechanism for synthesis of methoxymalonyl-ACP is found in the soraphen BGC (Figure 2.6 B) and although the pathways both result in the same product, spending the exact same substrates, there is a significant difference in the fact that the pathway is unrelated to an FkbH-like protein, and instead is initiated by a free-standing, single-module enzyme consisting of the three domains AT-ACP-MT.

(2R)-hydroxymalonyl-ACP, is an intermediate in the pathway of methoxymalonyl-ACP, deviating only in the fact that the side chain on the alpha carbon is not methylated by SAM. The BGC of zwittermycin A (BGC0001059) contains the genes *zmaN*, *zmaD*, *zmaG*, and *zmaE*, coding homologs of FkbH, FkbJ, FkbK, and FkbI, respectively (Figure 2.6 C) [33].

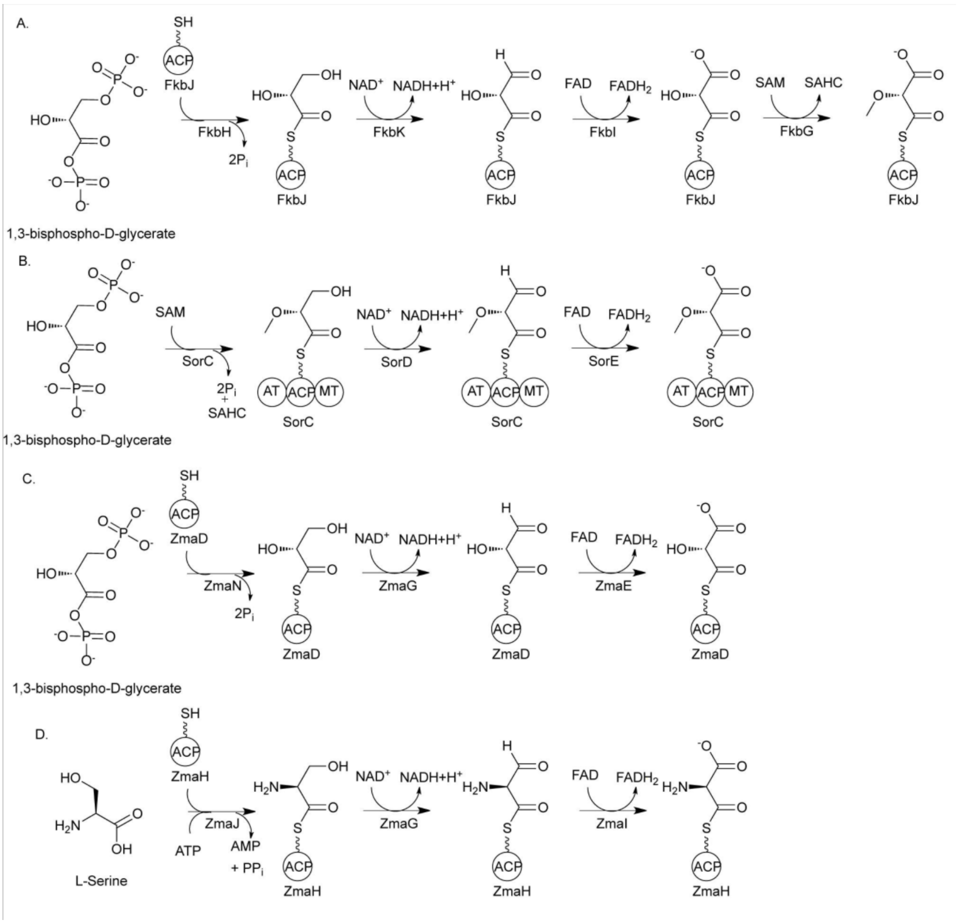


Figure (2.6) A and B: Pathways encoding synthesis of methoxymalonyl-ACP. both reactions use the same cofactors, abeit in a reverse order. C: Biosynthesis of hydroxymalonyl-ACP. Reactants and coreactants are identical to A, except for the methylation of the hydroxyl group on C3 by SAM. D: Biosynthesis of aminomalonyl-CoA.

2.5.3 Unique extender units

As of 2015, at least 7 BGC-specific extender units had been discovered (Table 2.6). As for hydroxymalonyl-ACP and methoxymalonyl-ACP, the synthesis of these extender units is also encoded by the BGCs they are used as extender units. These extender units are testaments to just how uncommon some extender units can be, and the possible utility of such systems if their mechanisms are completely understood. There has been postulated that other extender units theoretically could be synthesized in a similar fashion, however no other than those mentioned here have been found to date [5, 33]. Still, there is a consensus that many more such uncommon extender units may exist, as the following quote reveals:

”The finding by Alber and colleagues that Crotonyl-CoA reductase catalyzes not only the reduction of crotonyl-CoA but also its carboxylation raises the possibility that previously unknown extender units could be generated by an analogous reaction on any 2,3-desaturated acyl-CoA” (Chan et al. (2006))[33].)

Table (2.6) 7 extender units that have been found to be used in one specific BGC. These extender units cover the time periods 1969-2009[33], 2012-2015[5]. 2-carboxy-3-hydroxy-5-methylhexanoyl-CoA was discovered in 2011[56], but no other extender units are accounted for in this time period.

Extender unit	BGC
Benzylmalonyl-CoA	Splenocins
3-oxoadipyl-CoA	Pamamycins
Dichloropyrrolepropylmalonyl-ACP	Chlorizidine
Dimethylmalonyl-CoA	Yerziniabactin
(2S)-aminomalonyl-ACP	Zwittermycin A
chloroethylmalonyl-CoA	Salinosporamide
2-carboxy-3-hydroxy-5-methylhexanoyl-CoA	Leupyrrin

- **(2S)-Aminomalonyl-ACP** synthesis is initiated by the loading of an L-serine residue onto an ACP. Subsequent modification by various Fkb-like proteins yield the completed extender unit[33].
- **Chloroethylmalonyl-CoA** is derived from SAM. Coproducts are methinoine and adenine[33].
- **2-Carboxy-3-hydroxy-5-methylhexanoyl-CoA** is derived from isovaleryl-CoA and malonyl-CoA[56].
- **Benzylmalonyl-CoA** is derived from phenylalanine[5].
- **3-Oxoadipyl-CoA** is derived from succinyl-CoA and malonyl-CoA[5].
- **Dichloropyrrolepropylmalonyl-ACP** is derived from 4,5-dichloropyrrolyl-ACP and malonyl-CoA[5].
- **Dimethylmalonyl-CoA** is a twice methylated malonyl-CoA. Methyl groups are provided by SAM[5].

2.6 Core genes - Non ribosomal peptide synthase

”The extender unit synthesis genes and core genes of PKS have up untill now been the focus of attention. This section is dedicated to the core genes of NRPS. As the name implies, non ribosomal peptides (NRPs) are not synthesized by the ribosomal machinery, but instead by large multimodular enzymes in a similar fashion as T1PKS[57]. The main difference between NRPS and T1PKS is that while T1PKS products are synthesized from acyl-CoA extender units, NRPS are synthesized from amino acid extender units. The substrates used for NRPS are not limited to the 20 proteinogenic amino acids[3]. In fact more than 500 different amino acids have been recognized in nonribosomal peptides, and are the reason for the wide array of bioactivities these natural products exhibit, some of which are antitumor, antifungal, immunosupressant and antibiotic [57]”. (Fossheim - Project report, 2019)[1]

Although their secondary metabolite products are structurally different, the mechanism behind their synthesis is highly similar to T1PKS. The extending modules have the same structure; Condensing domain - Recruiting domain - Carrier protein. Extender units in the form of amino acids are recruited onto a peptidyl carrier protein (PCP) domain (analogous to ACP domains for PKS) by an Adenosine-Monophospate binding (A) domain (analogous to AT domains for PKS). Extender units recruited this way are then condensed onto the growing polypeptide chain by a condensation (C) domain [57]. The standard load, extender and terminating modules can be seen in Table 2.7. (As a clarifying note: T1PKS also bind extender units to the core intermediate through a condensation reaction. For NRPS this condensate is H_2O instead of CO_2 for T1PKS.) In addition to these obligatory domains, a module may contain one or more optional domains:

- **Formylation** (F) - Formylates (donates a CHO-group) to the polypeptide intermediate, using formyl-tetrahydrofolate (10-CHO-THF) as a cofactor [58].

- **Heterocyclization** (Cy) - into thioazolines or oxoazolines. Cy domain replaces the C domain [57].

- **Oxidation** (Ox) of thiazolines or oxazolines to thiazoles or oxazoles. The oxidation domain is inside a hybrid A-Ox domain [19].

- **Epimerizations** (E) of L-amino acids into D-amino acids. The condensation domain can act as a combined epimerisation/condensation domain [59].

- **N-methylation** (nMT) - Methylates the nitrogen atom of the peptide bond that is created by the C domain [57].

Examples of the resulting structures that are created from heterocyclisation domains, N-methylation domains, and condensation domains incorporating non-proteinogenic amino acids, is given in Figure 2.7.

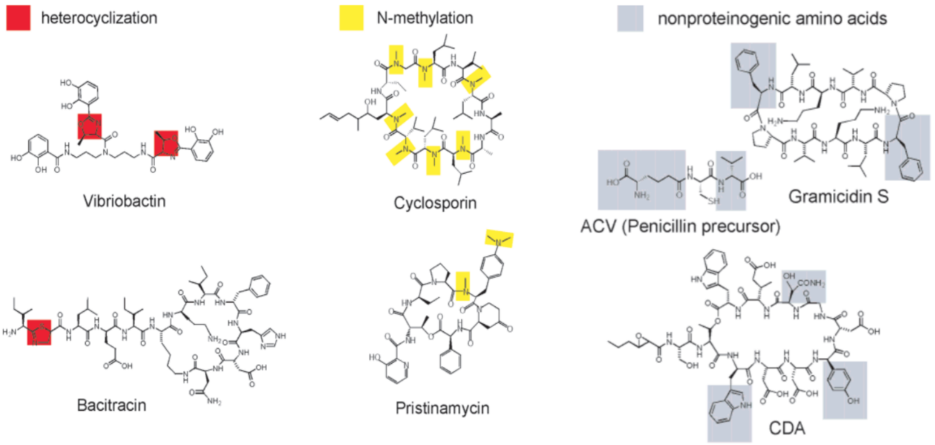


Figure (2.7) The effects of heterocyclisations, N-methylation and condensation domains incorporating non-proteinogenic amino acids in NRPS products. Figure gathered from [60]

Table (2.7) The three standard module structures for NRPS systems. Domains in brackets are optional domains. (C/Cy) is written to show that both condensation and heterocyclization domains are valid (albeit required) domains.

Module type	Domain sequence
Loader module	[F/nMT]-A-PCP-
Extending modules	-(C/Cy)-A-[nMT/Ox]-PCP-[E]-
End module	-(C/Cy)-A-[nMT/Ox]-PCP-[E]-(TE/Red)

Like PKS, NRPS domains can be coupled to a general reaction. The reactions catalyzed by each type of domain is given in Table 2.8.

Table (2.8) Reactions associated with each domain type for NRPS. The polypeptide intermediate is implied to take part in these reactions as well.

Domain	Abbreviation	Reaction
Condensation	C	Amino acid + ATP → H_2O + AMP + PPi
Heterocyclisation	Cy	→ H_2O
Formylation	F	10-CHO-THF → THF
Epimerization	E	No associated reaction
Oxidisation	Ox	O_2 →
Reduction	Red	NADPH → NADP
N-methyltransferase	nME	SAM → SAH

2.6.1 NRPS Extender units

NRPS may incorporate both proteinogenic and non-proteinogenic amino acids as extender units[57]. This section is meant to show the origins of non-proteinogenic amino acids, as well as which extender units CAN be predicted as substrates. One tool that is frequently used for analysing NRPS is NRPSpredictor[17]. In addition to finding NRPS BGCs, the tool also predicts substrate specificities of A domains. This prediction is based off two separate methods of prediction, both of which rely on experimentally curated data. Both methods of prediction are based on coupling the sequence of A domains against the substrate specificity of the A domain [17]. For the sake of simplicity, these methods are assumed to be black box models on the form of ”amino acid sequence goes in → prediction comes out”. They will further be reffered to as method 1 and method 2. Method 1 is based on more recently curated data, and thus contains A domains with additional specificities. These are given in Table 2.9. Method 2 contains 1546 A-domins and their specificities, given in Table 2.10. (The number of signatures is not representative for the number of times a module has been observed to incorporate that extender unit).

Table (2.9) #: number of signatures for the given amino acid in the database used for prediction method 2. DHPG and DPG reference the same substrate. Other than three-letter abbreviations for proteinogenic amino acids, abbreviations are: 2-OIA, 2-oxo-isovaleric-acid; 3-me-Glu, 3-methyl-glutamate; 4pPro, 4-propyl-proline; AAD, 2-amino-adipic acid; ABU, 2-amino-butyric acid; AEO, 2-amino-9,10-epoxy-8-oxodecanoic acid; a-HIA, α-hydroxy-isocaproic-acid; Ala-b, β-alanine; Ala-d, D-alanine; Alaninol; BHT, beta-hydroxy-tyrosine; Cap, caproic acid DAB, 2,4-diamino-butyric acid; DHB, 2,3-dihydroxy-benzoic acid; DHPH = DPG, 3,5-dihydroxy-phenylglycine; DHT, dehydro-threonine, 2,3-dehydroaminobutyric acid; DMA-TRP, N-(1,1-dimethyl-1-allyl)Trp; d-lyserg, D-lysergic acid; HAORN, N-hydroxy-N-acylOrnithine ;HFORN, L-N-hydroxy-N-formylornithine; HORN,L-N-hydroxyornithine; HPG, 4-hydoxy-phenyl-glycine; HYV-D, 2-hydroxy-valeric acid; Iva, isovaline; l-DAP, L-2,3-diaminopropionate; Lys-b, β-lysine; Orn, ornitine; PHG, phenyl-glycine; Pip, pipecolic acid; Sal, salicylic acid; TCL, (4S)-5,5,5-trichloro-leucine; Vol, valinol; V/I/Ai, val/Ile/allo-Ile; V/I, Val/Ile; A/G, Asn/Gln; T/T, trp/tyr. S/T: ser-thr Abbreviations are taken from [61].

AA	Frequency	AA	Frequency	AA	Frequency	AA	Frequency
Thr	35	Orn	12	Sal	2	DMA-Trp	1
Ala	35	Ile	11	Ser-Thr	2	Val/Ile	1
Leu	30	Gln	9	HAORN	2	Val/Ile/alloIle	1
Ser	29	DAB	8	HORN	2	allo-Thr	1
Val	25	Pip	7	PHG	1	Asn/Gln	1
Cys	24	BHT	7	His	1	H-Asn	1
Gly	20	DHPG	7	AEO	1	HFORN	1
HPG	19	Iva	7	4pPro	1	2-OIA	1
Asn	18	Arg	6	ABU	1	a-HIA	1
Pro	17	AAD	6	DPG	1	HYV-d	1
Tyr	16	Lys	5	3-Me-Glu	1	LDAP	1
DHB	15	Trp	5	TCL	1	Cap	1
Asp	14	DHT	4	d-Lyserg	1	b-Ala	1
Phe	13	Lys-b	2	Vol	1	Trp/Tyr	1
Glu	13	Ala-b	2	Alaninol	1		

Table (2.10) Frequency of A-domains with a certain specificity towards a substrate.

Abbreviation	Full name	Signatures	Exists in *S.Coelicolor*
A	Alanine	618	yes
DHB	2,3-dihydroxybenzoate	266	yes
F	Phenylalanine	99	yes
AAD	L-2-Aminoadipate	70	yes
L	Leucine	41	yes
T	Tyrosine	39	yes
V	Valine	38	yes
C	Cysteine	37	yes
S	Serine	34	yes
E	Glutamate	32	yes
D	Aspartate	28	yes
G	Glycine	27	yes
P	Proline	22	yes
N	Aspargine	22	yes
Y	Tyrosine	22	yes
HPG	4-hydroxy-phenyl-glycine	22	no
W	Tryptophan	16	yes
I	Isoleucine	15	yes
DAB	L-2,4-Diaminobutanoate	12	yes
ORN	Ornithine	17	yes
Q	Glutamine	10	yes
PIP	L-pipecolic acid	9	no
BHT	beta-hydroxy-tyrosine	9	no
DHPG	3,5-dihydroxy-phenyl-glycine	9	no
K	Lysine	8	yes
R	Arginine	6	yes
b-ala	β-alanine	5	yes
HORN	N5-hydroxy-L-ornithine	5	no
HYV-D	2-hydroxy-valeric acid	4	no
DHT	2,3-dehydroaminobutyric acid	4	no

2.7 Extender unit synthesis genes

The most frequently observed extender units that are not found in *S. coelicolor* are Hydroxy-phenyl-glycine (HPG), Di-hydroxy-phenyl-glycine (DHPG), β-hydroxy-tyrosine (BHT) and Pipecolic acid (PIP)[17]. As for the uncommon extender units found for PKS, BGCs do in nearly all cases encode for their synthesis, and they are synthesised from precursors that are found in the primary metabolism of cells [62]. This section is dedicated to elucidating the metabolic pathway of these extender units.

2.7.1 Hydroxy-phenyl-glycine (HPG)

Hydroxy-phenyl-glycine is synthesised in a cyclic reaction, with prephenate as the starting substrate. Prephenate - an intermediate in the shikimate pathway - is converted into *p*-hydroxyphenylpyruvate in a decarboxylation reaction. While several amide containing molecules can act as the NH_2-donating substrate, the most efficient donor in this pathway is tyrosine, as the deamination of tyrosine yields *p*-hydroxyphenylpyruvate and thus completing the reaction cycle. Other cofactors in the reaction are Flavine Mononucleotide (FMN), O_2 and NADH [63], which alongside prephenate and tyrosine are found as metabolites in *S. coelicolor* [29]. The reaction cycle is shown in Figure 2.8.

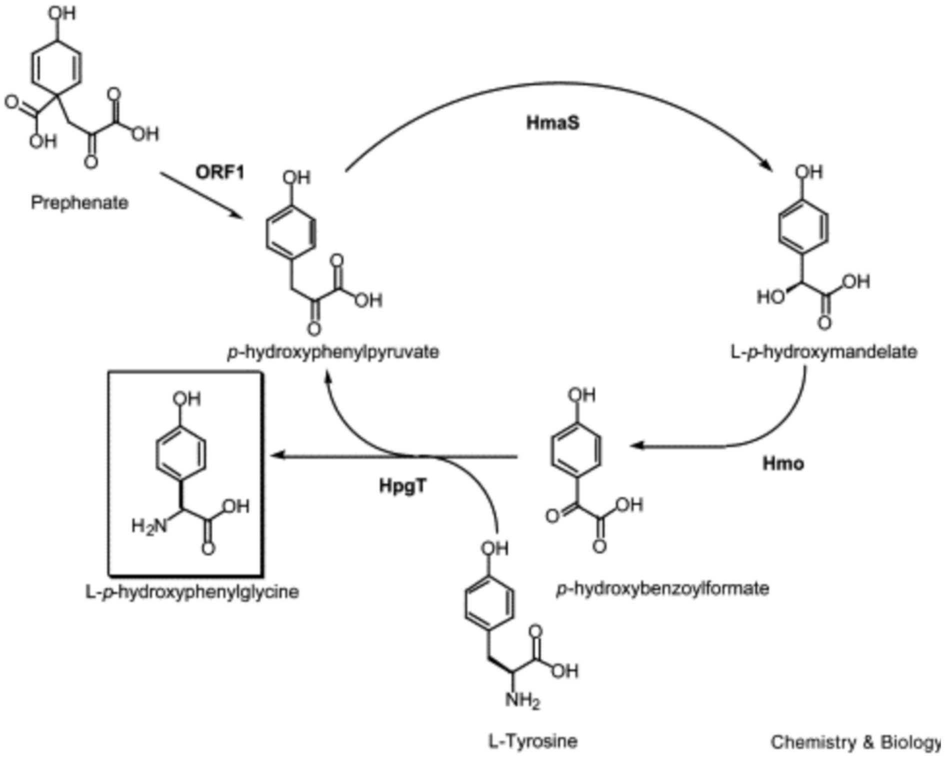


Figure (2.8) The cyclic pathway of HPG as proposed by Hubbard et al. (2000). Figure collected from [63].

2.7.2 Di-Hydroxy-phenyl-glycine (DHPG)

Although similar in name, the metabolic pathway encoding DHPG differs in all ways from that of HPG. Synthesis of DHPG begins with 4 Malonyl-CoA units forming a cyclic polyketide precursor. This precursor is then subjected to two condensation reactions, one oxidisation and a transamination by tyrosine, yielding DHPG. As for HPG, the tyrosine that reacts in this pathway forms *p*-hydroxyphenylpyruvate though this pathway [64].

2.7.3 β-hydroxy-tyrosine (BHT)

Single module enzymes catalysing the formation of uncommon extender units are not only found in PKS, as shown by synthesis of BHT. The oxidisation of tyrosine is catalyzed by a cycochrome p450 monooxygenase, using NADPH + H^+ The precursor for BHT is tyrosine, while cofactors that are used in the reaction are NADPH + H^+, O_2 and H_2O [65]. Synthesis of BHT is shown in Figure 2.9.

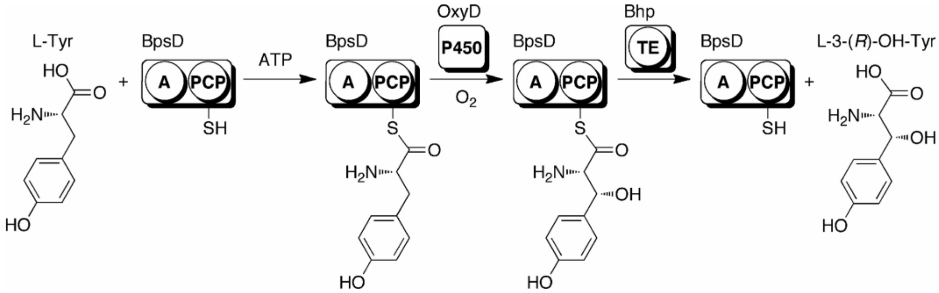


Figure (2.9) Biosynthesis of BHT. Figure collected from [65].

2.7.4 Pipecolic acid

Pipecolic acid is found as an extender unit in (among other secondary metabolites) rapamycin produced by streptomycete. The two main substrates that synthesise Pipecolic acid are pyruvate and lysine, which together with NADH yields Pipecolic acid, glycine and H_2O [66]. Biosynthesis of pipecolic acid is shown in Figure 2.10.

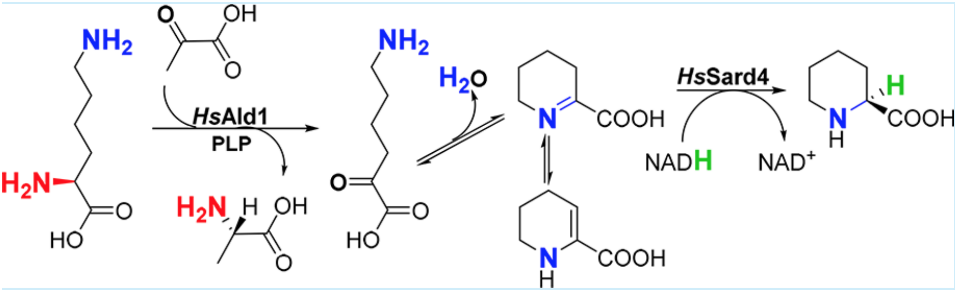


Figure (2.10) Biosynthesis of pipecolic acid. Figure collected from [66].

2.8 Alternative load modules

Untill now, the standard module structures for PKS and NRPS have been described. However, in some cases (for both NRPS and PKS), the load modules contain other domains than the traditional AT-ACP (for PKS) and A-PCP (for NRPS). Despite of the structural variety that is observed for these alternative loader modules, three domains stand out as having specific catalytic activities. These are the FkbH-like hydroxylase domains, GCN5-related N-acetyl transferase (GNAT) domains and Coenzyme A Ligase (CAL) domains[30]. Coenzyme A Ligase domains are often referred to as Acyl-CoA Ligase (AL) domains in literature, but will be referred to as CAL domains in this work, as this is the nomenclature that antiSMASH uses [23].

- **GCN5-related N-acetyl transferase loader modules** - GNATs decarboxylate malonyl-CoA and transfer the resulting acetyl moiety to an ACP to generate the starter group[30].
- **FkbH-like domain containing loader modules** - These starter modules incorporate a D-lactate moiety into the core structure. The precursor for this is 1,3-biphosphoglycerate [30].
- **Coenzyme A ligase loader modules** - Catalyze the incorporation of a wide variety of ACP-bound acyl units. Examples include fatty acids, acyl-CoA, 3-amino-5-hydroxybenzoate (AHBA) and shikimic acid[19]. In addition, these domains can sometimes catalyze the amination of the starting substrate, e.g. malonyl-CoA, using various NH_2-containing substrates[67].

2.8.1 Other loader modules

Exceptions to the traditional loader module structure also exists. For example, the load module of myxalamide (BGC0001022) which has the unusual structure of ACP-KS-AT, and is further confounding by the first extender module having the apparent domain sequence AT-DH-KR-ACP (lacking a KS domain)[68]. The domain sequence of the two modules can be seen in Figure 2.11). In some cases, the load module cannot be accounted for - Polyketide synthesis is initiated by the direct donation of an ACP bound acyl unit onto the first extending module of the BGC[30]. Table 2.11 gives a few examples of variety observed for starter modules, and the secondary metabolites they are associated with.

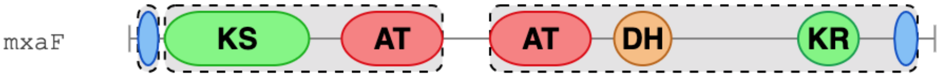


Figure (2.11) Domain structure of MxaF - the load module and first extending module of the myxalamide BGC. The load module structure does not follow the traditional sequence of AT-ACP. Figure obtained from [23].

Secondary metabolite	Domain sequence of loader module
Kirromycin	ACP (in core genes)
Glutarimides*	No apparent load module
Spliceostatin	DH-KR-FkbH-ACP
Oocydin	DH-FkbM-FkbH-ACP
lankacidin**	GNAT-ACP
Calyculin A	A-AMT-ACP
Bacilliaene***	CAL-ACP
Corallopyronin	KS_0-ACP
Malleilactone	TE-A-ACP
Oxazolomycin	F-A-ACP
Rhizopodin	F-MT-A-ACP

Table (2.11) Examples of the variety in loader modules that are found in transAT-PKS. Note the significant number of BGCs that have no apparent loading module [30]. * also: basiliskamide, pederin, onnamides, mycalamides, psymberin, diaphorin, nosperin, rhizoxin and bongkrekic acid. ** also: Enacyloxin, leinamycin. *** also: elansolid and albicidin

2.8.2 Synthesis of alternative starter units

Regardless of the domain sequence of the loader module, PKS and NRPS may incorporate a wide range of different starter units. The uncommon starter units that have been found for NRPS and PKS systems are given in Table 6.1 in Appendix. The precursors of the starter units are given in Table 6.2 in Appendix. Information on 1) if the starter unit is synthesised by genes found within the BGC it is observed, 2) if the starter unit itself is produced by *S. coelicolor*, and 3) if all precursor metabolites are found within *S. coelicolor*, is given in Table 6.3 in Appendix. In 24 out of 26 cases, the starter unit is either found in *S. coelicolor*, or the synthesis of the starter unit is encoded by the BGC *AND* all precursors required to synthesise the starter unit is found in *S. coelicolor*. In one case, the synthesis of the starter unit is uncertain, and in 1 case, the starter unit is synthesised by genes not found in the BGC [53, 69, 70]

2.9 Tailoring genes

The final type of genes apart from the core genes and extender unit synthesis genes, are the tailoring genes. After the polyketide/nonribosomal peptide core has been synthesized, these genes modify the core structure through cyclisation, oxidization, aminotransferase, halogenase, reducing, methylation, acyltransferase and glycosyltransferase reactions reactions[34–36]. Cyclisation, oxidisation, reducing and halogenase reactions are not included as they are typically only dependent on cofactors such as NADPH and ATP and thus represent a much smaller impact on the metabolic pathway than those that incorporate carbon atoms. Examples of substrates that are added in tailoring reactions are: Monosaccharides [71–73], 3,4-Dihydroxydipicolinate [74], aza-beta-tyrosine and 2-naphtonate [75], 2-amino-3-hydroxycyclopent-2-enone [38, 76, 77] and fumarate [38]. This section aims to describe the different substrates that are found to take part in the tailoring reactions.

2.9.1 Glycosyltransferase

Glycosyltransferases are highly prevalent in NRPS and PKS, adding a wide variety of monosaccharides to the core structure - 142 out of 1140 NRPS/PKS systems encode at least one glycosyltransferase[27]. The wide variety of glycosides is provided by other tailoring genes in the BGC such as glycosyl-specific methyltransferases, transaminases, epimerases, reductases, or oxidases [34, 35].

The adding of monosaccharides onto the core structure can be somewhat ambiguous. An example of this can be seen in the arimetamycin BGC which contains 3 glycosyltransferases. The BGC encodes 3 different secondary metabolites, one containing two glycosyl groups, and two containing one glcosyl group. The structure of arimetamycins can be seen in Figure 2.12.

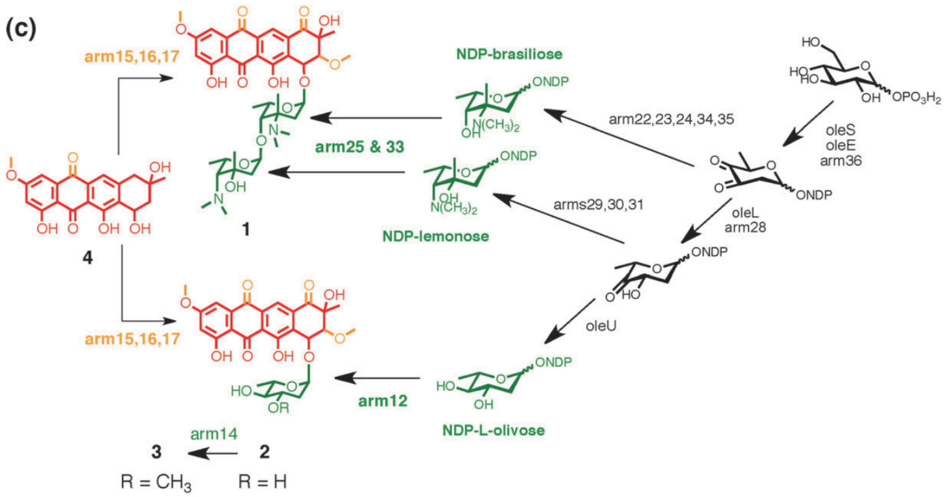


Figure (2.12) The tailoring reactions of aritmetamycin. The BGC encoding aritmetamycin contains 3 total glycosyltransferases, which produce 3 different secondary metabolites. One containing 2 glycosyl groups (1) and two containing one (2 and 3). Figure collected from [78]

Other examples of the varying activities that glycosyltransferases exhibit in polyketide and nonribosomal peptide synthesis can be seen in pathways of aculeximycin, Chromomycin C3 and Komodomycin B.

- The aculeximycin cluster encodes a total of 8 glycosyltransferases, whereas only 5 glycosyl groups are found on the completely assembled secondary metabolite. The reason for the redundant glycosyltransferases is in this case associated with self resistance to the secondary metabolite [72].
- During synthesis of Chromomycin A3, there is added 5 monosaccharides to the core structure, although only 4 glycosyltransferases are found within the cluster. This is due to one of the glycosyltransferases incorporating the same monosaccharide twice [79].

- The BGC encoding komodoquinone B contains 5 glycosyltransferases although komodoquinone B does not contain any glycosyl groups in its structure. However, komodoquinone B is essentially just the core structure of the anthracycline family of secondary metabolites. Anthracyclines that do contain glycosides exist, some of which are encoded by the same BGC as komodoquinone B [80].

2.9.2 Methylation

Methylation reactions are found throughout core genes, extender unit synthesis genes and tailoring reactions. The methylation of a substrate in a biosynthetic pathway consumes S-adenosyl-L-Methinoine (SAM), and releases S-Adenosyl-L-homocysteine (SAH). SAH is then recycled through a series of reactions (given on the next page) to obtain SAM, using folate as a cofactor[29] (Abbreviations are given in Table 2.12):

Abbreviation	Full name
X	To-be-methylated substrate
X-Me	Methylated substrate
SAM	S-Adenosyl-L-methinoine
SAH	S-Adenosyl-L-homocysteine
NAD/NADH	Nicotinamide adenine dinucleotide
THF	tetrahydrofolate
5,10-MTHF	5,10-Methylenetetrahydrofolate
5-MTHF	5-Methyltetrahydrofolate
Gly	Glycine
Met	Methinoine
H-Cys	Homocysteine
Adn	Adenine
P_i	Phosphate
PP_i	diphosphate

Table (2.12) Abbreviations used for metabolites in SAM cycle

$$\mathrm{SAM + X \rightarrow Me{-}X + SAH}$$

$$\mathrm{SAH + H_2O \rightarrow Adn + H{-}Cys}$$

$$\mathrm{Gly + NAD + THF \rightarrow 5{,}10{-}MTHF + NH_4 + CO_2 + NADH}$$

$$\mathrm{5{,}10{-}MTHF + NADH + 2H^+ \rightarrow 5\text{-}MTHF + NAD}$$

$$\mathrm{5\text{-}MTHF + H{-}Cys \rightarrow H^+ + Met + THF}$$

$$\mathrm{ATP + H_2O + Met \rightarrow SAM + P}_i\mathrm{ + PP}_i$$

Combining all reactions yield the total reaction for the recycling of SAM:

$$\mathrm{X{+}SAM{+}2H_2O{+}Gly{+}H^+{+}ATP \rightarrow Me{-}X{+}Adn{+}CO_2{+}NH_4{+}SAM{+}P}_i\mathrm{{+}PP}_i$$

In other words: Each time a substrate is methylated, one glycine is lost through this pathway, and one ammonium ion is created. In addition, one ATP is effectively broken down into its core parts PP_i, P_i and adenosine. The primary function of SAM outside of secondary metabolite synthesis is in the synthesis of nucleotides, meaning that it is an essential metabolite [81].

2.9.3 Acetylations

Acetylations of the core structure are not uncommon, and there is great variation in which acyl-substrates that are used. Examples of tailoring reactions are given in Table 2.13.

Table (2.13) Substrates that are added to the core structure in post-PKS tailoring reactions, and the secondary metabolites that incorporate the substrate. Abbreviations - HDIN: 3-hydroxy-7,8-dimethoxy-6-isopropoxy-2-naphthoic. *4-Hydroxy-3-iodo-5,6-dimethoxy-2-methylbenzoic acid. **2-methoxy-5-chloro-6 methylsalicyclic acid. ***This substrate also seen in ECO-02301, reductiomycin, limocrocin, mannumycin [82], asukamycin[76], Colabomycin E[83] and annimycin[77] BGCs. ****Homo-orselinic acid

Tailoring reaction substrate	BGC
Glycerate	Abyssomycin[84]
Cinnamic acid	Basiliskamides[30]
Carbamoyl-phosphate	Batumin[30]
Aspartate	kirromycin[30]
Malonyl/butyryl/propionyl-CoA	Hautermalides[30]
Cysteine	glutarimides[30]
Holothin*	Thiomarinols[30]
T1PKS-product*	Calicheamicin [85]
T1PKS-product**	Chlorothricin [86]
2-amino-3-hydroxy-cyclopenta-2-enone	Bafilomycin [38]***
Fumarate	Bafilomycin [38]
HDIN	Kedarcidin[75]
Aza-β-Tyrosine	Kedarcidin[75]
3,4-dihydroxydipicolinate	Rubradirin[74]
3-Amino-4-hyrdoxy-7-methoxycoumarin	Rubradirin [74]
T3PKS-product****	Tiacumicin B [87]
Acetyl/propionyl-CoA	midecamycin [34]
Isovaleryl-CoA	Carbomycin [34]

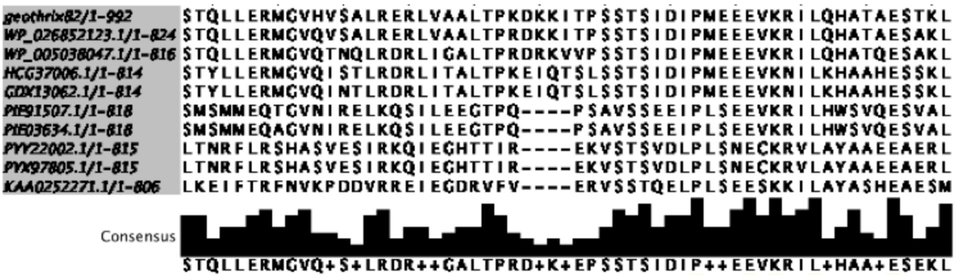
2.10 Identifying BGCs - antiSMASH

This section (2.9 Identifying BGCs - antiSMASH) is copied from the (Fossheim - Project report, 2019)[1]. Minor details have been added.

Several tools have been constructed in order to identify BGCs from genomic information such as CLUSEAN [88], ClustScan [89] and - the most recent and frequently updated detection tool - antiSMASH (antibiotics & Secondary Metabolite Analysis SHell), which utilizes the tools NCBI BLAST+, HMMer 3, Muscle 3, FastTree, PySVG and JQuery SVG [19]. This section focuses on the information antiSMASH gives on the BGC type T1PKS.

BGC detection

To detect BGCs, antiSMASH utilizes profile Hidden Markov Models (pHMMs) constructed from databases of reference genes specific for each class of BGCs (core genes) [19]. The probability for a given amino acid at a specific position in as well as the transition probability to an amino acid at the next position is determined by multiple alignment of these homologous reference genes, such as in Figure 2.13. The pHMM is created from these probabilities while taking gene deletions and insertions into account and describes the probability of going from one state to another, rather than the probability of being in a specific state. E.g. the probability of going from amino acid A to amino acid B. Not the probability that amino acid A is at position P in the sequence. This can be modeled as a network (Figure 2.14), where traversing an edge from one node to another has a certain probability. The pHMM can then be used to search a gene sequence for similar gene profiles. A probability score (E-value) is calculated and reflects the probability that the query sequence is not related to the sequences from which the pHMM was constructed [90]. There are also several heuristics that can be applied this process in order to minimize the computational power required [91].



```
geothrix82/1-992         STQLLERMGVHVSALRERLVAALTPKDKKITPSSTSIDIPMEEEVKRILQHATAESTKL
WP_026852123.1/1-824     STQLLERMGVQVSALRERLVAALTPRDKKITPSSTSIDIPMEEEVKRILQHATAESAKL
WP_005038047.1/1-816     STQLLERMGVQTNQLRDRLIGALTPRDRKVVPSSTSIDIPMEEEVKRILQHATQESAKL
HCG37006.1/1-814         STYLLERMGVQISTLRDRLITALTPKEIQTSLSSTSIDIPMEEEVKNILKHAAHESSKL
GDX13062.1/1-814         STYLLERMGVQINTLRDRLITALTPKEIQTSLSSTSIDIPMEEEVKNILKHAAHESSKL
PIE91507.1/1-818         SMSMMEQTGVNIRELKQSILEEGTPQ----PSAVSSEEIPLSEEVKRILHWSVQESVAL
PIE03634.1/1-818         SMSMMEQAGVNIRELKQSILEEGTPQ----PSAVSSEEIPLSEEVKRILHWSVQESVAL
PYY22002.1/1-815         LTNRFLRSHASVESIRKQIEGHTTIR----EKVSTSVDLPLSNECKRVLAYAAEEAERL
PYX97805.1/1-815         LTNRFLRSHASVESIRKQIEGHTTIR----EKVSTSVDLPLSNECKRVLAYAAEEAERL
KAA0252271.1/1-806       LKEIFTRFNVKPDDVRREIEGDRVFV----ERVSSTQELPLSEESKKILAYASHEAESM

               Consensus
                         STQLLERMGVQ+S+LRDR++GALTPRD+K+EPSSTSIDIP++EEVKRIL+HAA+ESEKL
```

Figure (2.13) Example of a multialignment used to generate pHMMs. Each letter corresponds to its amino acid letter-abbreviation. Dashes indicate no amino acid. Alignment was constructed using JalView [92].

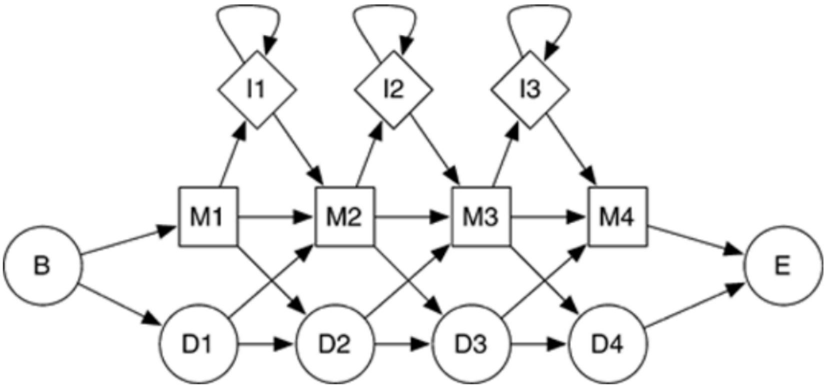


Figure (2.14) Representation of a pHMM network. B and E nodes represent beginning and end, and are dummy nodes. squares represent amino acid sequence, Circles represent deletions and diamonds represent insertions. Insertions are linked to themselves in order to account for any number of amino acid insertions in the sequence. Figure obtained from [91]

When antiSMASH identifies one of these pHMMs in the genome sequence, it returns the sequence of the signature, and a predetermined amount of nucleotides up- and downstream of the sequence. The amount of nucleotides vary depending on the type of BGC. For example for T1PKS this limit is 20 kBP. If more signatures are found within these boundaries, they are also included in the cluster, and a new boundary is set according to the predetermined amount from the newfound site [19].

Domain activity and stereochemistry

AT, KS, ACP and TE domains are assumed to always be active, but the reducing domains can be inactive depending on their stereochemistry. The KR-domains have six possible stereochemistries (A1, A2, B1, B2, C1, C2), which correspond to three possible ketoreduction outcomes: hydroxyl stereoisomer—A or B, or no reduction C. In the case where the stereochemistry is determined to be C, the domain is considered inactive. 88% of times, the activity of the KR domain is correctly predicted [93]. For DH and ER domains, the uncertainty is more significant. For ER domains, there has not been established a connection between the structure of the domain and its activity. However, very few ER domains are inactive in the presence of an active DH domain. Therefore all ER-domains are considered active, as long as there is an active DH and KR domain present in the module [93]. Active DH domains are accurately predicted as active, with only a 9% false prediction rate (i.e predicted as inactive when actually active). However, inactive DH domains are falsely predicted as active at a much higher rate, having an accuracy of only 63% [93].

Core structure predictions

For T1PKS, the core structure prediction is determined by the specificity of the AT domains, and the activities of the reducing domains. While the reducing domains have already been described, the main portion of the core structure is determined by the AT domain specificity - These domains are responsible for recruiting specific extender units that are used for synthesizing the polyketide. antiSMASH predicts this specificity through two separate methods, accepting it if the two methods reveal the same result. If the results do not coincide, the substrate specificity is annotated as ”pk”, meaning that the acyl group could not be predicted [19].

The first of the two methods is based on a 24 amino acid sequence in the AT domain. These 24 amino acids stem from two signature sequences in the CoA-ligase [19] active site of the AT domain. One is 13 amino acids long (the active site) and the other being 11 amino acids long, located slightly away from the active site [94]. As an example, the 13 amino acid domain specificities of methylmalonate and malonate are given in Table 2.14.

Table (2.14) The 13 amino acid signatures of methylmalonate and malonate. interchangable amino acids that still confer specificity are marked in color. Brackets display all possible amino acids that can be in place of the brackets, e.g. any of the amino acids L,V,I,F,A or M can replace the bracket [LVIFAM]. However, only one amino acid at a time

Extender unit	13 Amino acid signature
Methylmalonate	QQGHS[QMI]GR[S]H[T][NS]V
Malonate	QQGHS[LVIFAM]GR[FP]H[ANTGEDS][NHQ]V

The score for the AT-signature prediction is a fraction where the denominator is 24 (i.e. 13 + 11). There is also a possibility that the AT domain displays equal specificity for two or more amino acids. For example an AT-domain with an active site QQGHSIGRFHTNV can have a 95.83% match (23/24*100%) for both malonate and methylmalonate (given that the 11 amino acids long signatures are 100% specific for their respective extender units). The AT-signatures prediction method contains 624 AT-domain signatures, the frequency of which is given in Table 2.15.

Table (2.15) The number of AT domain signatures that correspond to a specific extender unit that exist in the AT-domain substrate database. Data collected from [23].

Extender unit	# of signatures
Malonyl-CoA	347
Methylmalonyl-CoA	234
Ethylmalonyl-CoA	15
Methoxymalonyl-ACP	10
Propionyl-CoA	4
Isobutyryl-CoA	4
2-methylbutyryl-CoA	2
Inactive	1
CHC-CoA	1
Trans-1,2-CPDA	1
Acetyl-CoA	1
Benzoyl-CoA	1
3-methylbutyryl-CoA	1
Cemal	1
2Rhydmal	1

The second method is based on the concept that extender units are recruited by the AT domains based on specific conserved amino acid residues in the sequence [95], a concept that has been backed by extensive investigation in mutagenesis, crystal structure and sequence analysis [19, 95–97]. It differs from the first method in that homology is also taken into account. From conserved residues on homologous sequences that are known to be specific to a certain extender unit, pHMMs are created, and a score is given according to how well the query fits the pHMM. Minowa method HMM profiles for extender units are built on the number of domain sequences given in Table 2.16

Table (2.16) The number of domain sequences that the pHMM of each domain specificity is built upon. NSEQ: Number of sequences used to build pHMM

Extender unit	NSEQ
Malonyl-CoA	266
Methylmalonyl-CoA	175
Methoxymalonyl-ACP	12
Ethylmalonyl-CoA	10
Isobutyryl-CoA	3
Propionyl-CoA	3
2-methylbutyryl-CoA	2
Acetyl-CoA	2
3-methyl-butyryl-CoA	1
Benzoyl-CoA	1
CHC-CoA	1
Fatty acid	1
Inactive	1
Trans-1,2-cpda	1

Predictions of the specificities of alternative loader module domain ”CAL_domain” is shown in Table 2.17.

Table (2.17) Number of domain sequences used to build pHMMs for the prediction of CAL domain specificities. ”NH2” refers to an unspecified ammonia donor [98]. ”Fatty acid” refers to different unspecified fatty acids

CAL specificity	NSEQ
Fatty acid	5
AHBA	4
Shikimic acid	3
Acetyl-CoA	1
NH_2	1

antiSMASH output

antiSMASH output is given to the user in the form of a GBK file for each BGC and three .txt files ”Known cluster BLAST”, ”Cluster BLAST” and ”SubClusterBLAST”.

- **Known cluster BLAST** - This file contains all similar BGCs that exist in the Minimum information about a Biosynthetic gene cluster (MiBIG) database, which consists of BGCs that have been analyzed experimentally. Each entry in this database satisfies the MIBiG standard. For most of the BGCs in this database, the metabolic pathway and end product has been determined [23, 27].

- **Cluster BLAST** - These results are procured in the same way as for the known cluster BLAST, except that the query BGC is examined against the antiSMASH database[25]. The database consists of antiSMASH results for all bacterial genomes marked as ”complete genome” available in the NCBI GenBank repository [23].

- **SubClusterBLAST** - The identified clusters are searched against a database containing operons involved in the biosynthesis of common secondary metabolite building blocks e.g. the biosynthesis of non-proteinogenic amino acids for NRPS, methoxymalonyl-ACP extender units and AHBA starter units for PKS [23]. Currently, SubClusterBLAST is able to detect the metabolic pathways behind 5 different starter units, as well as 5 pathways producing substrates that take part in tailoring reactions.
- **GBK file** - Each predicted BGC is given in the form of a GBK file containing
  - All genes inside the BGC, both their translated nucleotide-sequence and their gene identifier.
  - Secondary metabolism Clusters of Orthologous Groups of proteins (smCOG) definition of each gene (if available). smCOG definitions are determined by comparing each gene in the BGC against a database of COGs of proteins involved in secondary metabolism. This information is then used to provide an annotation of the putative function of the gene products.

2.11 Metabolic modeling

A genome-scale metabolic model (GEM) is an accurate description of the complete network of metabolic reactions that can occur in an organism, as determined by the enzyme encoding genes in the genome. These models are valuable tools that have been used in a range of different applications from strain engineering to modeling cancer metabolism [99]. Using one of several developed pipelines[100], one can obtain a draft metabolic network reconstruction that can be further curated into a high-quality model by gap-filling and validation against experimental data.

Flux balance analysis (FBA) is one of the flagship methods used to predict metabolic phenotypes using these models. FBA leverages linear programming to predict optimal flux states based on a given objective in a given nutritional environment under the assumption of steady-sate metabolite pools [101] Beyond FBA there is a wide range of more advanced algorithms that can include additional data, such as enzyme kinetics and omics data[102]).

Chapter 3

Methods

Based on the literature review and a careful analysis of the results provided by antiSMASH, an algorithm was constructed based on the rules and trends that were observed for each type of BGC. While some decisions in the algorithm are based on observations found by other researchers, others are based on a thorough comparison between the predicted functionality of genes and experimentally determined metabolic pathways. The results from these comparisons are detailed in Section 4.1. The input into the algorithm is the ".gbk"-file provided on the BGC as given by antiSMASH. Following this paragraph is a short pseudocode on how the algorithm works (names of variables differ from those in the actual code). A flowchart describing the main functionality is given in Figure 3.1.
The algorithm is implemented using python 3.7.1 and publicly available at:
`https://github.com/FredrikFossheim/MasterThesis`.

3.1 Pseudocode

3.1.1 Gathering information

```
information_dict = GetData(antiSMASH_output.gbk)
```

The first step is to scrape the ".gbk"-file for relevant information. "`information_dict`" is a dictionary, containing two entries: "`gene_information`" and "`core_information`".

`gene_information`

- **smCOG** - The smCOG definition for the gene, in a integer format. e.g. "smCOG 1109:8-amino-7-oxononanoate synthase" would be stored as 1109
- **Core gene** - The type and location of the core genes in the BGC. E.g. {"start": 2050, "end": 147976, "type": NRPS} If the gene is not a core gene, this is set to False.

- **Domains** - If the gene contains any domains (as predicted by antiSMASH) these are all stored here for the gene - The location, activity and type.
- **Strand** - The strand that the gene is on is necessary to know, as it affects the order that antiSMASH lists domains.

core_information

Contains the predicted extender unit and location for each module as identified by antiSMASH.

3.1.2 Building modules

```
module_list = find_modules(information_dict)
```

A list of modules is created based on rules that have been observed for NRPS and PKS. These modules include those that antiSMASH predicts, as well as other modules such as bridging modules, custom starter modules and oMT modules.

3.2 Finding tailoring reactions

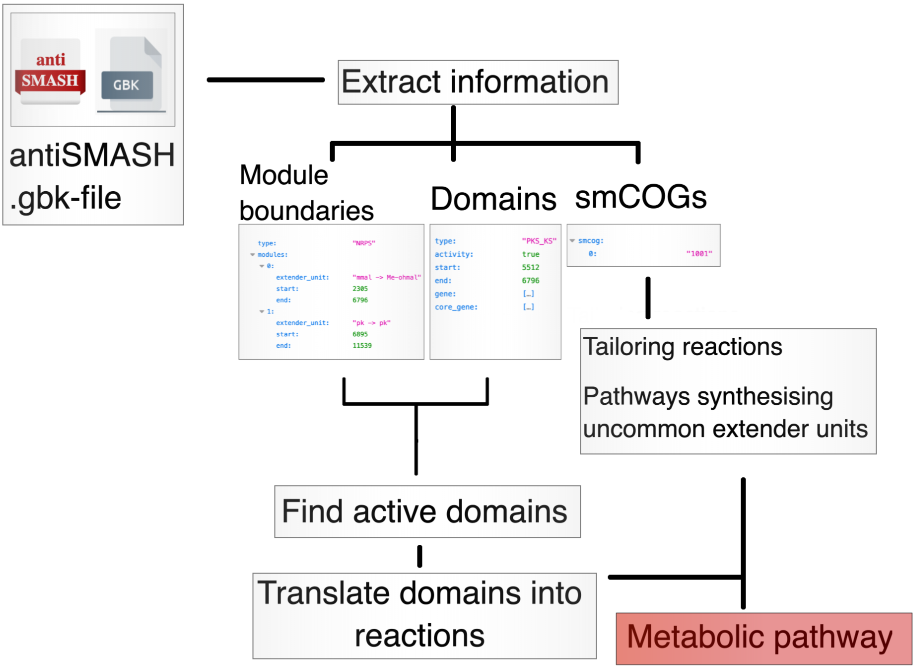
```
additional_reactions = find_additional_reactions(information_dict)
```

The smCOG definition of each gene is compared against a list of smCOGs that have been found to be related to certain tailoring reactions and extender unit synthesis pathways.

3.2.1 Creating the metabolic pathway

```
create_model(module_list, additional_reactions)
```

Finally, the lists of modules and additional reactions is converted into a metabolic model. The model is a set of reactions that together synthesise the secondary metabolite from primary metabolites.



```
  type:                "NRPS"
▾ modules:
  ▾ 0:
      extender_unit:   "mmal -> Me-ohmal"
      start:           2305
      end:             6796
  ▾ 1:
      extender_unit:   "pk -> pk"
      start:           6895
      end:             11539
```

```
type:                 "PKS_KS"
activity:             true
start:                5512
end:                  6796
gene:                 […]
core_gene:            […]
```

```
▾ smcog:
    0:                        "1001"
```

Figure (3.1) Flowchart outlining the function of the developed software, converting antiSMASH results into a metabolic pathway that can be included in any GEM.

3.3 Background for implementation of functions

This section describes more in-debt how the algorithm works, and the reasoning behind implementing the different functionalities of the algorithm. In the cases where this reasoning is based on our own findings, they are expanded upon in section 4.1.

3.3.1 FkbH-domains

The FkbH domain differs from other domains by being mentioned in literature describing all the three types of genes presented in theory - core genes, extender unit synthesis genes and tailoring genes. To define the role of detected FkbH domains, the antiSMASH output for all BGCs in the MiBIG database was searched for FkbH domains. The location of the FkbH domain on the BGC as well as the presence of other genes with specific smCOGs, resulted in rules that could accurately couple the FkbH domain to specific parts of the metabolic pathway: 1) A FkbH loader module that incorporates D-lactate into the core structure. 2) The synthesis of methoxymalonyl-ACP. 3) A tailoring reaction that adds glycerate to the synthesized core structure.

3.4 Core genes

Although antiSMASH predicts the domain sequence and whether or not they are part of a module, this information is stored separately in the output file of antiSMASH. The first process in finding the metabolic pathway that the core genes is responsible for, is therefore to rebuild the domain-module sequence as predicted by antiSMASH. Then additional modules are constructed from domains that fall short of antiSMASH detection rules for modules - bridging modules. In addition, modules containing an oMT domain are noted as being inactive. The process is illustrated in Figure 3.2

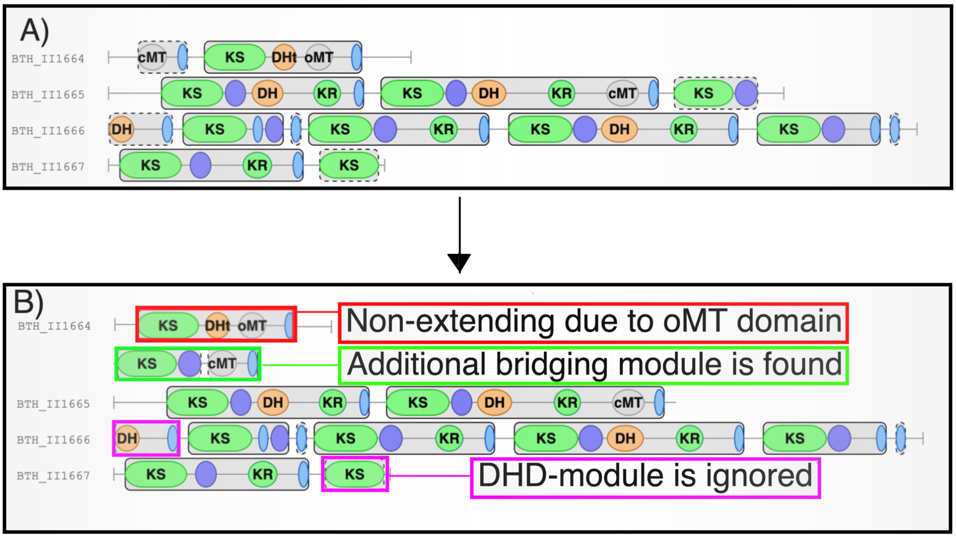


Figure (3.2) A) The reconstructed domain-module sequence as predicted by antiSMASH. B) The domain-module sequence after it has been processed by the algorithm. DHD-module = DeHydratase-Docking (inactive module). Note that the domain-module sequence is read from bottom to top and left to right. Figure based on elements from the thailandamide BGC[30], domain-module sequence representation was collected from antiSMASH[23].

Loader modules can be elusive due to variation in their domain sequence - some BGCs have clearly defined loader modules found on the first core gene of a linear PKS/NRPS, such as the formulaic AT-ACP loader module of bafilomycin (Figure 3.4)[23, 38]. Others, such as the BGC encoding maytancine are non linear, meaning that the loader module (in this case a CAL-ACP module) is not found on the first gene, albeit located at the beginning of a core gene (Figure 3.3)[23]. An example where no loader module can be found is lactimidomycin BGC. Here, the loading mechanism is somewhat unclear. Experimental analysis has revealed that it involves the direct acylation of the first KS-domain of the core genes by an acyl-ACP[30] (Figure 3.5).

3.4.1 Adding the loader module

After the improved module sequence has been found, the loader module is added. The different types of loader modules considered are: GNAT, FkbH, CAL (Section 2.8), A (Section 2.7), and AT (Section 2.2) loader modules. These are detected by first searching for a gene among the core genes that contains any of the alternate loader module-associated domains (FkbH, GNAT and CAL) located upstream of any extender modules. An example that satisfies these requirements is given in Figure 3.3. All domains preceding the first module on the gene are then converted into a module containing those domains.

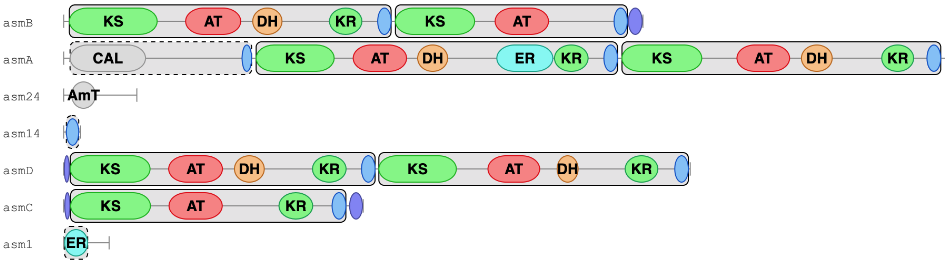


Figure (3.3) Domain-module sequence of maytancine. The loader module has been found to be the CAL-ACP module on AsmA.

If none of the alternate starting domains are found, the program searches for the sequence AT-ACP for PKS or A-PCP for NRPS. This would be the case for the bafilomycin domain-module sequence, given in Figure 3.4. The program would detect the AT-ACP domains on ADC79616.1, and substitute them with a module containing those two domains.

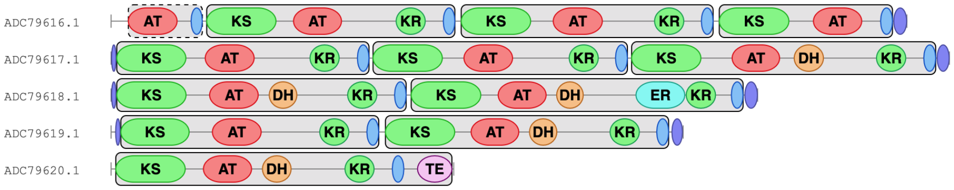


Figure (3.4) Domain sequence of the bafilomycin BGC. The loader module is the AT-ACP module circled with dotted lines on ADC79616.1

If neither of these are true, a custom loader module is added to the sequence of domains and modules. This would be the case for the domain sequence of lactimidomycin, for which no apparent loader module can be detected (Figure 3.5). (This is based on the assumption that the starter unit is loaded onto the first module by other means such as for C-PCP, KS_0 and ACP loader modules). In these cases, the starter unit is assumed to be malonyl-CoA.

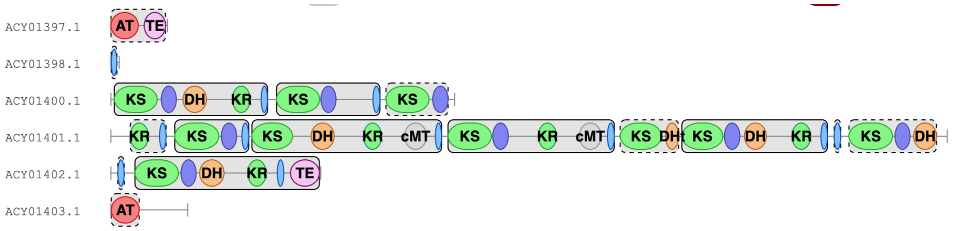


Figure (3.5) Domain sequence of the lactimidomycin BGC as predicted by antiSMASH. No loader module can be clearly defined for this cluster.

3.5 Extender unit prediction

Ideally, each module in the domain-module sequence is coupled to one specific predicted extender unit (e.g. malonyl-CoA or methylmalonyl-ACP for PKS, or tyrosine or alanine for NRPS). However in some cases, the substrate specificity of the module could not be predicted by antiSMASH. In this case, there is a need to change the substrate specificity of the module from ”unknown substrate” to something that can be added to the metabolic pathway. Another case where the substrate specificity of a module can be changed, is if the presence of a pathway synthesizing methoxymalonyl-CoA has been detected.

NRPS

In the case where the substrate specificity of an NRPS module could not be determined, the substrate specificity of the module is set to ”Unknown amino acid”. This is handled later when the metabolic model is created.

PKS

If the pathway of methoxymalonyl-ACP was detected (see FkbH-domains, section 4.1.1), the following actions are performed: If either of the two prediction methods for the substrate specificity (AT_signature-prediction and Minowa-prediction) predicted methoxymalonyl-ACP as the substrate for the module, the specificity of the module is set to methoxymalonyl-ACP. In addition, all modules with specificity towards ethylmalonyl-CoA are changed to methoxymalonyl-ACP, based on the finding that there is a high degree of uncertainty for predictions of this type.

3.6 Constructing metabolic pathway

To construct the part of the metabolic pathway for the core structure, the predicted domain-module sequence is used. All domains that are not part of modules are disregarded. Each module is firstly examined for being either active, inactive, or non-extending. If the module is inactive (DHD-modules), the algorithm moves on to the next module in the domain-module sequence. If the module is active, the extender unit that antiSMASH has predicted for the module is added to the metabolic pathway, along with the co-factors associated with the reaction. If the module is non-extending (oMT-modules), this step is skipped. Then, (for both active and non-extending modules) the reactions associated with all domains in the module are added to the pathway.

When the extender unit is added, an attempt is always made to add the extender unit as a metabolite that already exists in the model organism (e.g. *S. coelicolor*). However, in some cases the extender unit does not exist in the metabolic model. For PKS type extender units this is in most cases due to no consensus prediction from antiSMASH which leads to the extender unit being denoted as ”pk”. In these cases the extender unit is assumed to be malonyl-CoA. For NRPS type extender units, a custom extender unit ”Unknown amino acid” is added to the metabolic pathway. In addition, 20 reactions converting each of the proteinogenic amino acids into the ”Unknown amino acid” substrate is added to the model.

In the cases where the substrate specificity of the module is HPG, BHT, Pipecolic acid or DHPG, the pathways synthesizing these are also added to the model. This is based on the theory that states that these extender units are usually encoded by the BGC. This is also done for substrates that are specific to starter modules, i.e. the CAL domains with specificities towards AHBA, ”Fatty acid” and ”NH2”.

Lastly, the tailoring reactions that have been predicted based on the smCOG definitions of the genes in the cluster are added to the metabolic pathway. The metabolic pathway (and any other pathways synthesising extender units or substrates that are added in tailoring reactions) is saved as a COBRA model in .json format to a user-specified path.

Chapter 4

Results

The results section is divided into two parts. The first part - 4.1 ”Analysis of MiBiG clusters” - will present 1) the results that deal with analysis of the role of FkbH domains in BGCs, 2) the basis for how tailoring reactions are predicted based on smCOG definitions, and 3) analysis of the predicted substrate specificities of all modules that exist in the MIBiG database. The second part - 4.2 ”Comparison of constructed and experimentally determined pathways” - compares 8 pathways constructed by the algorithm against their experimentally determined counterparts.

4.1 Analysis of MiBiG clusters

To investigate the correlation between the smCOG of genes and non-core gene reactions, the antiSMASH output for all BGCs in the MIBiG database was downloaded and analyzed. MIBiG does not have all these files readily downloadable, so this process was done by web-scraping. This resulted in the acquisition of 1887 antiSMASH outputs which could be compared to experimentally determined data. This number does not agree with the total number of BGCs in MIBiG which is 1932. The reason for the disagreement between the two numbers was that the antiSMASH prediction does not exist for all BGCs in MIBiG. While the exact reason for discrepancy is not known, it did not cause any problems during analysis of individual clusters that were acquired this way.

4.1.1 Identification of FkbH domains

The search for FkbH domains in all BGCs resulted in finding that 59 out of the total 1887 successfully downloaded outputs contained an FkbH domain. Out of these 59, 51 were found within a BGC predicted as a NRPS or PKS cluster. Out of these 51 entries, 43 were non-redundant entries. These BGCs were then manually analyzed against their experimentally determined pathways. It was found that the presence of two genes with specific smCOG definitions gave a correct prediction of a specific reaction in 94% of cases.

Location

The location of the FkbH domain was found to be highly relevant with respect to its associated reaction. By examining the location of the FkbH-like domain, it was found that all FkbH-domains located within a core gene, was part of a loader unit (7/7 cases, Table 4.3). I.e an FkbH-domain was never part of an extending module when it was located within the core genes (This concept is illustrated in Figure 4.1).

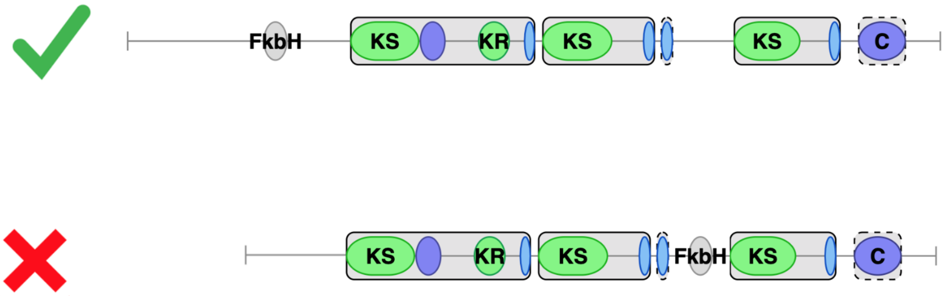


Figure (4.1) When a FkbH domain is found within the core genes, it is always found preceding a module.

Presence of other genes

When the FkbH domain was located outside the core genes of the BGC, this found to be indicative of either glycerate being added to the core structure by a tailoring enzyme (using 1,3-biphosphoglycerate as a precursor metabolite), *or* the synthesis of an uncommon extender unit (either methoxymalonyl-ACP or hydroxymalonyl-ACP). The presence of genes with specific smCOGs was used as a measure to differentiate between post-PKS addition of glycerate, and extender unit synthesis. When the FkbH domain was related to the synthesis of an extender unit, there was a gene with smCOG ”1095:3-hydroxybutyryl-CoA dehydrogenase” present in 23 out of 24 cases, as well as a gene with smCOG 1006:acyl-CoA_dehydrogenase in 24 out of 24 cases. Table 4.1). When the FkbH domain was associated with adding glycerate in a tailoring reaction, there was found a gene with smCOG ”3-oxoacyl-(acyl carrier protein) synthase III” present (10 out of 11 cases, Table 4.2).

In an effort to distinguish the extender units hydroxymalonyl-ACP and methoxymalonyl-ACP from each other, the BGC was examined for genes with function related to O-methylation. 1 out of 7 hydroxymalonyl-ACP synthesizing BGCs and 3 out of 18 BGCs synthesizing methoxymalonyl-ACP contained O-methylating domains, meaning that the two extender units could not be differentiated with this heuristic. Therefore, methoxymalonyl-ACP was always assumed to be synthesized because of its overall higher prevalence (7 hydroxymalonyl- ACP vs. 17 methoxymalonyl-ACP).

Table (4.1) Listing the BGCs in MIBiG that contains an FkbH domain and that synthesizes an uncommon extender unit. The presence of genes with smCOG definitions 1006: acyl-CoA_dehydrogenase, 1084: 3-oxoacyl-(acyl_carrier_protein)_synthase_III and 1095: 3-hydroxybutyryl-CoA_dehydrogenase is given in the columns with the respective smCOG numbers. Rows in red text highlight the cases where the presence or absence of genes with the specific smCOG definitions differed from the majority.

BGC#	Product	Substrate	1006	1084	1095
20	Actinosynnema[103]	Hydroxymalonyl-ACP	Yes	No	Yes
21	Apoptolidin[104]	Methoxymalonyl-ACP	Yes	No	Yes
28	Bafilomycin[38]	Methoxymalonyl-ACP	Yes	No	Yes
40	Concanamycin A[105]	Methoxymalonyl-ACP	Yes	No	Yes
65	Rustmicin[106]	Methoxymalonyl-ACP	Yes	No	Yes
66	Geldanamycin[107]	Methoxymalonyl-ACP	Yes	No	Yes
74	Herbimycin A[108]	Methoxymalonyl-ACP	Yes	No	Yes
78	Incednine[109]	Methoxymalonyl-ACP	Yes	No	Yes
90	Macbecin[110]	Methoxymalonyl-ACP	Yes	No	No
96	Midecamycin[111]	Methoxymalonyl-ACP	Yes	No	Yes
159	Tautomycin[112]	Methoxymalonyl-ACP	Yes	No	Yes
970	Chondrochloren A[113]	Methoxymalonyl-ACP	Yes	No	Yes
1034	Pellasoren[114]	Methoxymalonyl-ACP	Yes	No	Yes
1054	Xenocoumacin[115]	Hydroxymalonyl-ACP	Yes	No	Yes
1059	Zwittermycin A[116]	Hydroxymalonyl-ACP	Yes	No	Yes
1106	Oxazolomycin B[117]	Methoxymalonyl-ACP	Yes	Yes	Yes
1348	JBIR-100[51]	Methoxymalonyl-ACP	Yes	No	Yes
1511	Ansamitocin P-3[118]	Methoxymalonyl-ACP	Yes	No	Yes
1537	Butyrolactol A[119]	Hydroxymalonyl-ACP	Yes	No	Yes
1902	Bengamide[120]	Hydroxymalonyl-ACP	Yes	No	Yes
1956	Miharamycin A[121]	Hydroxymalonyl-ACP	Yes	No	Yes
1957	Amipurimycin[121]	Hydroxymalonyl-ACP	Yes	No	Yes
2011	Ansacarbamitocin A[122]	Methoxymalonyl-ACP	Yes	No	Yes
2033	Spiramycin[123]	Methoxymalonyl-ACP	Yes	No	Yes

Table (4.2) Listing the BGCs in MIBiG that contains an FkbH domain and that incorporates glycerate in a tailoring reaction. The presence of genes with smCOG definitions 1006:acyl-CoA_dehydrogenase, 1084:3-oxoacyl-(acyl_carrier_protein)_synthase_III and 1095:3-hydroxybutyryl-CoA_dehydrogenase is also indicated. Rows in red text highlight the cases where the presence or absence of genes with the specific smCOG definitions differed from the majority. 4H3H2HMe2HF5 = 4-hexadecanoyl-3-hydroxy-2-(hydroxymethyl)-2H-furan-5-one

BGC#	Product	Substrate	1006	1084	1095
1	Abyssomycin[84]	glycerate	No	Yes	No
36	Chlorothricin[86]	glycerate	No	Yes	No
82	Kijanimicin [71]	glycerate	Yes	No	No
133	Quartromicin[124]	glycerate	No	Yes	No
140	4H3H2HMe2HF5[124]	glycerate	No	Yes	No
162	Tetrocarcin[73]	glycerate	No	Yes	No
164	Tetronomycin[125]	glycerate	No	Yes	No
1004	Lobophorin B[126]	glycerate	No	Yes	No
1183	Lobophorin A[126]	glycerate	No	Yes	No
1204	Versipelostatin[127]	glycerate	Yes	Yes	Yes
1288	Maklamicin[128]	glycerate	No	Yes	No

Table (4.3) Listing the BGCs in MIBiG that contains an FkbH domain that can act in a loader module. The presence of genes with smCOG definitions 1006:acyl-CoA_dehydrogenase, 1084:3-oxoacyl-(acyl_carrier_protein)_synthase_III and 1095:3-hydroxybutyryl-CoA_dehydrogenase is also indicated

BGC#	Product	Substrate	1084	1006	1095
174	Byrostatin[129]	D-lactate	No	No	No
185	Tartrolon[130]	D-lactate	No	No	No
995	FR901464[131]	D-lactate	No	No	No
1113	Spliceostatin[132]	D-lactate	No	No	No
1114	Thailanstatin[133]	D-lactate	No	No	No
1031	Oocydin[30]	D-lactate	No	No	No
1350	Phormidolide[134]	D-lactate	No	No	No

4.1.2 Tailoring reactions

In addition to the tailoring reaction described in Section 4.1.1, glycosylations, and addition of 2-amino-3-hydroxycyclopent-2-enone to the core structure was found to be associated with observing specific smCOG definitions for genes.

Glycosylation of core structure

From manual examination of tailoring reactions involving glycosyltransferases, there appeared to be a relationship between the number of glycosyltransferases encoded by a BGC, and the number of glycosyl groups on the completely synthesised secondary metabolite.

In order to obtain more scientifically accurate data on this relationship, all BGCs of type PKS or NRPS in MIBiG (a total of 1140 BGCs) were first examined for genes with smCOG definition 1062:glycosyltransferase. This resulted in detecting 142 BGCs containing at least one such gene. The number of glycosyl groups on the experimentally determined secondary metabolite was found by manual curation. As this was a tedious process, this was not done for all of the 142 that contained a glycosyltransferase. Instead, to obtain a significant sample size for this analysis, the 35 first BGCs (sorted by MIBiG identifier number) containing at least one glycosyltransferase were selected for analysis. In addition, all BGCs containing 4 or more glycosyltransferases in all of MIBiG were included the afforementioned group. In total, this gave a sample size of 40 out of 142 BGCs. Analysis of these 40 clusters did indeed reveal correlation between the number of glycosyltransferases encoded by the BGC and the number of glycosyl groups that had been experimentally determined to be added to the core structure (Table 4.4 and Figure 4.2). These results lead to the following reaction being added to the constructed metabolic pathway, for each glycosyltransferase that is found in a BGC:

$$\text{Glucose}-\text{6-phosphate} + [\text{Core intermediate 1}\,] \rightarrow [\text{Core intermediate 2}\,] + \text{P}_i$$

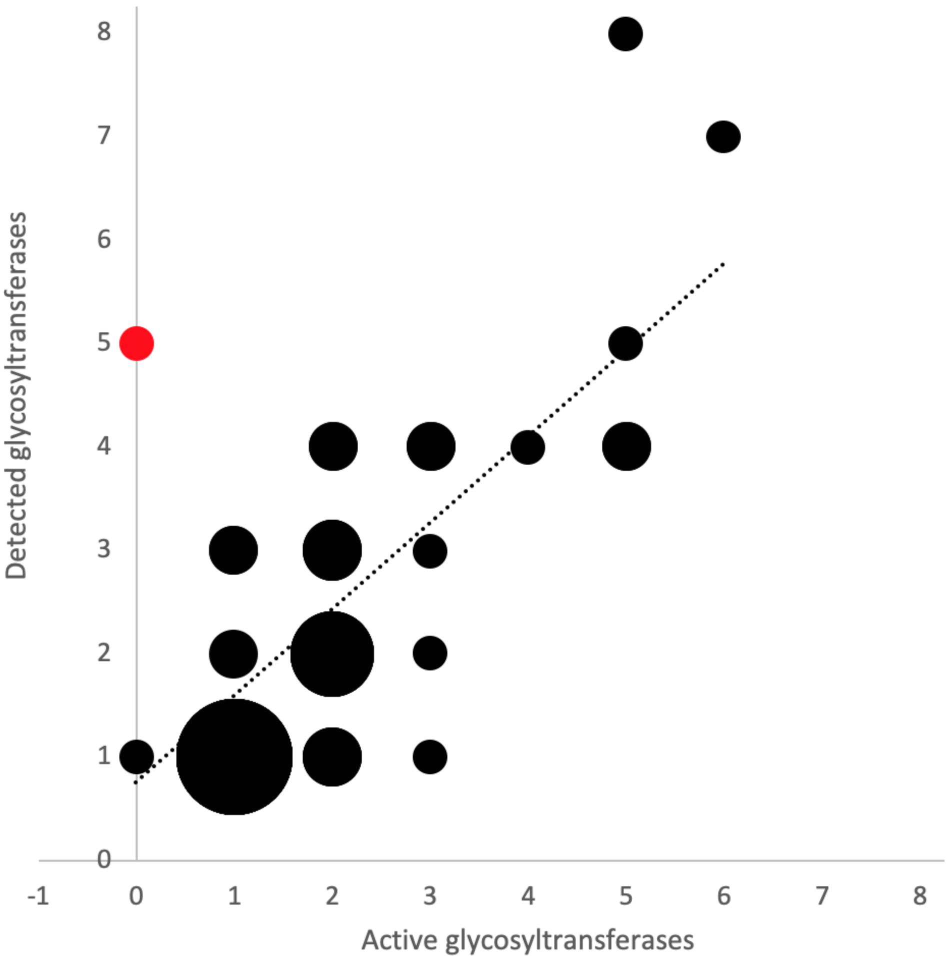


Figure (4.2) Y = 0,84x + 0,75 → R^2 = 0.49. When komodoquinone B (point (0,5), marked in red on the Figure is excluded, these became Y = 0,95x + 0,40 → R^2 = 0.64 . Points are scaled so that their areas are proportional to the frequency of observation.

Table (4.4) The number of detected glycosyltransferases (Det) found within a BGC, and the actual number of glycosyl groups on the experimentally determined structure of the secondary metabolite it encodes (Act).

BGC#	Secondary metabolite	Act	Det	Act - Det
2	Aculeximycin[72]	5	8	3
21	Apoptolidin[104]	3	2	-1
33	Calicheamicin[85]	4	4	0
34	Candicidin[135]	1	1	0
35	Chalcomycin[136]	2	1	-1
36	Chlorothricin[86]	2	2	0
42	Cremimycin[137]	1	1	0
52	ECO-02301[138]	1	1	0
54	Erythromycin B[139]	2	2	0
78	Inecidine[109]	2	3	1
81	Kedarcidin[75]	2	2	0
82	Kijanimicin[71]	5	5	0
85	Lankamycin[140]	2	3	1
92	Megalomicins[141]	3	3	0
96	Midecamicin[111]	2	1	-1
102	Mycinamicin II[142]	2	1	-1
105	Nanchangmycin[143]	1	1	0
108	Natamycin[144]	1	1	0
115	Nystatin A1[145]	1	1	0
136	Rifamycin[108]	1	1	0
141	Rubradirin[74]	1	2	1
148	A83543A[146]	2	2	0
151	Stambomycin A[147]	1	1	0
162	Tetrocarcin A[73]	5	4	-1
165	Tiacumicin B[148]	2	2	0
167	Vicenistatin[149]	1	1	0
197	Aranciamycin[150]	1	1	0
198	Arenimycin A/B/C[151]	2	2	0
199	ArimetamycinA[78]	2	3	1
199	ArimetamycinB[78]	1	3	2
199	ArimetamycinC[78]	1	3	2
200	Arixanthomycin A[152]	1	2	1
203	BE-7585A[153]	3	1	-2
208	Chelocardin[154]	0	1	1
210	Chromomycin A3[79]	5	4	-1
1183	Lobophorin[126]	3	4	1
1452	Sipanmycin[155]	2	4	2
1522	Auroramycin[156]	2	4	2
1619	Ibomycin[157]	6	7	1
1851	Komodoquinone B[80]	0	5	5
2033	Spiramycin[123]	3	4	1

4.1.3 2-Amino-3-hydroxy-cyclopenta-2-enone

Post core structure synthesis addition of 2-amino-3-hydroxy-cyclopenta-2-enone stood out among tailoring reactions, as it was found as a substrate in the metabolic pathway of several BGCs. The gene synthesising 5-aminolevulinate (the precursor of 2-amino-3-hydroxy-cyclopenta-2-enone) was found to have smCOG 1109:8-amino-7-oxononanoate synthase in the bafilomycin gene cluster. The antiSMASH output of all MIBiG clusters was therefore searched for genes with this smCOG. In order to further investigate the role of the genes with smCOG 1109:8-amino-7-oxononanoate synthase, the smCOGs of neighboring genes was found to be relevant. If a neighboring gene had smCOG 1002:AMP-dependent ligase and synthetase, this was indicative of a reaction involving glycine and acyl units in 7 out of 8 cases (Tables 4.5 and 4.6). Interestingly, not all the reaction pathways these gene were involved in were identical. Instead, they utilised highly similar substrates (malonyl-CoA vs. succinyl-CoA).

Table (4.5) BGCs containing a gene with smCOG 1109:”8-amino-7-oxononanoate synthase”. * Does the ”8-amino-7-oxononanoate synthase”-gene neighbor a gene with putative AMP-binding function?. ** Does the reaction utilise glycine + acyl unit(s) as substrates? Data acquired from [27]

BGC #	BGC	*	**
28	Bafilomycin	Yes	Yes
62	Fumonisin	No	No
91	Marineosin	Yes	Yes
187	Asukamycin	Yes	Yes
213	Colabomycin E	Yes	Yes
1063	Undecylprodigiosin	Yes	Yes
1215	Conglobatin	No	No
1298	Annimycin	Yes	Yes
1420	Myxochromide	Yes	No
1633	Ketomemicin B3	No	No
1740	Phthoxazolin	Yes	Yes

Table (4.6) The reactions that the gene with smCOG 1109:8-amino-7-oxononanoate synthase participates in. ”PK” implies that the polyketide intermediate participates in the reaction.

BGC	Associated reaction
Bafilomycin[38]	PK_1 + Gly + succinyl-CoA + ATP → PK_2 + AMP + PP_i + CO_2 + H_2O + CoA
Fumonisin[158]	Ala + PK→PK-NH2+ CO_2
Fumonisin[158]	Ala + PK→PK-NH2+ CO_2
Marineosin[159]	PK_1 + Gly + 2 malonyl-CoA + ATP → PK_2 + AMP + PP_i + $2CO_2$ + H_2O + 2CoA
Asukamycin[76]	PK_1 + Gly + succinyl-CoA + ATP → PK_2 + AMP + PP_i + CO_2 + H_2O + CoA
Colabomycin E[83]	PK_1 + Gly + succinyl-CoA + ATP → PK_2 + AMP + PP_i + CO_2 + H_2O + CoA
Undecylprodigiosin[160]	PK_1 + Gly + 6 malonyl-CoA + ATP → PK_2 + AMP + PP_i + $6CO_2$ + H_2O + 6CoA
Conglobatin[161]	not characterised
Annimycin[77]	PK_1 + Gly + succinyl-CoA + ATP → PK_2 + AMP + PP_i + CO_2 + H_2O + CoA
Myxochromide[162]	not characterised
Ketomemicin B3[163]	PK + Phe → CO_2 + PK
Phthoxazolin[164]	PK_1 + Gly + succinyl-CoA + ATP → PK_2 + AMP + PP_i + CO_2 + H_2O + CoA

4.2 Extender unit specificities

In order to better understand which extender unit predictions can be expected from antiSMASH, the antiSMASH outputs for all BGCs in MIBiG was analyzed, and the sum of predictions for each extender unit was found. The general overview of the results from this data gathering is presented in Figure 4.3 for NRPS and Figure 4.4 for PKS. The specific number of occurrences for each prediction is given in Table 6.4 in Appendix. In total, antiSMASH provided a prediction for 4645 out of 5629 modules, which translates into a rate of 83%.

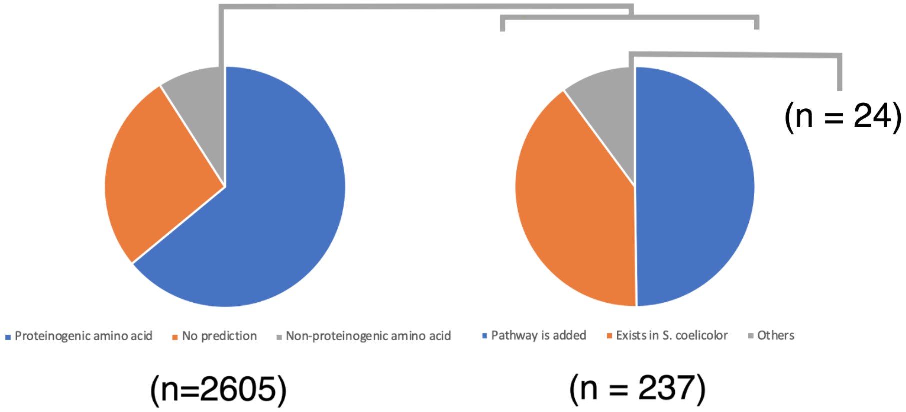


Figure (4.3) Pie chart representation of substrate predictions for NRPS modules. Number of predictions for each pie chart area: (Left) Proteinogenic amino acid = 1667, No prediction = 701, Non-proteinogenic amino acid = 237 (Right) Pathway [synthesising the nonproteinogenic amino acid] is added [to the metabolic pathway] = 118, [The predicted nonproteinogenic amino acid is synthesised by, and therefore] Exists in *S. coelicolor* = 95, Others = 24.

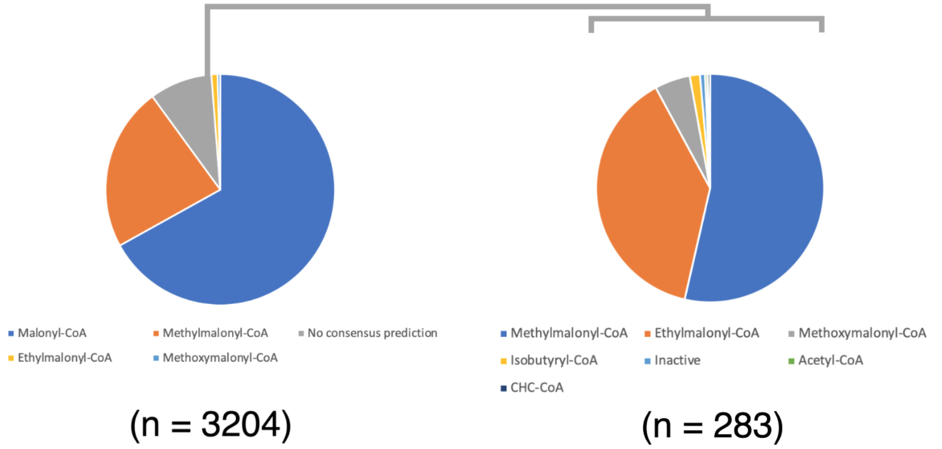


Figure (4.4) Pie chart representation of substrate predictions for PKS modules. Number of predictions for each pie chart area: (Left) Malonyl-CoA = 2146, Methylmalonyl-CoA = 735, No consensus prediction = 283, Ethylmalonyl-CoA = 28, Methoxymalonyl-ACP = 12. Right: Minowa predictions in the case where minowa and AT_signature predictions disagreed. Methylmalonyl-CoA = 149, Ethylmalonyl-CoA = 107, methoxymalonyl-ACP = 14, Isobutyryl-CoA = 4, Inactive = 2, Acetyl-CoA = 1, CHC-CoA = 1.

4.3 Comparison of predicted and experimentally determined clusters

In order to test the viability of the algorithm, 8 different BGCs were examined. The predicted activity of each domain for all BGCs was aligned against its experimentally determined activity. Only the alignment for bafilomycin is shown in this section, the rest are given in Appendix (Tables 6.5 - 6.11). In this project we have used python 3.7.1 and the COBRApy v0.17.1 python package to run FBA with a GEM with heterologous BGCs inserted into the model. We have used the recently updated GEM of *S. coelicolor* as a template model [165]. This GEM was selected because the majority of PKS/NRPS BGCs have been found within its genus, and because the GEM of *S. coelicolor* is of high quality. The FBA results presented in this section were obtained by first setting the objective function of the GEM to production of biomass in order to find the maximum production of biomass. Then, the lower bound of the biomass flux was set to 95% of the maximum biomass flux, and the objective function was changed to maximise production of the secondary metabolite. The results of the FBA are summarised in Table 4.16. The differences between the predicted and experimentally determined pathway that are specific for each BGC will be briefly discussed in this chapter. Differences between constructed and experimentally determined pathways in general are discussed in chapter 5.

4.3.1 Bafilomycin B1

Bafilomycin B1 is a cluster found in *Streptomyces lohii* and *Kitosatospora setae*. This allows for the comparison of BGCs with different nucleotide sequences, but with identical experimentally determined pathways. The experimentally determined biosynthetic pathway can be seen in Figure 2.1. The alignment of the predicted vs experimentally determined pathway is given in Table 4.8. A summary of the precursor substrates used for its synthesis is given in Table 4.7.

The pathway of the BGC found in *S. lohii* was for the most part correctly predicted, except for the prediction of methylmalonyl-CoA instead of methoxymalonyl-ACP in module 11. The same error was observed for K.Setae, in addition to the incorrect prediction of malonyl-CoA instead of methylmalonyl-CoA in module 9. The addition of 2-amino-3-hydroxy-cyclopenta-2-enone in a tailoring reaction is predicted for both K.Setae and S.Lohii, based on the smCOG definition of the genes found in the cluster. This reaction is also observed for the real bafilomycin cluster. One tailoring reaction is neglected, which is the binding of fumaryl-AMP to the core structure.

Table (4.7) Comparison between experimentally determined (Real) and constructed pathway, showing the number of different substrate molecules that are used for synthesis of bafilomycin B1.

Substrate	Real S.Lohii	Constructed S.Lohii	Constructed K.Setae
Isopropyl-CoA	1	0	0
Malonyl-CoA	2	4	5
Methylmalonyl-CoA	7	7	6
Methoxymalonyl-ACP	2	1	1
NADPH + H+	11	11	11
H2O	-4	-4	-4

Table (4.8) Comparison between the computationally predicted and experimentally determined pathway of bafilomycin (BGC0000028) in *S. lohii* as well as the predicted pathway of *K. setae*. Each complete module is separated with horisontal lines. ”Real” represents the reaction that has been experimentally determined to occur. ”Constructed” represents the reactions that the software includes in the constructed metabolic pathway. Domains that have their reactions incorrectly predicted are highlighted in red. (The polyketide intermediate is also a substrate in all reactions)

Module #	Domain sequences			Domain reactions		
	Real *S. lohii*	Constructed *S. lohii*	Constructed *K. setae*	Real *S. lohii*	Constructed *S. lohii*	Constructed *K. setae*
Load	AT	AT	AT	Isopropyl-CoA → CO_2 + CoA	Malonyl-CoA	Malonyl-CoA
1	KS KR	KS KR	KS KR	Methylmalonyl-CoA NADPH + H+ → NADP+	Methylmalonyl-CoA NADPH + H+ → NADP+	Methylmalonyl-CoA NADPH + H+ → NADP+
2	KS KR	KS KR	KS KR	Malonyl-CoA → CO_2 + CoA NADPH + H+ → NADP+	Malonyl-CoA → CO_2 + CoA NADPH + H+ → NADP+	Malonyl-CoA → CO_2 + CoA NADPH + H+ → NADP+
3	KS	KS	KS	Methylmalonyl-CoA → CO_2 + CoA	Methylmalonyl-CoA → CO_2 + CoA	Methylmalonyl-CoA → CO_2 + CoA
4	KS KR	KS KR	KS KR	Methylmalonyl-CoA → CO_2 + CoA NADPH + H+ → NADP+	Methylmalonyl-CoA → CO_2 + CoA NADPH + H+ → NADP+	Methylmalonyl-CoA → CO_2 + CoA NADPH + H+ → NADP+
5	KS KR	KS KR	KS KR	Methoxymalonyl-ACP → CO_2 + ACP NADPH + H+ → NADP+	Methoxymalonyl-ACP → CO_2 + ACP NADPH + H+ → NADP+	Methoxymalonyl-ACP → CO_2 + ACP NADPH + H+ → NADP+
6	KS DH KR	KS DH KR	KS DH KR	Malonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+	Malonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+	Malonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+
7	KS DH KR	KS DH KR	KS DH KR	Methylmalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+	Methoxymalonyl-CoA→ CO_2 + CoA → H2O NADPH + H+ → NADP+	Methoxymalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+
8	KS DH ER KR	KS DH ER KR	KS DH ER KR	Methylmalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+ NADPH + H+ → NADP+	Methylmalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+ NADPH + H+ → NADP+	Methylmalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+ NADPH + H+ → NADP+
9	KS KR	KS KR	KS KR	Methylmalonyl-CoA → CO_2 + CoA NADPH + H+ → NADP+	Methylmalonyl-CoA → CO_2 + CoA NADPH + H+ → NADP+	Malonyl-CoA → CO_2 + CoA NADPH + H+ → NADP+
10	KS DH KR	KS DH KR	KS DH KR	Methylmalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+	Methylmalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+	Methylmalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+
11	KS DH KR TE	KS DH KR TE	KS DH KR TE	Methoxymalonyl-ACP → CO_2 + ACP → H2O NADPH + H+ → NADP+ H2O →	Methylmalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+ H2O →	Methylmalonyl-CoA → CO_2 + CoA → H2O NADPH + H+ → NADP+ H2O →

The flux through real pathway was found to be nearly identical to the predicted pathways of both *K. setae* and *S. lohii*, both having a flux ratio of 1.0 between experimentally determined and predicted pathway.

4.3.2 Difficidin

Difficidin is a transAT-PKS found in the FZB42 strain of *Bacillus velezensis*. It incorporates an apparent acrylyl group as its starter unit, a substrate not found in the *S. coelicolor* model. However, the origin of the acrylyl moiety has been attributed to the reduction of a three-carbon glycosidic pathway intermediate by an enzyme encoded within the BGC [52]. For the purposes of comparison, the three carbon unit is assumed to be 1,3-biphosphoglycerate. The predicted domains and modules from antiSMASH is shown in Figure 4.5. The experimentally determined biosynthetic pathway is shown in Figure 4.6. A summary of the precursor substrates used for its synthesis is given in Table 4.9.

Table (4.9) Summary of the substrates that are used for real and predicted pathway of difficidin. 1,3-biphosphoglycerate is used in place of the unknown glycosidic pathway intermediate that creates the acrylyl moiety. The negative amounts of H_2O result from H_2O being produced by the reaction catalysed by DH domains.

Substrate	Real	Constructed
1,3-biphosphoglycerate	1	0
Malonyl-CoA	12	12
NADPH + H^+	13	11
H_2O	-7	-8
SAM	3	3

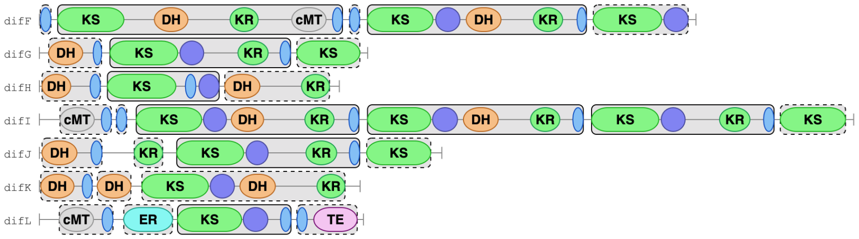


Figure (4.5) antiSMASH representation of all the domains in the modular transAT-PKS of difficidin.

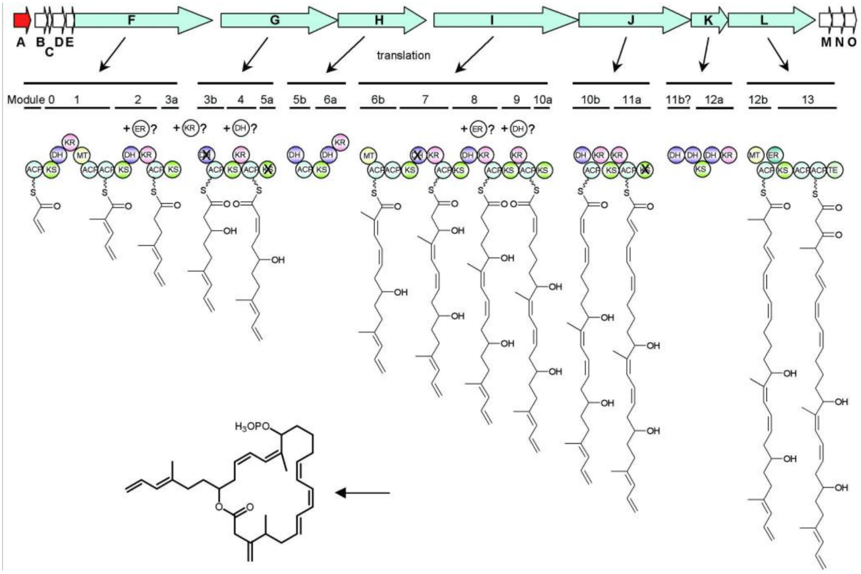


Figure (4.6) Experimentally determined pathway of difficidin. Figure collected from [52]

In addition to the modules predicted by antiSMASH, our software manages to predict some active bridging modules, namely one between difF and difG as well as one between difK and difL (see Figure 4.5), which have both been found experimentally to be active. Other possible bridging modules are the ones between difG and difH, difI and difJ, and difJ and difK, however these are excluded by the software, as their function is predicted to be solely related to collating the mega enzyme from the translated core genes - They are DHD-modules. One of these DHD-modules has been experimentally determined to be active (between difI and difJ). The alignment of the predicted vs experimentally determined pathway is given in Table 6.5 in Appendix.

The most relevant difference is found at the first KS-domain, where the presumed 1,3-biphosphoglycerate is used as the starter unit, whereas the algorithm assumes malonyl-CoA as the starter unit because it cannot locate a loader module. 3 (NADPH + H$^+$) is not included in the constructed pathway, as two free-standing ER domains and one free-standing KR domain take part in the reaction. The ratio between fluxes is 1.1, and is due to the predicted and experimentally determined pathway only differing by the addition of 1,3-biphosphoglycerate and 2 (NADPH + H$^+$).

4.3.3 Oocydin

The pathway of Oocydin contains 16 total KS domains, 7 out of which have been experimentally found to be inactive. As antiSMASH does not currently have any way to detect inactive KS-domains, this necessarily leads to the incorrect prediction of 7 extending steps that should ideally not have been included. The the experimentally determined pathway with inactive KS domains can be seen in Figure 4.7. The cluster contains an FkbM-domain which antiSMASH also does not predict (as it contains no rules for their detection), leading to one SAM being excluded from the substrate base of the constructed pathway. The incorrect prediction of the activities of 6 KS domains is clearly reflected in the observed flux ratio of 0.67. A summary of the precursor substrates used for its synthesis is given in Table 4.10. The full alignment of domain activities is given in Table 6.6 in Appendix.

Table (4.10) The number of different substrate molecules that are used for synthesis of oocydin

Substrate	Real	Constructed
1,3-biphosphoglycerate	1	1
Malonyl-CoA	9	16
NADPH + H+	7	6
SAM	2	1
H2O	-4	-3

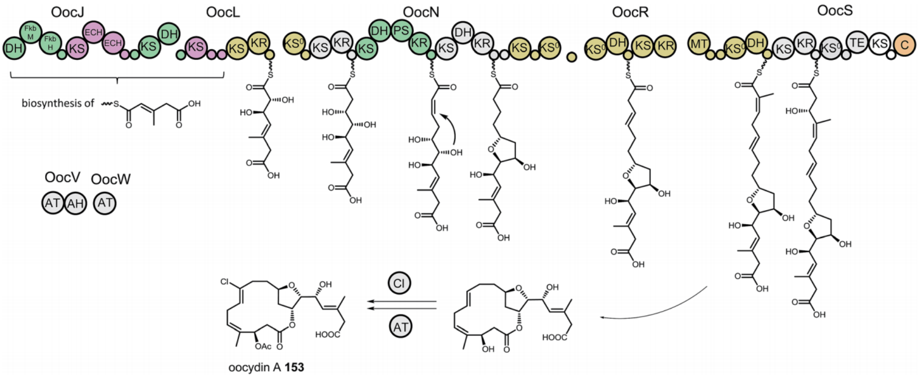


Figure (4.7) Experimentally determined pathway of Oocydin. Figure collected from [30].

4.3.4 Leupyrrin

Leupyrrin is a hybrid NRPS/T1PKS cluster. The experimentally determined pathway of the core structure is given in Figure 4.8, and the tailoring reactions is given in Figure 4.9. Domain sequence as predicted by antiSMASH is given in Figure 4.10. Alignment of predicted domain activities against experimentally determined activities is given in Table 6.7. For leupyrrin and the clusters in Sections 4.3.5 - 4.3.8, there were predicted many different substrates. Because of this, it was found more relevant to present an overview of the predictions based on the type of domain responsible for the reaction e.g. correct prediction of extender units and correct prediction of cofactor-associated reactions. For leupyrrin, these are given in Table 4.11.

Table (4.11) Correct predictions and incorrect predictions for extender units and cofactor-associated reactions for constructed vs. experimentally determined pathway of leupyrrin.

	Correct	Incorrect	Total
Extender units	4	3	7
Cofactors	11	0	11
Sum	15	3	18

The pathway of leupyrrin is one that highlights the complexity that can be found for metabolic pathways encoded by BGCs. First of all, the mechanism behind the loading of the starter unit is non-traditional. The A domain on Leu5 binds proline to a carrier protein which subsequently transferred to the PCP domain on Leu9 by Leu6. As the algorithm only looks for A-PCP as a starter module, the C-PCP module is not detected. Instead the starter unit is assumed as malonyl-CoA. Secondly, the leupyrrin cluster incorporates a novel extender unit into its structure, which the algoritm is not able to predict and instead assumes malonyl-CoA as the extender unit. Malonyl-CoA and isovaleryl-CoA are the substrates used for the synthesis of this extender unit, so the effective difference in the pathway is the exclusion of isovaleryl-CoA. Finally, three tailoring reactions are not accounted for, missing the substrates 4-Methyl-2-oxopentanoate, acetyl-CoA and geranyl diphosphate.

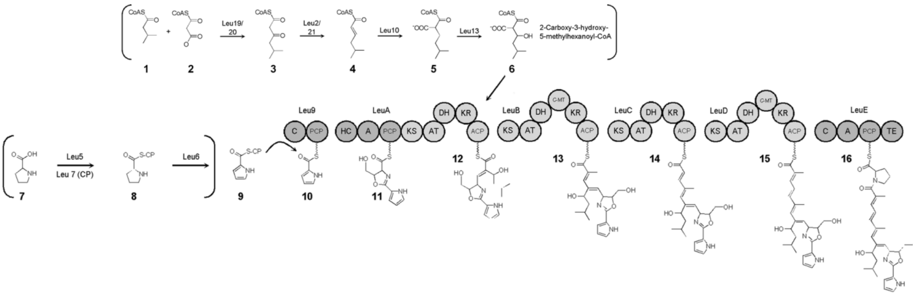


Figure (4.8) The biosynthetic pathway of the core structure encoded by the leupyrrin BGC (10-16). Non-core genes are responsible for the synthesis of the unusual metabolite 2-carboxy-3-hydroxy-5-methylhexanoyl-coa (6) which is used as an extender unit in the polyketide synthesis. Figure collected from [56]

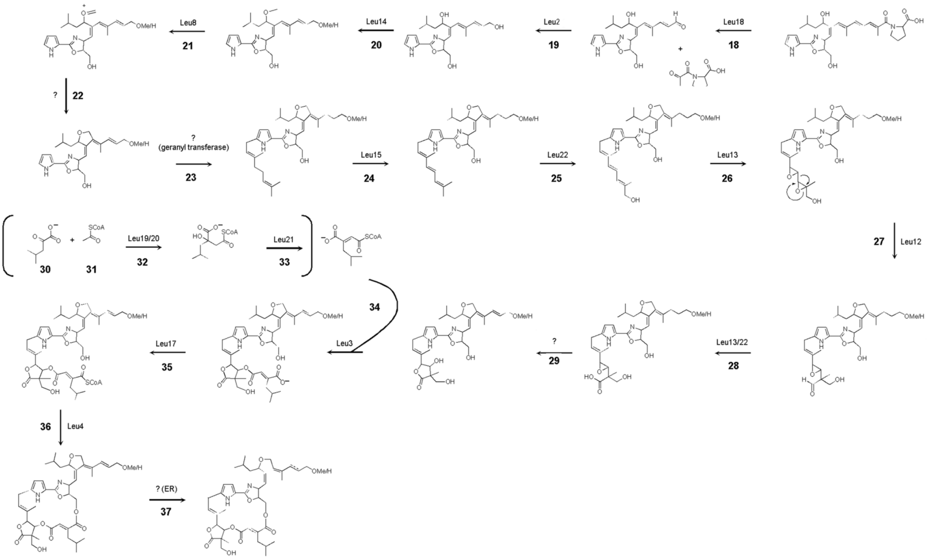


Figure (4.9) The biosynthetic pathway of the tailoring reactions encoded by the leupyrrin BGC (10-16). Figure collected from [56]

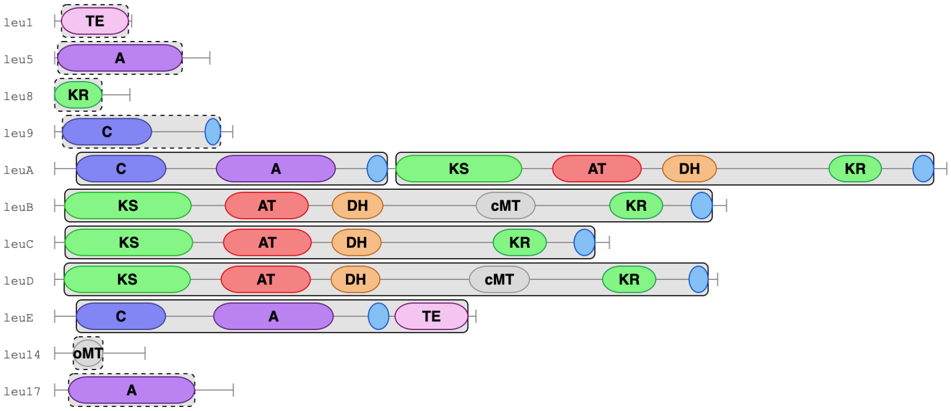


Figure (4.10) All domains in the leupyrrin BGC as predicted by antiSMASH.

Because of all the missing tailoring reactions, the flux ratio between real and predicted pathway for leupyrrin is the highest that is observed for all clusters: 2.6.

4.3.5 Anabaenopeptin

Anabaenopeptin is another example of strange behaviour in the modular structure of NRPS. In this BGC there are two consecutive genes in the BGC that each may function as the loader unit for the synthesis, albeit only one at a time (Figure 4.11). This leads to antiSMASH predicting an 8 amino acid long oligopeptide when the real secondary metabolite only consists of 6. The number of correct predictions for extender units and cofactor-associated reactions is given in Table 4.12.

Table (4.12) Correct predictions and incorrect predictions for extender units and cofactor-associated reactions for constructed vs. experimentally determined pathway of anabaenopeptin.

	Correct	Incorrect	Total
Extender units	5	3	8
Cofactors	1	0	1
Sum	6	3	9

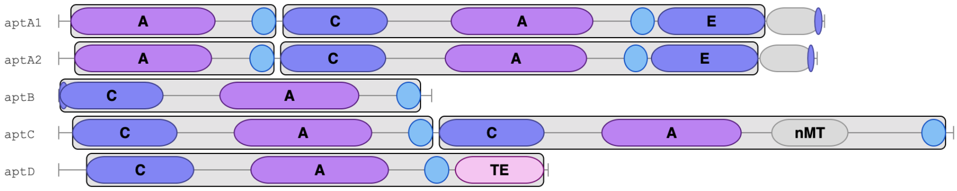


Figure (4.11) Modular structure of the anabaenopeptin BGC. Both genes aptA1 and aptA2 may function as the first of the proteins that together make up the giant multimodular enzyme, however only one at the same time. the load module of aptA1 (first module from the left) may recruit both L-arginine and L-lysine as the starter unit, while the load module of aptA2 is specific to L-tyrosine

Four of the modules are predicted as having unknown specificity, while the other four have their specificity correctly predicted (Table 6.9 in Appendix). FBA revealed a ratio of 0.40 between real and constructed pathway. One would expect the flux to be higher for the predicted flux because of the two additional amino acids that are required in the predicted pathway. However, a 60% decrease in flux is unprecedented. This can only arise from there being associated an additional ”cost” with how the prediction method handles unknown amino acid predictions. Due to time limitations, this has not been further investigated.

4.3.6 Tolaasin

There is little to note about the predicted pathway of tolaasin. There are no tailoring reactions, and no uncommon loader modules in this BGC. The substrate specificities of 16 out of the 18 modules are correctly predicted. Domain reaction alignments are given in Table 6.8 in Appendix. This is reflected by the flux ratio between real and predicted pathway which is 1.0. The number of correct predictions for extender units and cofactor-associated reactions is given in Table 4.13

Table (4.13) Correct predictions and incorrect predictions for extender units and cofactor-associated reactions for constructed vs. experimentally determined pathway of tolaasin.

	Correct	Incorrect	Total
Extender units	16	2	18
Cofactors	0	0	0
Sum	16	2	18

4.3.7 Geldanamycin

Geldanamycin is a T1PKS that incorporates methoxymalonyl-ACP into its structure. It is the only BGC among the 8 mentioned here where the predicted pathway completely agrees with the experimentally determined pathway. Domain reaction allignments are given in Table 6.10. As the two pathways are identical, the flux ratio is necessarily 1.0 as well. The number of correct predictions for extender units and cofactor-associated reactions is given in Table 4.14

Table (4.14) Correct predictions and incorrect predictions for extender units and cofactor-associated reactions for constructed vs. experimentally determined pathway of geldanamycin.

	Correct	Incorrect	Total
Extender units	7	0	7
Cofactors	16	0	16
Sum	23	0	23

4.3.8 Oxazolomycin

Oxazolomycin is a transAT-PKS that also includes methoxymalonyl-ACP as one of its extender units. However, the algorithm fails to include methoxymalonyl-ACP as a substrate in the constructed metabolic pathway, as the rules implemented for swapping extender unit specificities do not trigger. In addition, the prediction fails to recognise one formylation domain, as well as one MT domain, leading to the predicted flux of the secondary metabolite to be higher than the real flux. the flux ratio of 1.8 is so high because the predicted pathway fails to include formylation. If the formylation reaction is excluded from the real pathway, the flux ratio becomes 0.8 between experimentally determined and predicted pathway. Domain reaction alignments are given in Table 6.11 in appendix. The number of correct predictions for extender units and cofactor-associated reactions is given in Table 4.15

Table (4.15) Correct predictions and incorrect predictions for extender units and cofactor-associated reactions for constructed vs. experimentally determined pathway of oxazolomycin.

	Correct	Incorrect	Total
Extender units	11	1	7
Cofactors	21	3	16
Sum	32	4	36

4.3.9 Reaction prediction accuracies

In total 177 out of 215 domains had its associated reaction correctly predicted. Out of these 215 domains, 92 were associated with an extender unit and 123 were associated with cofactors. Pie chart representations of the results are given in Figure 4.12.

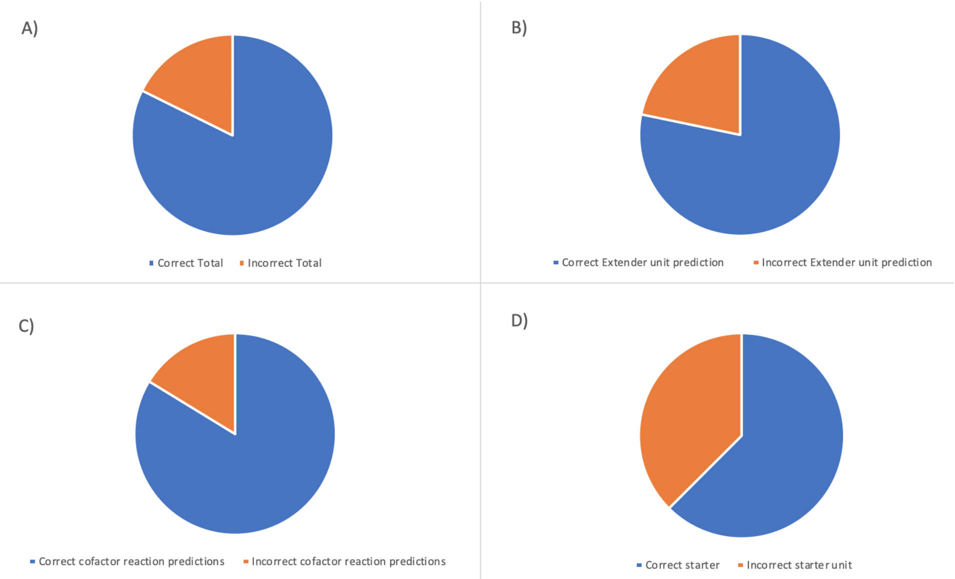


Figure (4.12) Predicted activity VS. experimentally determined activity was examined for 215 domains from 8 BGCs. A) 177 out of 215 total domains were correctly predicted (82%). B) 72 out of 92 predictions for the extender unit was correct (78%). C) 103 out of 123 non-extending domains was correctly predicted (84%). D) 5 out of 8 starter units were correctly predicted (63%).

4.3.10 Flux balance analysis

We used flux balance analysis to estimate the maximal flux through each metabolic pathway as predicted by our software and as detailed in the literature. By comparing the predicted fluxes for each of the 8 BGCs we get a quantitative evaluation of the accuracy of the algorithm complementary to the qualitative comparison of starter units, extender units and cofactors given above. Flux balance analysis comparisons for the 8 clusters examined is given in Table 4.16. The specific flux is not that interesting, the point of interest lies in comparing real against predicted flux, i.e. the ratio. For 4 of the 8 cluster the predicted flux is identical, or very similar, for the predicted and experimentally determined pathways. For the remaining 4 clusters the predicted flux is not as accurate, but at least in the right order of magnitude. One due to inability to predict inactive KS domains, one due to not exhibiting collinearity, one due to inability to predict tailoring reactions, and one due to a formylation domain not being detected by antiSMASH.

Table (4.16) Flux balance analysis for the predicted VS experimentally determined pathways. ”Real” is the flux of product through the experimentally determined pathway. ”Constructed” is the flux through the constructed pathway. The ratio of predicted/real flux is also given. Units are mmol per gram dry weight of cell per hour. * Bafilomycin from *S. lohii*. ** Bafilomycin from *K. setae*

BGC	Real	Predicted	Ratio
Bafilomycin*	0.0046	0.0047	1.0
Bafilomycin**	0.0046	0.0046	1.0
Difficidin	0.0054	0.0061	1.1
Oocydin	0.0076	0.0051	0.67
Leupyrrin	0.0031	0.0083	2.6
Anabaenopeptin	0.0139	0.0055	0.40
Tolaasin	0.0023	0.0024	1.0
Geldanamycin	0.0063	0.0063	1.0
Oxazolomycin	0.0029	0.0052	1.8

Chapter 5

Discussion

This chapter aims to discuss:

- 1) Which extender units the algorithm supports, and which will be disregarded when constructing the metabolic pathway.
- 2) The ratio between real and predicted flux, as well as the variance in prediction quality that is observed for different clusters.
- 3) The disregarding of most tailoring reactions, and to what degree they affect the constructed metabolic pathway.
- 4) The limitations of this method of constructing the metabolic pathway of a BGC, and some suggestions on how it can be improved.

5.1 Substrate predictions

There was found an average 82% overall accuracy in predicting the correct reaction associated for a PKS/NRPS domain, 78% accuracy in predicting substrate specificity for extender units, and 84% accuracy for cofactor-associated reactions. antiSMASH predicts the specificity of a module at a rate of 83% (i.e. gives a prediction at all, not necessarily a correct prediction). Because there is some uncertainty associated with the prediction method that antiSMASH uses, the 78% correct prediction rate for extender units is most likely inflated by having (randomly) selected our 8 BGCs among those that antiSMASH has made accurate predictions for. The correct prediction rate is presented anyways, in order to confirm that the algorithm does what it is supposed to do, while also confirming that reactions related to non-essential domains are predicted at an acceptable rate. As the predictions on substrate specificity is made by antiSMASH, the uncertainty in the predictions of specific substrates can be found by reviewing the antiSMASH documentation ([19–23]).

5.1.1 Extender units

To introduce this subsection, an example scenario is given: A module is predicted to incorporate the nonproteinogenic amino acid HAORN into the core structure. However, HAORN does not exist in the *S. coelicolor* GEM as a metabolite. How does the algorithm handle this case, and does this happen often enough that it will significantly impact the quality of the metabolic pathways that are constructed?

The extender unit predictions that can be directly added to the metabolic pathway are: proteinogenic amino acids, malonyl-CoA, ethylmalonyl-CoA, methylmalonyl-CoA, methoxymalonyl-ACP, HPG, BHT, Pipecolic acid and DHPG. The latter 5 do not exist in the *S. coelicolor* GEM, so when they are added, they are assumed to be synthesised by other genes in the BGC, and a reaction pathway that synthesises them is therefore added to the constructed metabolic pathway.

For NRPS, 1904 out of 2605 modules had their substrate specificities predicted (Figure 4.3). Out of these 1904, 1667 of the predictions were proteinogenic amino acids. This left 237 modules that were predicted as non-proteinogenic amino acids. 95 of the non-proteinogenic amino acids exist in the *S. coelicolor* as primary metabolites (e.g. Ornithine, DPG etc.), and 118 are accounted for by adding the pathway that synthesises them to the constructed metabolic pathway (i.e. HPG, DHPG, Pipecolic acid and BHT). This leaves only 24 modules with other predictions. These were therefore deemed rare enough that they could be ignored. The solution to this was to effectively set the substrate specificity of these modules to ”unknown amino acid”, and treating them as such, described in Section 3.6. There are still 701 cases where the substrate specificity cannot be predicted, but this is a problem that lies with the substrate prediction method, and not with the method of constructing the pathway.

For the PKS predictions, there were found 283 modules that had no prediction out of 3204 total domains (Figure 4.4). These modules were all assumed to incorporate malonyl-CoA, defaulting to the ”AT-signature” method of prediction. Out of the 283 modules, only 8 (Isobutyryl-CoA = 4, Inactive = 2, Acetyl-CoA = 1, CHC-CoA = 1) were predicted to have a substrate specificity other than malonyl-CoA, methylmalonyl-coa, methoxymalonyl-ACP, ethylmalonyl-CoA by any of the two methods of prediction (minowa and AT-specificity). It was therefore deemed unnecessary to include specific rules to add these uncommon extender units to the constructed metabolic pathways.

5.1.2 Starter units

Out of the 26 different alternative starter units that have been observed for NRPS/PKS (Tables 6.1, 6.2 and 6.3 in Appendix), the synthesis of 24 of them are included within the BGC. Except for AHBA, none of the other starter units in the table can be predicted by the algorithm.

Although it is unfortunate that the substrate specificity of 3 out of 8 starter modules are incorrectly predicted, the substitution of a starter unit with e.g. malonyl-CoA does not seem to completely obfuscate the pathway, as can be seen for the pathway of bafilomycin. Here, the starter unit isopropyl-CoA was incorrectly predicted as malonyl-CoA. Still, the ratio between the real and predicted pathway was close to 1.

In addition to proteinogenic amino acids, and other products of glycolysis, a common substrate for synthesising these alternative starter units is shikimic acid. Shikimic acid is also a precursor for the most frequently predicted nonproteinogenic amino acid HPG, as well as tyrosine. Tyrosine is further a precursor to BHT, another common nonproteinogenic amino acid. These results show that shikimic acid may be an interesting target metabolite with respect to metabolic engineering of BGCs for the overproduction of certain secondary metabolites.

5.2 Flux balance analysis

BGCs Bafilomycin, Difficidin, Tolaasin and geldanamycin showed a close approximation to the experimentally determined pathway. On the other hand, Oocydin and anabaenopeptin gave around half, while leupyrrin and oxazolomycin gave twice the flux of the experimentally determined pathway. While the production rate of the secondary metabolite is somewhat arbitrary as a measure for the prediction accuracy, it shows that there is a relatively large variance in the accuracy of the constructed metabolic pathways. This section attempts to describe different types of clusters and how well the algorithm can construct their pathways.

5.2.1 BGCs exhibiting collinearity

The metabolic pathway of some BGCs, such as difficidin, can be determined with great accuracy because they abide the collinearity rule described in the Theory. In addition, there are no tailoring reactions associated with the pathway of the secondary metabolite. Adding to that, the difficidin BGC is a transAT-PKS, meaning that there is no ambiguity as to which extender units is incorporated into its core structure.

Slightly more complicated are the collinear T1PKS and NRPS BGCs without any tailoring genes - those that are ambiguous as to which extender units are incorporated. Two examples of such BGCs are geldanamycin and tolaasin. Here there is uncertainty in the predictions that antiSMASH gives which could lead to the incorrect prediction of the pathway.

Even more uncertainty is associated with the pathway of BGCs that encode tailoring enzymes, such as leupyrrin and bafilomycin. For leupyrrin, 4-Methyl-2-oxopentanoate is one of the substrates that is added in a tailoring step but is completely ignored by this algorithm. This results in the predicted production of secondary metabolite being nearly three times higher than the real. In the case of bafilomycin, one of the tailoring reactions is actually predicted, reflected in the observed flux ration of 1.0.

5.2.2 BGCs defying collinearity rule

In all examples mentioned previously, there has been a fairly coherent sequence of domains and modules that can be intuitively understood by reading the sequence of domains. This is not the case for all BGCs, as some do not follow the collinearity rule. One such BGC can be seen in Figure 5.1, containing core genes that have several loader and terminating modules. These types of BGCs will most likely have their metabolic pathway ”overestimated” - more precursor metablites are added to the constructed pathway than is observed for the real pathway. One such example that has been discussed is the anabaenopeptin cluster. For this cluster, two more extender units are added to the constructed metabolic pathway than has been experimentally determined (section 5.1.5, Anabaenopeptin).

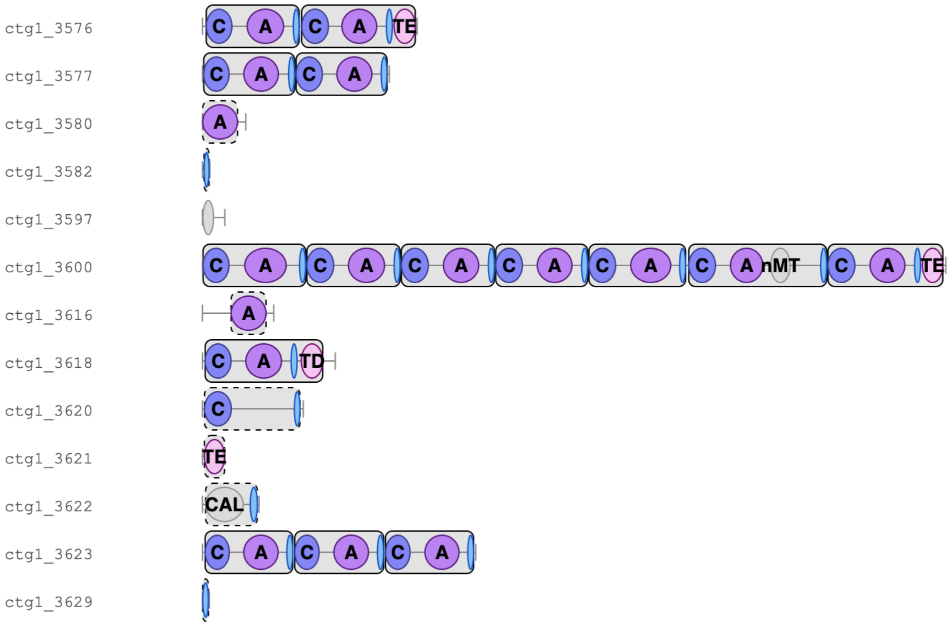


Figure (5.1) A non-characterised BGC found within *Actinosynnema pretosium*. The BGC contains three possible terminating modules (ctg_3576, ctg_3600, ctg_3618) as well as a free-standing TE-domain (ctg_3621). Several valid loader modules are also present (C-PCP on ctg_3620, CAL-PCP on ctg_3622, as well as two free-standing A-domains on ctg_3580 and ctg_3616) suggesting that the BGC may encode several secondary metabolites.

5.2.3 Pseudo-BGCs

On the opposite side of the spectrum, the metabolic pathway can be severely underestimated for certain BGCs. When running antiSMASH on the genome of *Pseudomonas costantinii* - the same organism as the tolaasin cluster was obtained from, 13 other clusters were identified. Among these, an NRPS cluster with a most similar cluster to pyoverdin was found. This cluster stood out due to its relatively small size compared to its most related cluster, having only 9% similar genes. Upon further investigation, another NRPS cluster was found, closely resembling the other segment of the pyoverdin cluster, located around 400kbp upstream of the first cluster. In MIBiG the pyoverdin cluster entry is from the genome of *Pseudomonas protegens* Pf-5. For the MIBiG entry, the distance between the two segments of the cluster is much smaller - around 100 kbp. A comparison between the pyoverdin cluster(s) in *P. costantinii* and *P. protegens* is given in Figure 5.2.

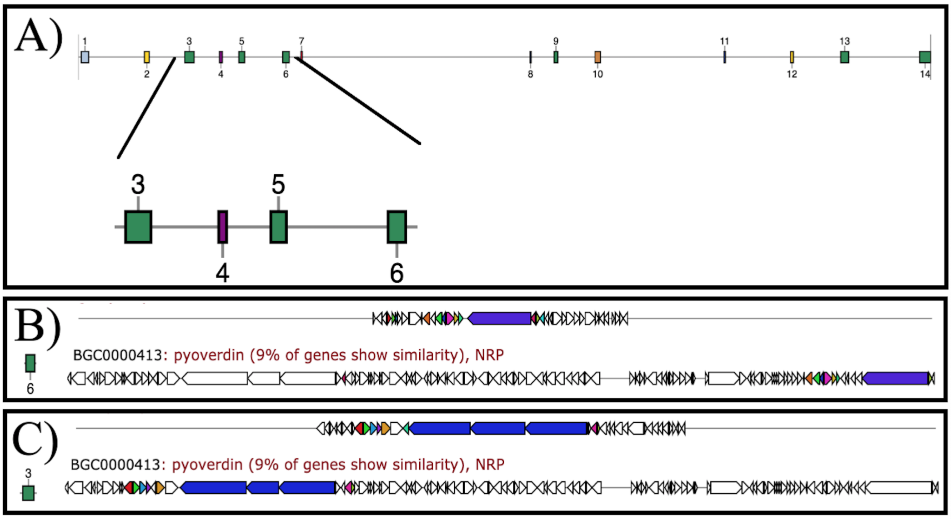

A)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
3
5
4
6
B)
6
BGC0000413: pyoverdin (9% of genes show similarity), NRP
C)
3
BGC0000413: pyoverdin (9% of genes show similarity), NRP

Figure (5.2) A) Location of the 14 BGCs found within *P. costantinii*. The two BGCs in question (BGC 3 and BGC 6) are highlighted. B) BGC 6 and its similarity to the pyoverdin BGC. C) BGC 3 and its similarity to the pyoverdin BGC.

In this case, predicting the pathway of any of the two clusters would lead to the exclusion of the part of the pathway that the other BGC encodes. Little can be done in these cases, except for being aware of their existence.

5.2.4 Glycosyltransferases

There is assumed a one-to-one relationship between number of glycosyltransferases and the number of glycosyl units on the secondary metabolite. Although the R^2 value of 0.49 is not perfect (Figure 4.2), the main take-away should be that there in nearly all cases is incorporated a glycosyl group when a glycosyltransferase is found in the BGC. As long as the substrate is included at all, it will affect which metabolic engineering measures will increase the production of the secondary metabolite.

Arimetamycin 2 and komodomycin B are both examples of exceptions to the one-to-one relationship, however, they are both part of clusters that can synthesize several secondary metabolites. Predicting whether or not a BGC encodes several secondary metabolite products is not discussed further, and is simply noted as one of the many sources of uncertainty associated with modeling biosynthesis of secondary metabolites.

5.2.5 Other tailoring reactions

While there exist other tailoring reactions that require additional substrates and cofactors, none of these are included in the construction of metabolic pathway. smCOG definitions also proved to be too general, as the genes encoding the enzymes catalysing these reactions were assigned broad definitions such as ”acyl-coa ligase” and ”AMP-dependent synthase and ligase" which appear frequently in BGCs and are therefore not good indicators of specific reactions.

5.3 AntiSMASH limitations

As the algorithm relies heavily on the predictions made by antiSMASH, most of the uncertainty in the metabolic pathway lies in these. The databases used to search for AT-signarures and build pHMMs for the minowa prediction method are nearly a decade old. Since then, many AT domains with specificity towards uncommon extender units have been discovered. for instance, the AT-specificity database contains 12 signatures that are specific towards methoxymalonyl-ACP. There are 17 BGCs displayed in Table 4.1 that encorporate methoxymalonyl-ACP into the polyketide structure, and some of them have multiple methoxymalonyl-ACP specific AT-domains, such as geldanamycin and bafilomycin that each have 2 such domains. Another example is that antiSMASH does not predict hydroxymalonyl-ACP as an extender unit, although there are at least 4 BGCs that are known to do so. Still, this extender unit is nowhere to be found in the databases of both minowa and AT-specificity prediction.

During analysis of all clusters in MIBiG, there were found 98 putative CAL domain loader modules. The pHMMs that recognise CAL domain specificities in antiSMASH are built on only 14 CAL domain sequences. Examining the specificities of more CAL domains could lead to observing CAL domains with additional specificities, as well as increased prediction accuracy for the specificity of these domains.

5.3.1 KS-Domains

KS-domains are frequently determined as inactive. One extreme case is the Oocydin BGC in which 7 out of the 16 KS-domains that are part of complete modules (e.g. is followed by an ACP domain preceding the next KS domain) do not extend the polyketide intermediate. Inactive KS-domains are most frequently found in TransAT-PKS systems, and while there are some of them that can be determined to be inactive (rules related to DHD-and oMT-modules), there is still a significant number that are incorrectly predicted as active. One solution to this could be to integrate the functionality of transator - an application that is able to predict such inactive KS domains - into antiSMASH.

5.4 SubClusterBLAST

Unfortunately, SubClusterBLAST does not function properly, leading to difficulty in adding tailoring reactions to the constructed pathway. In theory, the pathway associated with a subcluster could be included into the metabolic pathway. This would account for much better prediction of e.g. specific glycosyl groups that are found in glycosides. This is relevant as the observed glycosyl groups are frequently methylated and aminated, consuming additional substrates such as glycine and SAM, both of which have a non-negligible impact on the metabolic pathway. In addition, it would allow for better prediction of the metabolic pathway of BGCs that incorporate uncommon substrates, either as starter units, or through tailoring reactions.

5.5 Further work

Although the developed software can predict the metabolic pathways with a reasonable accuracy for a range of different BGCs, there is still room for improvement. Because our software relies on the accuracy and output from antiSMASH, many of the suggested improvements should be addressed within the antiSMASH software, These include:

- Updating antiSMASH in order to better predict the substrate specificity of domains, specifically AT, A and CAL domains. Although AT and A domains can be found with ease, their substrate specificities are associated with a fair amount of uncertainty. This is less relevant for malonyl-CoA and methylmalonyl-CoA extender units, as the current methods of detection are built on many such domain sequences. However for more uncommon extender units (e.g. isobutyryl-CoA and methoxymalonyl-ACP), there seems to be much room for improvement by adding only a few of the known AT-domain specificities.
- Updating (and fixing) the SubClusterBLAST functionality of antiSMASH to include more of the subclusters that are related to synthesizing starter units or those that are involved in tailoring reactions. As the algorithm does not currently utilize SubClusterBLAST results from antiSMASH in any way, this would also have to be implemented somehow.
- Implement functionality to detect inactive KS domains in antiSMASH (e.g. transator).

Although not directly related to this project, there is also a need to better be able to model secondary metabolism. Without such methods in place, there is less to be gained from being able to predict the pathway of secondary metabolites in the first place. Finally, we suggest performing FBA on the constructed metabolic pathways in other GEMs in order to deduce the viability of different heterologous host organisms.

Chapter 6

Conclusion

We have in this work built an algorithm that converts information of predicted biosynthetic gene clusters (BGCs) as provided by antiSMASH into metabolic pathways for use in genome scale metabolic models (GEMs). We report an average 82% overall accuracy in predicting the correct reaction associated for a PKS/NRPS domain, 78% accuracy in predicting substrate specificity for extender units, and 84% for cofactor-associated reactions. The algorithm will in all cases for *S. coelicolor* produce a pathway that is feasible, based on the assumption that the metabolic pathway encoded a BGC will contain all genes necessary to produce a secondary metabolite. Overall, there is large variance in the quality of the metabolic pathways that are constructed - some are highly accurate, while others are fairly inaccurate.

In an attempt to predict the more elusive tailoring reactions, we uncovered a relationship between the presence of genes with specific predicted functions, and their roles. We have also attempted to unravel the consequences that wrongful predictions entail, as well as the parts of the constructed metabolic pathways that are neglected by the algorithm. This is achieved through careful examination of experimentally determined metabolic pathways. We find that the metabolites that are missing from the pathway are largely synthesized from primary metabolites, by genes encoded by the BGC, meaning that their exclusion should hopefully not obfuscate the constructed metabolic pathway.

In order to better construct metabolic pathways from BGCs, we suggest that the SubClusterBLAST functionality of antiSMASH is expanded to include additional known tailoring reactions that are found for PKS/NRPS. Also, we suggest updating the databases used for prediction of NRPS/PKS module domain specificity so that the predictions that antiSMASH makes - and in turn the metabolic pathways that the algorithm produces - are more true to their real life counterparts.

Bibliography

1. Fossheim, F. A. *Predicting metabolic pathways form identified biosynthetic gene clusters* tech. rep. (Norwegian University of Science and Technology, Dec. 2019).
2. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* **79,** 629–661. ISSN: 0163-3864. https://doi.org/10.1021/acs.jnatprod.5b01055 (2016).
3. Strieker, M., Tanović, A. & Marahiel, M. A. Nonribosomal peptide synthetases: structures and dynamics. *Current Opinion in Structural Biology* **20,** 234–240. https://doi.org/10.1016/j.sbi.2010.01.009 (Apr. 2010).
4. Risdian, C., Mozef, T. & Wink, J. Biosynthesis of Polyketides in Streptomyces. *Microorganisms* **7.** ISSN: 2076-2607. https://www.mdpi.com/2076-2607/7/5/124 (2019).
5. Ray, L. & Moore, B. S. Recent advances in the biosynthesis of unusual polyketide synthase substrates. *Natural product reports* **33.** 26571143[pmid], 150–161. ISSN: 1460-4752. https://www.ncbi.nlm.nih.gov/pubmed/26571143 (2016).
6. Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Molecular Microbiology* **25,** 619–637. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2958.1997.4821861.x. https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.1997.4821861.x (1997).
7. Martinet, L. *et al.* A Single Biosynthetic Gene Cluster Is Responsible for the Production of Bagremycin Antibiotics and Ferroverdin Iron Chelators. *mBio* **10** (ed Wright, G. D.) eprint: https://mbio.asm.org/content/10/4/e01230-19.full.pdf. https://mbio.asm.org/content/10/4/e01230-19 (2019).
8. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic acids research* **43,** D213–D221 (2014).

9. Wang, B., Guo, F., Dong, S.-H. & Zhao, H. Activation of silent biosynthetic gene clusters using transcription factor decoys. *Nature Chemical Biology* **15,** 111–114. ISSN: 1552-4469. https://doi.org/10.1038/s41589-018-0187-0 (2019).

10. Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nature chemical biology* **11.** 26284661[pmid], 625–631. ISSN: 1552-4469. https://www.ncbi.nlm.nih.gov/pubmed/26284661 (2015).

11. Gräslund, S. *et al.* Protein production and purification. *Nature Methods* **5,** 135–146. ISSN: 1548-7105. https://doi.org/10.1038/nmeth.f.202 (2008).

12. Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biology* **20.** https://doi.org/10.1186/s13059-019-1730-3 (June 2019).

13. Kim, B., Kim, W. J., Kim, D. I. & Lee, S. Y. Applications of genome-scale metabolic network model in metabolic engineering. *Journal of Industrial Microbiology & Biotechnology* **42,** 339–348. ISSN: 1476-5535. https://doi.org/10.1007/s10295-014-1554-9 (2015).

14. King, Z. A., Lloyd, C. J., Feist, A. M. & Palsson, B. O. Next-generation genome-scale models for metabolic engineering. *Current Opinion in Biotechnology* **35.** Chemical biotechnology • Pharmaceutical biotechnology, 23 –29. ISSN: 0958-1669. http://www.sciencedirect.com/science/article/pii/S0958166914002316 (2015).

15. Tran, P. N., Yen, M.-R., Chiang, C.-Y., Lin, H.-C. & Chen, P.-Y. Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. eng. *Applied microbiology and biotechnology* **103.** 30859257[pmid], 3277–3287. ISSN: 1432-0614. https://pubmed.ncbi.nlm.nih.gov/30859257 (2019).

16. Walker, M. C., Mitchell, D. A. & van der Donk, W. A. Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2020/03/15/2020.03.13.990614.full.pdf. https://www.biorxiv.org/content/early/2020/03/15/2020.03.13.990614 (2020).

17. Röttig, M. *et al.* NRPSpredictor2–a web server for predicting NRPS adenylation domain specificity. eng. *Nucleic acids research* **39.** 21558170[pmid], W362–W367. ISSN: 1362-4962. https://pubmed.ncbi.nlm.nih.gov/21558170 (2011).

18. Helfrich, E. J. N. *et al.* Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nature Chemical Biology* **15,** 813–821. ISSN: 1552-4469. https://doi.org/10.1038/s41589-019-0313-7 (2019).

19. Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research* **39,** W339–W346. ISSN: 0305-1048. eprint: http://oup.prod.sis.lan/nar/article-pdf/39/suppl_2/W339/18784565/gkr466.pdf. https://doi.org/10.1093/nar/gkr466 (June 2011).

20. Blin, K. *et al.* antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Research* **41,** W204–W212. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/41/W1/W204/16943787/gkt449.pdf. https://doi.org/10.1093/nar/gkt449 (May 2013).

21. Weber, T. *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research* **43,** W237–W243. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/43/W1/W237/23238788/gkv437.pdf. https://doi.org/10.1093/nar/gkv437 (May 2015).

22. Blin, K. *et al.* antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Research* **45,** W36–W41. ISSN: 0305-1048. eprint: http://oup.prod.sis.lan/nar/article-pdf/45/W1/W36/18137297/gkx319.pdf. https://doi.org/10.1093/nar/gkx319 (Apr. 2017).

23. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research* **47,** W81–W87. ISSN: 0305-1048. eprint: http://oup.prod.sis.lan/nar/article-pdf/47/W1/W81/28879835/gkz310.pdf. https://doi.org/10.1093/nar/gkz310 (Apr. 2019).

24. Weber, T. & Kim, T. Y. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic and Systems Biotechnology* **1** (Feb. 2016).

25. Blin, K. *et al.* The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research* **47,** D625–D630. ISSN: 0305-1048. eprint: http://oup.prod.sis.lan/nar/article-pdf/47/D1/D625/27437289/gky1060.pdf. https://doi.org/10.1093/nar/gky1060 (Nov. 2018).

26. Wilkinson, B. & Micklefield, J. Mining and engineering natural-product biosynthetic pathways. *Nature Chemical Biology* **3,** 379–386. ISSN: 1552-4469. https://doi.org/10.1038/nchembio.2007.7 (2007).

27. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research.* https://doi.org/10.1093/nar/gkz882 (Oct. 2019).

28. Bérdy, J. Bioactive Microbial Metabolites. *The Journal of Antibiotics* **58,** 1–26. ISSN: 1881-1469. https://doi.org/10.1038/ja.2005.1 (2005).

29. Kumelj, T., Sulheim, S., Wentzel, A. & Almaas, E. Predicting Strain Engineering Strategies Using iKS1317: A Genome-Scale Metabolic Model of Streptomyces coelicolor. *Biotechnology Journal* **14,** 1800180. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/biot.201800180.https://onlinelibrary.wiley.com/doi/abs/10.1002/biot.201800180 (2019).

30. Helfrich, E. J. N. & Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural Product Reports* **33,** 231–316. https://doi.org/10.1039/c5np00125k (2016).

31. Schwarzer, D. & Marahiel, M. A. Multimodular biocatalysts for natural product assembly. *Naturwissenschaften* **88,** 93–101. ISSN: 1432-1904. https://doi.org/10.1007/s001140100211 (2001).

32. Amoutzias, G. D., Van de Peer, Y. & Mossialos, D. Evolution and taxonomic distribution of nonribosomal peptide and polyketide synthases. *Future Microbiology* **3.** PMID: 18505401, 361–370. eprint: https://doi.org/10.2217/17460913.3.3.361. https://doi.org/10.2217/17460913.3.3.361 (2008).

33. Chan, Y. A., Podevels, A. M., Kevany, B. M. & Thomas, M. G. Biosynthesis of polyketide synthase extender units. eng. *Natural product reports* **26.** 19374124[pmid], 90–114. ISSN: 0265-0568. https://pubmed.ncbi.nlm.nih.gov/19374124 (2009).

34. Rix, U., Fischer, C., Remsing, L. L. & Rohr, J. Modification of post-PKS tailoring steps through combinatorial biosynthesis. *Natural Product Reports* **19,** 542–580. https://doi.org/10.1039/b103920m (July 2002).

35. Olano, C., Mendez, C. & Salas, J. Post-PKS tailoring steps in natural product-producing actinomycetes from the perspective of combinatorial biosynthesis. *Natural product reports* **27,** 571–616 (Apr. 2010).

36. Walsh, C. T. *et al.* Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. *Current Opinion in Chemical Biology* **5,** 525–534. https://doi.org/10.1016/s1367-5931(00)00235-0 (Oct. 2001).

37. Koskinen, A. M. P. & Karisalmi, K. Polyketide stereotetrads in natural products. *Chem. Soc. Rev.* **34,** 677–690. http://dx.doi.org/10.1039/B417466F (8 2005).

38. Nara, A. *et al.* Characterization of bafilomycin biosynthesis in Kitasatospora setae KM-6054 and comparative analysis of gene clusters in Actinomycetales microorganisms. *The Journal of Antibiotics* **70,** 616–624. ISSN: 1881-1469. https://doi.org/10.1038/ja.2017.33 (2017).

39. Gallo, A., Ferrara, M. & Perrone, G. Phylogenetic study of polyketide synthases and nonribosomal peptide synthetases involved in the biosynthesis of mycotoxins. eng. *Toxins* **5.** 23604065[pmid], 717–742. ISSN: 2072-6651. https://pubmed.ncbi.nlm.nih.gov/23604065 (2013).

40. Keatinge-Clay, A. T. The structures of type I polyketide synthases. *Natural Product Reports* **29,** 1050. https://doi.org/10.1039/c2np20019h (2012).

41. Kasaragod, P., Schmitz, W., Hiltunen, J. K. & Wierenga, R. K. The isomerase and hydratase reaction mechanism of the crotonase active site of the multifunctional enzyme (type-1), as deduced from structures of complexes with 3S-hydroxy-acyl-CoA. *The FEBS Journal* **280,** 3160–3175.

42. Moretto, L., Heylen, R., Holroyd, N., Vance, S. & Broadhurst, R. W. Modular type I polyketide synthase acyl carrier protein domains share a common N-terminally extended fold. *Scientific Reports* **9,** 2325. ISSN: 2045-2322. https://doi.org/10.1038/s41598-019-38747-9 (2019).

43. Jenke-Kodama, H., Sandmann, A., Muller, R. & Dittmann, E. Evolutionary Implications of Bacterial Polyketide Synthases. *Molecular Biology and Evolution* **22,** 2027–2039. https://doi.org/10.1093/molbev/msi193 (June 2005).

44. Tran, L., Broadhurst, R. W., Tosin, M., Cavalli, A. & Weissman, K. J. Insights into Protein-Protein and Enzyme-Substrate Interactions in Modular Polyketide Synthases. *Chemistry & Biology* **17,** 705–716. https://doi.org/10.1016/j.chembiol.2010.05.017 (July 2010).

45. Liao, J.-L. *et al.* Chimeric 6-methylsalicylic acid synthase with domains of acyl carrier protein and methyltransferase from Pseudallescheria boydii shows novel biosynthetic activity. *Microbial Biotechnology* **12,** 920–931. https://doi.org/10.1111/1751-7915.13445 (June 2019).

46. Yadav, G., Gokhale, R. S. & Mohanty, D. Computational Approach for Prediction of Domain Organization and Substrate Specificity of Modular Polyketide Synthases. *Journal of Molecular Biology* **328,** 335 –363. ISSN: 0022-2836. http://www.sciencedirect.com/science/article/pii/S0022283603002328 (2003).

47. Lohman, J. R. *et al.* Structural and evolutionary relationships of “AT-less” type I polyketide synthase ketosynthases. *Proceedings of the National Academy of Sciences* **112,** 12693–12698. https://doi.org/10.1073/pnas.1515460112 (Sept. 2015).

48. Cummings, M., Breitling, R. & Takano, E. Steps towards the synthetic biology of polyketide biosynthesis. *FEMS microbiology letters* **351.** 24372666[pmid], 116–125. ISSN: 1574-6968. https://www.ncbi.nlm.nih.gov/pubmed/24372666 (2014).

49. Chan, Y. A. *et al.* Hydroxymalonyl-acyl carrier protein (ACP) and aminomalonyl-ACP are two additional type I polyketide synthase extender units. *Proceedings of the National Academy of Sciences* **103,** 14349–14354. ISSN: 0027-8424. eprint: https://www.pnas.org/content/103/39/14349.full.pdf. https://www.pnas.org/content/103/39/14349 (2006).

50. Khosla, C., Gokhale, R. S., Jacobsen, J. R. & Cane, D. E. Tolerance and Specificity of Polyketide Synthases. *Annual Review of Biochemistry* **68.** PMID: 10872449, 219–253. eprint: https://doi.org/10.1146/annurev.biochem.68.1.219. https://doi.org/10.1146/annurev.biochem.68.1.219 (1999).

51. Molloy, E. M., Tietz, J. I., Blair, P. M. & Mitchell, D. A. Biological characterization of the hygrobafilomycin antibiotic JBIR-100 and bioinformatic insights into the hygrolide family of natural products. *Bioorganic & medicinal chemistry* **24.** 27234886[pmid], 6276–6290. ISSN: 1464-3391. `https://www.ncbi.nlm.nih.gov/pubmed/27234886` (2016).

52. Chen, X.-H. *et al.* Structural and Functional Characterization of Three Polyketide Synthase Gene Clusters in Bacillus amyloliquefaciens FZB 42. *Journal of bacteriology* **188,** 4024–36 (July 2006).

53. Ishida, K., Lincke, T. & Hertweck, C. Assembly and Absolute Configuration of Short-Lived Polyketides fromBurkholderia thailandensis. *Angewandte Chemie International Edition* **51,** 5470–5474. `https://doi.org/10.1002/anie.201200067` (Apr. 2012).

54. Moldenhauer, J., Chen, X.-H., Borriss, R. & Piel, J. Biosynthesis of the Antibiotic Bacillaene, the Product of a Giant Polyketide Synthase Complex of the trans-AT Family. *Angewandte Chemie International Edition* **46,** 8195–8197. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200703386`. `https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200703386` (2007).

55. Subramanian, V., Dubini, A. & Seibert, M. in, 399–422 (Oct. 2012).

56. Kopp, M. *et al.* Insights into the complex biosynthesis of the leupyrrins in Sorangium cellulosum So ce690. *Mol. BioSyst.* **7,** 1549–1563. `http://dx.doi.org/10.1039/C0MB00240B` (5 2011).

57. Challis, G. L. & Naismith, J. H. Structural aspects of non-ribosomal peptide biosynthesis. *Current Opinion in Structural Biology* **14,** 748–756. `https://doi.org/10.1016/j.sbi.2004.10.005` (Dec. 2004).

58. Schoenafinger, G., Schracke, N., Linne, U. & Marahiel, M. A. Formylation Domain: An Essential Modifying Enzyme for the Nonribosomal Biosynthesis of Linear Gramicidin. *Journal of the American Chemical Society* **128,** 7406–7407. ISSN: 0002-7863. `https://doi.org/10.1021/ja0611240` (2006).

59. Scherlach, K. *et al.* Biosynthesis and Mass Spectrometric Imaging of Tolaasin, the Virulence Factor of Brown Blotch Mushroom Disease. *ChemBioChem* **14,** 2439–2443. `https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.201300553` (2013).

60. Sieber, S. A. & Marahiel, M. A. Molecular Mechanisms Underlying Nonribosomal Peptide Synthesis: Approaches to New Antibiotics. *Chemical Reviews* **105,** 715–738. ISSN: 0009-2665. `https://doi.org/10.1021/cr0301191` (2005).

61. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D. H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). eng. *Nucleic acids research* **33.** 16221976[pmid], 5799–5808. ISSN: 1362-4962. `https://pubmed.ncbi.nlm.nih.gov/16221976` (2005).

62. Walsh, C. T., O'Brien, R. V. & Khosla, C. Nonproteinogenic amino acid building blocks for nonribosomal peptide and hybrid polyketide scaffolds. eng. *Angewandte Chemie (International ed. in English)* **52.** 23729217[pmid], 7098–7124. ISSN: 1521-3773. `https://pubmed.ncbi.nlm.nih.gov/23729217` (2013).

63. Hubbard, B. K., Thomas, M. G. & Walsh, C. T. Biosynthesis of L-p-hydroxyphenylglycine, a non-proteinogenic amino acid constituent of peptide antibiotics**Supported in part by NIH grant GM 49338. *Chemistry Biology* **7,** 931 –942. ISSN: 1074-5521. `http://www.sciencedirect.com/science/article/pii/S1074552100000430` (2000).

64. Chen, H., Tseng, C. C., Hubbard, B. K. & Walsh, C. T. Glycopeptide antibiotic biosynthesis: enzymatic assembly of the dedicated amino acid monomer (S)-3,5-dihydroxyphenylglycine. eng. *Proceedings of the National Academy of Sciences of the United States of America* **98.** 11752437[pmid], 14901–14906. ISSN: 0027-8424. `https://pubmed.ncbi.nlm.nih.gov/11752437` (2001).

65. Cryle, M. J., Meinhart, A. & Schlichting, I. Structural characterization of OxyD, a cytochrome P450 involved in beta-hydroxytyrosine formation in vancomycin biosynthesis. eng. *The Journal of biological chemistry* **285.** 20519494[pmid], 24562–24574. ISSN: 1083-351X. `https://pubmed.ncbi.nlm.nih.gov/20519494` (2010).

66. Xu, B. *et al.* Insights into Pipecolic Acid Biosynthesis in Huperzia serrata. *Organic Letters* **20,** 2195–2198. ISSN: 1523-7060. `https://doi.org/10.1021/acs.orglett.8b00523` (2018).

67. Shen, B. *et al.* Cloning and Characterization of the Bleomycin Biosynthetic Gene Cluster from Streptomyces verticillus ATCC15003. *Journal of Natural Products* **65,** 422–431. ISSN: 0163-3864. `https://doi.org/10.1021/np010550q` (2002).

68. Mapp, A. K. & Heathcock, C. H. Total Synthesis of Myxalamide A. *The Journal of Organic Chemistry* **64,** 23–27. ISSN: 0022-3263. `https://doi.org/10.1021/jo9813742` (1999).

69. Moore, B. S. & Hertweck, C. Biosynthesis and attachment of novel bacterial polyketide synthase starter units. *Nat. Prod. Rep.* **19,** 70–99. `http://dx.doi.org/10.1039/B003939J` (1 2002).

70. Ray, L. & Moore, B. S. Recent advances in the biosynthesis of unusual polyketide synthase substrates. eng. *Natural product reports* **33.** 26571143[pmid], 150–161. ISSN: 1460-4752. `https://pubmed.ncbi.nlm.nih.gov/26571143` (2016).

71. Zhang, H. *et al.* Elucidation of the kijanimicin gene cluster: insights into the biosynthesis of spirotetronate antibiotics and nitrosugars. eng. *Journal of the American Chemical Society* **129.** 17985890[pmid], 14670–14683. ISSN: 1520-5126. `https://pubmed.ncbi.nlm.nih.gov/17985890` (2007).

72. Rebets, Y. *et al.* Complete genome sequence of producer of the glycopeptide antibiotic Aculeximycin Kutzneria albida DSM 43870T, a representative of minor genus of Pseudonocardiaceae. *BMC Genomics* **15,** 885. ISSN: 1471-2164. https://doi.org/10.1186/1471-2164-15-885 (2014).

73. Fang, J. *et al.* Cloning and Characterization of the Tetrocarcin A Gene Cluster from Micromonospora chalcea NRRL 11289 Reveals a Highly Conserved Strategy for Tetronate Biosynthesis in Spirotetronate Antibiotics. *Journal of Bacteriology* **190,** 6014–6025. ISSN: 0021-9193. eprint: https://jb.asm.org/content/190/17/6014.full.pdf. https://jb.asm.org/content/190/17/6014 (2008).

74. Kim, C.-G. *et al.* Biosynthesis of rubradirin as an ansamycin antibiotic from Streptomyces achromogenes var. rubradiris NRRL3061. *Archives of Microbiology* **189,** 463–473. https://doi.org/10.1007/s00203-007-0337-3 (Dec. 2007).

75. Lohman, J. R. *et al.* Cloning and sequencing of the kedarcidin biosynthetic gene cluster from Streptoalloteichus sp. ATCC 53650 revealing new insights into biosynthesis of the enediyne family of antitumor antibiotics. eng. *Molecular bioSystems* **9.** 23360970[pmid], 478–491. ISSN: 1742-2051. https://pubmed.ncbi.nlm.nih.gov/23360970 (2013).

76. Rui, Z. *et al.* Biochemical and genetic insights into asukamycin biosynthesis. eng. *The Journal of biological chemistry* **285.** 20522559[pmid], 24915–24924. ISSN: 1083-351X. https://pubmed.ncbi.nlm.nih.gov/20522559 (2010).

77. Kalan, L. *et al.* A Cryptic Polyene Biosynthetic Gene Cluster in Streptomyces calvus Is Expressed upon Complementation with a Functional bldA Gene. *Chemistry Biology* **20,** 1214 –1224. ISSN: 1074-5521. http://www.sciencedirect.com/science/article/pii/S1074552113003414 (2013).

78. Kang, H.-S. & Brady, S. F. Arimetamycin A: Improving Clinically Relevant Families of Natural Products through Sequence-Guided Screening of Soil Metagenomes. *Angewandte Chemie International Edition* **52,** 11063–11067. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201305109. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201305109 (2013).

79. Menéndez, N. *et al.* Biosynthesis of the Antitumor Chromomycin A¡sub¿3¡/sub¿ in ¡em¿Streptomyces griseus¡/em¿: Analysis of the Gene Cluster and Rational Design of Novel Chromomycin Analogs. *Chemistry & Biology* **11,** 21–32. ISSN: 1074-5521. https://doi.org/10.1016/j.chembiol.2003.12.011 (2004).

80. Grocholski, T. *et al.* Evolutionary Trajectories for the Functional Diversification of Anthracycline Methyltransferases. eng. *ACS chemical biology* **14.** 30995392[pmid], 850–856. ISSN: 1554-8937. https://pubmed.ncbi.nlm.nih.gov/30995392 (2019).

81. Zhao, M., Wijayasinghe, Y. S., Bhansali, P., Viola, R. E. & Blumenthal, R. M. A surprising range of modified-methionyl S-adenosylmethionine analogues support bacterial growth. eng. *Microbiology (Reading, England)* **161.** 25717169[pmid], 674–682. ISSN: 1465-2080. https://pubmed.ncbi.nlm.nih.gov/25717169 (2015).

82. Zhang, W., Bolla, M. L., Kahne, D. & Walsh, C. T. A three enzyme pathway for 2-amino-3-hydroxycyclopent-2-enone formation and incorporation in natural product biosynthesis. eng. *Journal of the American Chemical Society* **132.** 20394362[pmid], 6402–6411. ISSN: 1520-5126. https://pubmed.ncbi.nlm.nih.gov/20394362 (2010).

83. Petříčková, K. *et al.* Biosynthesis of Colabomycin E, a New Manumycin-Family Metabolite, Involves an Unusual Chain-Length Factor. *ChemBioChem* **15,** 1334–1345. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cbic.201400068. https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.201400068 (2014).

84. Gottardi, E. M. *et al.* Abyssomicin biosynthesis: formation of an unusual polyketide, antibiotic-feeding studies and genetic analysis. *Chembiochem : a European journal of chemical biology* **12.** 21656887[pmid], 1401–1410. ISSN: 1439-7633. https://www.ncbi.nlm.nih.gov/pubmed/21656887 (2011).

85. Zhang, C. *et al.* Biochemical and structural insights of the early glycosylation steps in calicheamicin biosynthesis. eng. *Chemistry & biology* **15.** 18721755[pmid], 842–853. ISSN: 1074-5521. https://pubmed.ncbi.nlm.nih.gov/18721755 (2008).

86. Jia, X.-Y. *et al.* Genetic Characterization of the Chlorothricin Gene Cluster as a Model for Spirotetronate Antibiotic Biosynthesis. *Chemistry Biology* **13,** 575 – 585. ISSN: 1074-5521. http://www.sciencedirect.com/science/article/pii/S1074552106001219 (2006).

87. Xiao, Y. *et al.* Characterization of Tiacumicin B Biosynthetic Gene Cluster Affording Diversified Tiacumicin Analogues and Revealing a Tailoring Dihalogenase. *Journal of the American Chemical Society* **133,** 1092–1105. https://doi.org/10.1021/ja109445q (Feb. 2011).

88. Weber, T. *et al.* CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *Journal of Biotechnology* **140.** Functional Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology, 13 –17. ISSN: 0168-1656. http://www.sciencedirect.com/science/article/pii/S0168165609000078 (2009).

89. Starcevic, A. *et al.* ClustScan : an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Research* **36,** 6882–6892. ISSN: 0305-1048. eprint: http://oup.prod.sis.lan/nar/article-pdf/36/21/6882/16748984/gkn685.pdf. https://doi.org/10.1093/nar/gkn685 (Oct. 2008).

90. Schuster-Böckler, B., Schultz, J. & Rahmann, S. HMM Logos for visualization of protein families. *BMC bioinformatics* **5.** 14736340[pmid], 7–7. ISSN: 1471-2105. https://www.ncbi.nlm.nih.gov/pubmed/14736340 (2004).

91. Murrell, B. *Structure discovery in hidden Markov models.* in (2009).

92. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191. https://doi.org/10.1093/bioinformatics/btp033 (Jan. 2009).

93. Starcevic, A. *et al.* ClustScan : an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Research* **36,** 6882–6892. https://doi.org/10.1093/nar/gkn685 (Oct. 2008).

94. Yadav, G., Gokhale, R. S. & Mohanty, D. Computational Approach for Prediction of Domain Organization and Substrate Specificity of Modular Polyketide Synthases. *Journal of Molecular Biology* **328,** 335–363. https://doi.org/10.1016/s0022-2836(03)00232-8 (Apr. 2003).

95. Minowa, Y., Araki, M. & Kanehisa, M. Comprehensive Analysis of Distinctive Polyketide and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes. *Journal of Molecular Biology* **368,** 1500–1517. https://doi.org/10.1016/j.jmb.2007.02.099 (May 2007).

96. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry & Biology* **6,** 493–505. https://doi.org/10.1016/s1074-5521(99)80082-9 (Aug. 1999).

97. Keatinge-Clay, A. T. *et al.* Catalysis, Specificity, and ACP Docking Site of Streptomyces coelicolor Malonyl-CoA:ACP Transacylase. *Structure* **11,** 147–154. https://doi.org/10.1016/s0969-2126(03)00004-2 (Feb. 2003).

98. Du, L., Sánchez, C., Chen, M., Edwards, D. J. & Shen, B. The biosynthetic gene cluster for the antitumor drug bleomycin from Streptomyces verticillus ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. *Chemistry Biology* **7,** 623 –642. ISSN: 1074-5521. http://www.sciencedirect.com/science/article/pii/S1074552100000119 (2000).

99. Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biology* **20,** 121. ISSN: 1474-760X. https://doi.org/10.1186/s13059-019-1730-3 (2019).

100. Mendoza, S. N., Olivier, B. G., Molenaar, D. & Teusink, B. A systematic assessment of current genome-scale metabolic reconstruction tools. eng. *Genome biology* **20.** 31391098[pmid], 158–158. ISSN: 1474-760X. https://pubmed.ncbi.nlm.nih.gov/31391098 (2019).

101. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? eng. *Nature biotechnology* **28.** 20212490[pmid], 245–248. ISSN: 1546-1696. https://pubmed.ncbi.nlm.nih.gov/20212490 (2010).

102. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology* **10,** 291–305. ISSN: 1740-1534. https://doi.org/10.1038/nrmicro2737 (2012).

103. Yu, T.-W. *et al.* The biosynthetic gene cluster of the maytansinoid antitumor agent ansamitocin from Actinosynnema pretiosum. *Proceedings of the National Academy of Sciences* **99,** 7968–7973. ISSN: 0027-8424. eprint: https://www.pnas.org/content/99/12/7968.full.pdf. https://www.pnas.org/content/99/12/7968 (2002).

104. Du, Y. *et al.* Biosynthesis of the apoptolidins in Nocardiopsis sp. FU 40. *Tetrahedron* **67.** 2010 Tetrahedron Prize for Creativity in Organic Chemistry, Treasure from Microorganism: Discovery, Chemicalbiology and Total Synthesis, Satoshi Omura, 6568 –6575. ISSN: 0040-4020. http://www.sciencedirect.com/science/article/pii/S0040402011007964 (2011).

105. Haydock, S. F. *et al.* Organization of the biosynthetic gene cluster for the macrolide concanamycin A in Streptomyces neyagawaensis ATCC 27449. *Microbiology* **151,** 3161–3169. ISSN: 1350-0872. https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.28194-0 (2005).

106. Mandala, S. M. *et al.* Rustmicin, a Potent Antifungal Agent, Inhibits Sphingolipid Synthesis at Inositol Phosphoceramide Synthase. *Journal of Biological Chemistry* **273,** 14942–14949. https://doi.org/10.1074/jbc.273.24.14942 (June 1998).

107. Patel, K. *et al.* Engineered Biosynthesis of Geldanamycin Analogs for Hsp90 Inhibition. *Chemistry Biology* **11,** 1625 –1633. ISSN: 1074-5521. http://www.sciencedirect.com/science/article/pii/S1074552104002960 (2004).

108. in. *Alkaloids* (eds Funayama, S. & Cordell, G. A.) 219 –232 (Academic Press, Boston, 2015). ISBN: 978-0-12-417302-6. http://www.sciencedirect.com/science/article/pii/B9780124173026000131.

109. Takaishi, M., Kudo, F. & Eguchi, T. Biosynthetic pathway of 24-membered macrolactam glycoside incednine. *Tetrahedron* **64,** 6651 –6656. ISSN: 0040-4020. http://www.sciencedirect.com/science/article/pii/S0040402008009125 (2008).

110. Hatano, K., Muroi, M., Higashide, E. & Yoneda, M. Biosynthesis of Macbecin. *Agricultural and Biological Chemistry* **46,** 1699–1702. eprint: https://doi.org/10.1080/00021369.1982.10865313. https://doi.org/10.1080/00021369.1982.10865313 (1982).

111. Cong, L. & Piepersberg, W. Cloning and Characterization of Genes Encoded in dTDP-D-mycaminose Biosynthetic Pathway from a Midecamycin-producing Strain, Streptomyces mycarofaciens. *Acta Biochimica et Biophysica Sinica* **39,** 187–193. ISSN: 1672-9145. eprint: https://academic.oup.com/abbs/article-pdf/39/3/187/906/39-3-187.pdf. https://doi.org/10.1111/j.1745-7270.2007.00265.x (Mar. 2007).

112. Li, W., Ju, J., Rajski, S. R., Osada, H. & Shen, B. Characterization of the Tautomycin Biosynthetic Gene Cluster fromStreptomyces spiroverticillatusUnveiling New Insights into Dialkylmaleic Anhydride and Polyketide Biosynthesis. *Journal of Biological Chemistry* **283,** 28607–28617. https://doi.org/10.1074/jbc.m804279200 (Aug. 2008).

113. Rachid, S., Scharfe, M., Blöcker, H., Weissman, K. J. & Müller, R. Unusual Chemistry in the Biosynthesis of the Antibiotic Chondrochlorens. *Chemistry Biology* **16,** 70 –81. ISSN: 1074-5521. http://www.sciencedirect.com/science/article/pii/S1074552108004511 (2009).

114. Jahns, C. *et al.* Pellasoren: Structure Elucidation, Biosynthesis, and Total Synthesis of a Cytotoxic Secondary Metabolite fromSorangium cellulosum. *Angewandte Chemie International Edition* **51,** 5239–5243. https://doi.org/10.1002/anie.201200327 (Apr. 2012).

115. Masschelein, J., Jenner, M. & Challis, G. Antibiotics from Gram-negative bacteria: A comprehensive overview and selected biosynthetic highlights. *Natural Product Reports* **34** (June 2017).

116. Kevany, B. M., Rasko, D. A. & Thomas, M. G. Characterization of the Complete Zwittermicin A Biosynthesis Gene Cluster from Bacillus cereus. *Applied and Environmental Microbiology* **75,** 1144–1155. ISSN: 0099-2240. eprint: https://aem.asm.org/content/75/4/1144.full.pdf. https://aem.asm.org/content/75/4/1144 (2009).

117. Helfrich, E. J. N. & Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural Product Reports* **33,** 231–316. https://doi.org/10.1039/c5np00125k (2016).

118. Lin, J., Bai, L., Deng, Z. & Zhong, J.-J. Effect of Ammonium in Medium on Ansamitocin P-3 Production by Actinosynnema pretiosum. *Biotechnology and Bioprocess Engineering* **15,** 119–125 (Feb. 2010).

119. Harunari, E., Komaki, H. & Igarashi, Y. Biosynthetic origin of butyrolactol A, an antifungal polyketide produced by a marine-derived Streptomyces. eng. *Beilstein journal of organic chemistry* **13.** 28382182[pmid], 441–450. ISSN: 1860-5397. https://pubmed.ncbi.nlm.nih.gov/28382182 (2017).

120. Wenzel, S. C. *et al.* Production of the Bengamide Class of Marine Natural Products in Myxobacteria: Biosynthesis and Structure–Activity Relationships. *Angewandte Chemie International Edition* **54,** 15560–15564. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201508277 (2015).

121. Romo, A. J. *et al.* The Amipurimycin and Miharamycin Biosynthetic Gene Clusters: Unraveling the Origins of 2-Aminopurinyl Peptidyl Nucleoside Antibiotics. *Journal of the American Chemical Society* **141,** 14152–14159. ISSN: 0002-7863. https://doi.org/10.1021/jacs.9b03021 (2019).

122. Li, X., Wu, X. & Shen, Y. Identification of the Bacterial Maytansinoid Gene Cluster asc Provides Insights into the Post-PKS Modifications of Ansacarbamitocin Biosynthesis. *Organic Letters* **21.** PMID: 31299158, 5823–5826. eprint: https://doi.org/10.1021/acs.orglett.9b01891. https://doi.org/10.1021/acs.orglett.9b01891 (2019).

123. Karray, F. *et al.* Organization of the biosynthetic gene cluster for the macrolide antibiotic spiramycin in Streptomyces ambofaciens. *Microbiology* **153,** 4111–4122. ISSN: 1350-0872. https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.2007/009746-0 (2007).

124. He, H.-Y. *et al.* Quartromicin Biosynthesis: Two Alternative Polyketide Chains Produced by One Polyketide Synthase Assembly Line. *Chemistry Biology* **19,** 1313 – 1323. ISSN: 1074-5521. http://www.sciencedirect.com/science/article/pii/S1074552112003250 (2012).

125. Demydchuk, Y. *et al.* Analysis of the Tetronomycin Gene Cluster: Insights into the Biosynthesis of a Polyether Tetronate Antibiotic. *ChemBioChem* **9,** 1136–1145. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cbic.200700715. https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.200700715 (2008).

126. Zhang, C., Ding, W., Qin, X. & Ju, J. Genome Sequencing of Streptomyces olivaceus SCSIO T05 and Activated Production of Lobophorin CR4 via Metabolic Engineering and Genome Mining. eng. *Marine drugs* **17.** 31635159[pmid], 593. ISSN: 1660-3397. https://pubmed.ncbi.nlm.nih.gov/31635159 (2019).

127. Hashimoto, T. *et al.* Biosynthesis of Versipelostatin: Identification of an Enzyme-Catalyzed [4+2]-Cycloaddition Required for Macrocyclization of Spirotetronate-Containing Polyketides. *Journal of the American Chemical Society* **137,** 572–575. ISSN: 0002-7863. https://doi.org/10.1021/ja510711x (2015).

128. Daduang, R. *et al.* Characterization of the biosynthetic gene cluster for maklamicin, a spirotetronate-class antibiotic of the endophytic Micromonospora sp. NBRC 110955. *Microbiological Research* **180,** 30–39. https://doi.org/10.1016/j.micres.2015.07.003 (Nov. 2015).

129. Buchholz, T. J. *et al.* Polyketide β-Branching in Bryostatin Biosynthesis: Identification of Surrogate Acetyl-ACP Donors for BryR, an HMG-ACP Synthase. *Chemistry & Biology* **17,** 1092–1100. https://doi.org/10.1016/j.chembiol.2010.08.008 (Oct. 2010).

130. Elshahawi, S. I. *et al.* Boronated tartrolon antibiotic produced by symbiotic cellulose-degrading bacteria in shipworm gills. *Proceedings of the National Academy of Sciences* **110,** E295–E304. ISSN: 0027-8424. eprint: https://www.pnas.org/content/110/4/E295.full.pdf. https://www.pnas.org/content/110/4/E295 (2013).

131. Zhang, F. *et al.* Cloning and Elucidation of the FR901464 Gene Cluster Revealing a Complex Acyltransferase-less Polyketide Synthase Using Glycerate as Starter Units. *Journal of the American Chemical Society* **133,** 2452–2462. ISSN: 0002-7863. https://doi.org/10.1021/ja105649g (2011).

132. Eustáquio, A. S., Janso, J. E., Ratnayake, A. S., O'Donnell, C. J. & Koehn, F. E. Spliceostatin hemiketal biosynthesis in Burkholderia spp. is catalyzed by an iron/-ketoglutarate–dependent dioxygenase. *Proceedings of the National Academy of Sciences* **111,** E3376–E3385. ISSN: 0027-8424. eprint: https://www.pnas.org/content/111/33/E3376.full.pdf. https://www.pnas.org/content/111/33/E3376 (2014).

133. Liu, X. *et al.* Genomics-guided discovery of thailanstatins A, B, and C As pre-mRNA splicing inhibitors and antiproliferative agents from Burkholderia thailandensis MSMB43. eng. *Journal of natural products* **76.** 23517093[pmid], 685–693. ISSN: 1520-6025. https://pubmed.ncbi.nlm.nih.gov/23517093 (2013).

134. Bertin, M. J. *et al.* The Phormidolide Biosynthetic Gene Cluster: A trans-AT PKS Pathway Encoding a Toxic Macrocyclic Polyketide. eng. *Chembiochem : a European journal of chemical biology* **17.** 26769357[pmid], 164–173. ISSN: 1439-7633. https://pubmed.ncbi.nlm.nih.gov/26769357 (2016).

135. Gil, J. & Campelo-Diez, A. Candicidin biosynthesis in Streptomyces griseus. *Applied Microbiology and Biotechnology* **60,** 633–642. ISSN: 1432-0614. https://doi.org/10.1007/s00253-002-1163-9 (2003).

136. Ward, S. L. *et al.* Chalcomycin Biosynthesis Gene Cluster from Streptomyces bikiniensis: Novel Features of an Unusual Ketolide Produced through Expression of the chm Polyketide Synthase in Streptomyces fradiae. *Antimicrobial Agents and Chemotherapy* **48,** 4703–4712. ISSN: 0066-4804. eprint: https://aac.asm.org/content/48/12/4703.full.pdf. https://aac.asm.org/content/48/12/4703 (2004).

137. Amagai, K., Takaku, R., Kudo, F. & Eguchi, T. A Unique Amino Transfer Mechanism for Constructing the -Amino Fatty Acid Starter Unit in the Biosynthesis of the Macrolactam Antibiotic Cremimycin. *ChemBioChem* **14,** 1998–2006. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cbic.201300370. https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.201300370 (2013).

138. Zhang, W., Bolla, M. L., Kahne, D. & Walsh, C. T. A three enzyme pathway for 2-amino-3-hydroxycyclopent-2-enone formation and incorporation in natural product biosynthesis. eng. *Journal of the American Chemical Society* **132.** 20394362[pmid], 6402–6411. ISSN: 1520-5126. https://pubmed.ncbi.nlm.nih.gov/20394362 (2010).

139. Zhang, H., Wang, Y., Wu, J., Skalina, K. & Pfeifer, B. A. Complete Biosynthesis of Erythromycin A and Designed Analogs Using E. coli as a Heterologous Host. *Chemistry Biology* **17,** 1232 –1240. ISSN: 1074-5521. `http://www.sciencedirect.com/science/article/pii/S1074552110003984` (2010).

140. Arakawa, K., Kodama, K., Tatsuno, S., Ide, S. & Kinashi, H. Analysis of the loading and hydroxylation steps in lankamycin biosynthesis in Streptomyces rochei. eng. *Antimicrobial agents and chemotherapy* **50.** 16723550[pmid], 1946–1952. ISSN: 0066-4804. `https://pubmed.ncbi.nlm.nih.gov/16723550` (2006).

141. Volchegursky, Y., Hu, Z., Katz, L. & McDaniel, R. Biosynthesis of the anti-parasitic agent megalomicin: transformation of erythromycin to megalomicin in Saccharopolyspora erythraea. *Molecular Microbiology* **37,** 752–762. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2958.2000.02059.x`. `https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.2000.02059.x` (2000).

142. Anzai, Y. *et al.* Function of Cytochrome P450 Enzymes MycCI and MycG in Micromonospora griseorubida, a Producer of the Macrolide Antibiotic Mycinamicin. *Antimicrobial agents and chemotherapy* **56,** 3648–56 (Apr. 2012).

143. Liu, T., Lin, X., Zhou, X., Deng, Z. & Cane, D. E. Mechanism of thioesterase-catalyzed chain release in the biosynthesis of the polyether antibiotic nanchangmycin. eng. *Chemistry & biology* **15.** 18482697[pmid], 449–458. ISSN: 1074-5521. `https://pubmed.ncbi.nlm.nih.gov/18482697` (2008).

144. Aparicio, J. F. *et al.* Biotechnological production and application of the antibiotic pimaricin: biosynthesis and its regulation. eng. *Applied microbiology and biotechnology* **100.** 26512010[pmid], 61–78. ISSN: 1432-0614. `https://pubmed.ncbi.nlm.nih.gov/26512010` (2016).

145. Bruheim, P. *et al.* Chemical Diversity of Polyene Macrolides Produced by Streptomyces noursei ATCC 11455 and Recombinant Strain ERD44 with Genetically Altered Polyketide Synthase NysC. *Antimicrobial agents and chemotherapy* **48,** 4120–9 (Dec. 2004).

146. Evans, D. A. & Black, W. C. Total synthesis of (+)-A83543A [(+)-lepicidin A]. *Journal of the American Chemical Society* **115,** 4497–4513. ISSN: 0002-7863. `https://doi.org/10.1021/ja00064a011` (1993).

147. Song, L. *et al.* Cytochrome P450-mediated hydroxylation is required for polyketide macrolactonization in stambomycin biosynthesis. *The Journal of Antibiotics* **67,** 71–76. ISSN: 1881-1469. `https://doi.org/10.1038/ja.2013.119` (2014).

148. Xiao, Y. *et al.* Characterization of Tiacumicin B Biosynthetic Gene Cluster Affording Diversified Tiacumicin Analogues and Revealing a Tailoring Dihalogenase. *Journal of the American Chemical Society* **133,** 1092–1105. ISSN: 0002-7863. `https://doi.org/10.1021/ja109445q` (2011).

149. Ogasawara, Y. *et al.* Cloning, Sequencing, and Functional Analysis of the Biosynthetic Gene Cluster of Macrolactam Antibiotic Vicenistatin in ¡em¿Streptomyces halstedii¡/em¿. *Chemistry & Biology* **11,** 79–86. ISSN: 1074-5521. https://doi.org/10.1016/j.chembiol.2003.12.010 (2004).

150. Luzhetskyy, A. *et al.* Cloning and Heterologous Expression of the Aranciamycin Biosynthetic Gene Cluster Revealed a New Flexible Glycosyltransferase. *ChemBioChem* **8,** 599–602. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cbic.200600529. https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.200600529 (2007).

151. Jensen, P. R., Moore, B. S. & Fenical, W. The marine actinomycete genus Salinispora: a model organism for secondary metabolite discovery. eng. *Natural product reports* **32.** 25730728[pmid], 738–751. ISSN: 1460-4752. https://pubmed.ncbi.nlm.nih.gov/25730728 (2015).

152. Kang, H.-S. & Brady, S. F. Arixanthomycins A-C: Phylogeny-guided discovery of biologically active eDNA-derived pentangular polyphenols. eng. *ACS chemical biology* **9.** 24730509[pmid], 1267–1272. ISSN: 1554-8937. https://pubmed.ncbi.nlm.nih.gov/24730509 (2014).

153. Sasaki, E., Ogasawara, Y. & Liu, H.-w. A Biosynthetic Pathway for BE-7585A, a 2-Thiosugar-Containing Angucycline-Type Natural Product. *Journal of the American Chemical Society* **132,** 7405–7417. ISSN: 0002-7863. https://doi.org/10.1021/ja1014037 (2010).

154. Lukežič, T. *et al.* Identification of the chelocardin biosynthetic gene cluster from Amycolatopsis sulphurea: a platform for producing novel tetracycline antibiotics. *Microbiology* **159,** 2524–2532. ISSN: 1350-0872. https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.070995-0 (2013).

155. Malmierca, M. G. *et al.* New Sipanmycin Analogues Generated by Combinatorial Biosynthesis and Mutasynthesis Approaches Relying on the Substrate Flexibility of Key Enzymes in the Biosynthetic Pathway. *Applied and Environmental Microbiology* **86** (ed Drake, H. L.) ISSN: 0099-2240. eprint: https://aem.asm.org/content/86/3/e02453-19.full.pdf. https://aem.asm.org/content/86/3/e02453-19 (2020).

156. Yeo, W. L. *et al.* Biosynthetic engineering of the antifungal, anti-MRSA auroramycin. *Microbial Cell Factories* **19,** 3. ISSN: 1475-2859. https://doi.org/10.1186/s12934-019-1274-y (2020).

157. Robbins, N. *et al.* Discovery of Ibomycin, a Complex Macrolactone that Exerts Antifungal Activity by Impeding Endocytic Trafficking and Membrane Function. *Cell Chemical Biology* **23,** 1383–1394. ISSN: 2451-9456. https://doi.org/10.1016/j.chembiol.2016.08.015 (2016).

158. Burgess, K. M. N., Renaud, J. B., McDowell, T. & Sumarah, M. W. Mechanistic Insight into the Biosynthesis and Detoxification of Fumonisin Mycotoxins. *ACS Chemical Biology* **11,** 2618–2625. ISSN: 1554-8929. https://doi.org/10.1021/acschembio.6b00438 (2016).

159. Salem, S. M. *et al.* Elucidation of final steps of the marineosins biosynthetic pathway through identification and characterization of the corresponding gene cluster. eng. *Journal of the American Chemical Society* **136.** 24575817[pmid], 4565–4574. ISSN: 1520-5126. https://pubmed.ncbi.nlm.nih.gov/24575817 (2014).

160. Williamson, N. R. *et al.* Biosynthesis of the red antibiotic, prodigiosin, in Serratia: identification of a novel 2-methyl-3-n-amyl-pyrrole (MAP) assembly pathway, definition of the terminal condensing enzyme, and implications for undecylprodigiosin biosynthesis in Streptomyces. *Molecular Microbiology* **56,** 971–989. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2958.2005.04602.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2958.2005.04602.x (2005).

161. Zhou, Y. *et al.* Iterative Mechanism of Macrodiolide Formation in the Anticancer Compound Conglobatin. *Chemistry Biology* **22,** 745 –754. ISSN: 1074-5521. http://www.sciencedirect.com/science/article/pii/S1074552115001945 (2015).

162. Burgard, C. *et al.* Genomics-Guided Exploitation of Lipopeptide Diversity in Myxobacteria. *ACS Chemical Biology* **12,** 779–786. ISSN: 1554-8929. https://doi.org/10.1021/acschembio.6b00953 (2017).

163. Kawata, J., Naoe, T., Ogasawara, Y. & Dairi, T. Biosynthesis of the Carbonylmethylene Structure Found in the Ketomemicin Class of Pseudotripeptides. *Angewandte Chemie International Edition* **56,** 2026–2029. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201611005. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201611005 (2017).

164. Suroto, D. A., Kitani, S., Arai, M., Ikeda, H. & Nihira, T. Characterization of the biosynthetic gene cluster for cryptic phthoxazolin A in Streptomyces avermitilis. *PLOS ONE* **13,** 1–17. https://doi.org/10.1371/journal.pone.0190973 (Jan. 2018).

165. Sulheim, S. *et al.* Genome-scale Model Constrained by Proteomics Reveals Metabolic Changes in Streptomyces coelicolor M1152 Compared to M145. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2020/03/26/796722.full.pdf. https://www.biorxiv.org/content/early/2020/03/26/796722 (2020).

Appendix

The various starter units that have been observed for loader modules

Table (6.1) Starter units encountered in various secondary metabolites. (CHC: Cyclohexanecarboxylate, PABA: P-aminobenzoate, 3,5-AHBA: 3-amino-5-hydroxybenzoate, DHCH: 3,4,-dihydrocyclohexanecarboxylate, 3,4-AHBA: 3-amino-4-hydroxybenzoate, DHBA: 3,5-Dihydroxybenzoate) 1-18 from [69], 19-25 from [70], 26 from [53]

index	Starter unit incorporated*	Secondary metabolite(s)
1	Isobutyryl-coa	Virginiamycin, Avermectins, Manumycins
2	(S)-2-methylbutyryl-coa	Avermectins, Manumycins
3	Isovaleryl-coa	Manumycins, Myxothiazol
4	Various three carbon units	Aplasmomycin, Boromycin, Tartrolon B
5	CHC	Asukamycin, Ansatrienin
6	DHCHC	Rapamycin, Ascomycin, Tacrolimus
7	Benzoate	Soraphen A
8	PABA	Candicidin, FR-008
9	3,5-AHBA	Ansamycins
10	3,4-AHBA	Manumycins
11	DHBA	Kendomycin
12	Phenylacetate	Ripostatin
13	Phenylacetate	Microcystin-LR, Nodularin
14	Beta alanine	Fluvirucins
15	4-aminobutyrate	Marginolactones
16	3-amino-2-methylpropionate	Vicenistatin
17	Beta phenylalanine	Hitachimycin
18	Proline-derived	Calcimycin, Indanomycin
19	p-nitrobenzoic acid	Aureothin
20	Pyrrolyl-CP	Pyileuteorin, Marinopyrrole A
21	4,5-dichloropyrrolyl-CP	Chlorizidine A
22	4-guanidinobutyril-CoA	Azalomycin F3a
23	5-Hexynoyl-Coa	Jamaicamide
24	Benzylmalonyl-CoA	Splenocins
25	3-oxoadipyl-CoA	Pamamycins
26	4-hydroxyphenylacetate	Thailandamide

Table (6.2) The main substrates for synthesis of the different starter units observed for T1PKS. Cofactors and CoA have been ignored. PEP: Phosphoenoylpyruvate, E4P: erythrose-4-phosphate. * *via* citric acid cycle, through L-aspartate. ***via* citric acid cycle, through L-glutamate. ***Starter unit is either synthesized from glycerol and directly incorporated, or incorporated as phenylalanine and subsequently modified. ****Is synthesized by a fatty-acyl-CoA ligase, a membrane-bound fatty acid desaturase and an ACP, all found within the jamaicamide BGC. 1-18 from [69], 19-25 from [70]

Index	Starter unit incorporated	Derived from
1	Isobutyryl-coa	L-Valine
2	(S)-2-Methylbutyryl-CoA	L-Isoleucine
3	Isovaleryl-coa	L-Leucine
4	Various three carbon (3C) units	Glycerol
5	CHC	Shikaiic acid
6	DHCHC	Shikimic acid
7	Benzoate*	Glycerol
8	PABA	Chorismic acid
9	3,5-AHBA	Shikimic acid
10	3,4-AHBA	Succinate + Glycerol + Pyruvate
11	DHBA	4 x Malonyl-coa
12	Phenylacetate	Phenylalanine + something else
13	Phenylacetate	Phenylalanine + Pyruvate
14	β-Alanine*	Acetyl-coa
15	4-Aminobutyrate	Ornithine or Arginine
16	3-Amino-2-methylpropionate**	Acetyl-coa
17	β-Phenylalanine***	Glycerol
18	Proline-derived	Proline
19	p-Nitrobenzoic acid	Chorismic acid
20	Pyrrolyl-CP and	Proline
21	4,5-Dichloropyrrolyl-CP	Proline
22	4-Guanidinobutyril-CoA	L-arginine
23	5-Hexynoyl-CoA****	Acetyl-CoA
24	Benzylmalonyl-CoA	L-phenylalanine
25	3-Oxoadipyl-CoA	Succinate
26	4-Hydroxyphenylacetate	Unknown

Table (6.3) * Are the enzymes responsible for synthesising the starter unit encoded by the BGC?, ** Is the starter unit found within *S. coelicolor*?, *** Are the precursor metabolites found in *S. coelicolor*?

index	Starter unit incorporated	*	**	***
1	Isobutyryl-coa	yes	yes	yes
2	(S)-2-methylbutyryl-coa	yes	yes	yes
3	Isovaleryl-coa	yes	yes	yes
4	Various three carbon units	yes	no	yes
5	CHC	yes	no	yes
6	DHCHC	no	no	yes
7	Benzoate	yes	no	yes
8	PABA	yes	no	yes
9	3,5-AHBA	yes	no	yes
10	3,4-AHBA	yes	no	yes
11	DHBA	yes	no	yes
12	Phenylacetate	yes	yes	yes
13	Phenylacetate	yes	yes	yes
14	Beta alanine	no	yes	yes
15	4-aminobutyrate	yes	no	yes
16	3-amino-2-methylpropionate	yes	no	yes
17	Beta phenylalanine	no	yes	yes
18	Proline-derived	no	yes	yes
19	p-nitrobenzoic acid	yes	yes	yes
20	Pyrrolyl-CP and	yes	no	yes
21	4,5-dichloropyrrolyl-CP	yes	no	yes
22	4-guanidinobutyril-CoA	yes	no	yes
23	5-Hexynoyl-Coa	yes	no	yes
24	Benzylmalonyl-CoA	yes	no	yes
25	3-oxoadipyl-CoA	yes	yes	yes
26	4-hydroxyphenylacetate	no	no	unknown

Predicted substrate specificities for all modules in all BGCs in the MIBiG database

Table (6.4) The substrate specificities of all modules in the MIBiG database, as predicted by antiSMASH. PA = Proteinogenic amino acid, NPA = non-proteinogenic amino acid, AU = Acyl unit, LM = Loader Module, # = number of times a module with a given specificity was detected for all modules found in the MIBiG database. A note on the column showing the specificities of alternative loader modules: specificities other than FkbH and GNAT result from CAL domains. Mal = Malonyl-CoA, Mmal = Methylmalonyl-CoA, Emal = ethylmalonyl-CoA, Mxmal = methoxymalonyl-ACP

NRPS				PKS		NRPS/PKS	
PA	#	NPA	#	AU	#	LM	#
Unknown	701	HPG	60	Mal	2146	NH_2	46
Thr	229	DAB	48	Mmal	735	Fatty acid	25
Val	188	Orn	31	Custom starter	581	AHBA	24
Ser	175	BHT	26	Unknown	278	FkbH	16
Leu	154	DHPG	21	Emal	28	GNAT	12
Gly	145	PIP	11	Mxmal	12	Acetyl-CoA	2
Ala	119	Ala-b	7			Shikimic acid	1
Pro	106	Iva	6				
Cys	90	HAORN	5				
Phe	90	AAD	4				
Asn	74	β-Ala	3				
Ile	72	DMT-TRP	3				
Asp	71	Abu	2				
Tyr	67	3-Me-Glu	2				
Gln	41	HASN	2				
Glu	26	HYV-d	2				
Trp	7	Alaninol	1				
Arg	6	Cap	1				
Lys	5	LDAP	1				
His	1						
Ser-Thr	1						

Comparisons between computationally constructed and experimentally determined pathways

For all tables in this section: Each complete module is separated with horisontal lines. All ACP, A and AT-domains are excluded from the sequences. KS and C domains are indicated as catalysing the extensions of the core structure even though the A and AT-domains are responsible for recruiting the extender unit. ”Apparent” represents the true order of domains, regardless of activity. ”Effective” represents the reaction that has been experimentally determined to occur. ”Predicted” represents the reactions in the constructed pathway

Table (6.5) Comparison between the computationally predicted and experimentally determined pathway of difficidin (BGC0000176).

Apparent	Effective	Predicted	Effective	Predicted
KS	KS	KS	1,3-biphosphoglycerate → 2 P_i	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
MT	MT	MT	SAM → SAH	SAM → SAH
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
	ER (free-standing)		NADPH + H^+ → NADP+	
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH				
	KR (free-standing)		NADPH + H^+ → NADP+	
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
	DH (free-standing)		→ H_2O	
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS			DHD	DHD
DH			DHD	DHD
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
MT	MT	MT	SAM → SAH	SAM → SAH
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH		DH		→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
	ER (free-standing)		NADPH + H^+ → NADP+	
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
	DH (free-standing)		→ H_2O	
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS		Malonyl-CoA → CO_2 + CoA	DHD
DH	DH		→ H_2O	DHD
KR	KR		NADPH + H^+ → NADP+	DHD
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
		DH		→ H_2O
KS			DHD	DHD
DH			DHD	DHD
DH				
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
MT	MT	MT	SAM → SAH	SAM → SAH
ER	ER	ER	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
TE	TE	TE	H_2O →	

Table (6.6) Alignment of experimentally determined VS predicted predicted domain activities of oocydin

Actual	Effective	Predicted	Effective	Predicted
DH				H_2O
FkbH	FkbH	FKBH	1,3-biphosphoglycerate → 2 pi	1,3-biphosphoglycerate → 2 pi
FkbM	FkbM		SAM → SAH	
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
ECH	ECH	ECH		
ECH	ECH	ECH		
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	
	KR (free-standing)		NADPH + H^+ → NADP+	
		KS		Malonyl-CoA → CO_2 + CoA
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS_0		KS		Malonyl-CoA → CO_2 + CoA
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
PS	PS	PS		
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
KS_0		KS		Malonyl-CoA → CO_2 + CoA
KS_0		KS		Malonyl-CoA → CO_2 + CoA
DH	DH		→ H_2O	
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
MT	MT	MT	SAM → SAH	SAM → SAH
KS_0		KS		Malonyl-CoA → CO_2 + CoA
DH		DH		
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS_0		KS		Malonyl-CoA → CO_2 + CoA
TE	TE	TE	H_2O →	H_2O →
KS		KS		Malonyl-CoA → CO_2 + CoA
C				

Table (6.7) Alignment of experimentally determined VS predicted predicted domain activities of leupyrrin

real	predicted	real	predicted
C	AT	Prolyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
Cy	Cy	Threonine → H_2O	Cysteine → H_2O
KS	KS	2c3h5m-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	→ H_2O	→ H_2O
KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	→ H_2O	→ H_2O
CMT	cMT	SAM → SAH	SAM → SAH
KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	→ H_2O	→ H_2O
KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	→ H_2O	→ H_2O
CMT	cMT	SAM → SAH	SAM → SAH
KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
C	Condensation	Proline → H_2O	Proline → H_2O
TE	Thioesterase	H_2O →	H_2O →

Table (6.8) Alignment of experimentally determined VS predicted predicted domain activities of tolaasin

Real	Predicted	Real	Predicted
C	C	Threonine → H_2O	Threonine → H_2O
C	C	Proline → H_2O	Proline → H_2O
C	C	Serine → H_2O	Serine → H_2O
C	C	Leucine → H_2O	Leucine → H_2O
C	C	Valine → H_2O	Valine → H_2O
C	C	Serine → H_2O	Serine → H_2O
C	C	Leucine → H_2O	Leucine → H_2O
C	C	Valine → H_2O	Valine → H_2O
C	C	Valine → H_2O	Valine → H_2O
C	C	Glutamine → H_2O	Glutamine → H_2O
C	C	Leucine → H_2O	Leucine → H_2O
C	C	Valine → H_2O	Valine → H_2O
C	C	Threonine → H_2O	Threonine → H_2O
C	C	Threonine → H_2O	Threonine → H_2O
C	C	Leucine → H_2O	Leucine → H_2O
C	C	L-homoserine → H_2O	Unknown → H_2O
C	C	L-2,4-diaminobutanoate → H_2O	L-2,4-diaminobutanoate → H_2O
C	C	Lysine → H_2O	Unknown → H_2O

Table (6.9) Alignment of experimentally determined VS predicted predicted domain activities of anabaenopeptin

Apparent	Real	Predicted	Real	Predicted
C		C		Unknown → H_2O
C		C		Unknown → H_2O
C	C	C	Tyrosine → H_2O	Tyrosine → H_2O
C	C	C	Lysine → H_2O	Unknown → H_2O
C	C	C	Valine → H_2O	Valine → H_2O
C	C	C	Tyrosine → H_2O	Unknown → H_2O
C MT	C MT	C MT	Alanine → H_2O SAM → SAH	Alanine → H_2O SAM → SAH
C	C	C	Phenylalanine → H_2O	Phenylalanine → H_2O

Table (6.10) Alignment of experimentally determined VS predicted domain activities of geldanamycin

Apparent	Real	Predicted	Real	Predicted
CAL	CAL	CAL	AHBA	AHBA
KR	KR	KR		
KS	KS	KS	Methylmalonyl-CoA → CO_2 + CoA	Methylmalonyl-CoA → CO_2 + CoA
AT	AT	AT		
DH	DH	DH	→ H_2O	→ H_2O
ER	ER	ER	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Methoxymalonyl-ACP → CO_2 + ACP	Methoxymalonyl-ACP → CO_2 + ACP
AT	AT	AT		
DH	DH	DH	→ H_2O	→ H_2O
ER	ER	ER	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Methylmalonyl-CoA → CO_2 + CoA	Methylmalonyl-CoA → CO_2 + CoA
AT	AT	AT		
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Methylmalonyl-CoA → CO_2 + CoA	Methylmalonyl-CoA → CO_2 + CoA
AT	AT	AT		
DH	DH	DH	→ H_2O	→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Methoxymalonyl-ACP → CO_2 + ACP	Methoxymalonyl-ACP → CO_2 + ACP
AT	AT	AT		
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
AT	AT	AT		
DH	DH	DH	→ H_2O	→ H_2O
ER	ER	ER	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Methylmalonyl-CoA → CO_2 + CoA	Methylmalonyl-CoA → CO_2 + CoA
AT	AT	AT		
DH	DH	DH	→ H_2O	→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+

Table (6.11) Alignment of experimentally determined VS predicted predicted domain activities of oxazolomycin

Apparent	Real	Predicted	Real	Predicted
F	F		10-CHO-THF → THF	
A	A	A	Glycine→ H_2O	Glycine→ H_2O
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH		
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
MT	MT	MT	SAM→SAH	SAM→SAH
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
MT	MT	MT	SAM→SAH	SAM→SAH
C	C	C	Glycine→ H_2O	Glycine→ H_2O
A	A	A		
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
DH	DH	DH	→ H_2O	→ H_2O
KR	KR		NADPH + H^+ → NADP+	
MT	MT		SAM→SAH	
KS_0				
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Methoxymalonyl-ACP → CO_2 + ACP	Malonyl-CoA → CO_2 + CoA
KS_0				
DH	DH	DH	→ H_2O	→ H_2O
ER	ER	ER	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KR	KR	KR	NADPH + H^+ → NADP+	NADPH + H^+ → NADP+
KS	KS	KS	Malonyl-CoA → CO_2 + CoA	Malonyl-CoA → CO_2 + CoA
MT	MT	MT	SAM→SAH	SAM→SAH
C	C	C	Glycine→ H_2O	Glycine→ H_2O
A	A	A		
MT	MT	MT	SAM→SAH	SAM→SAH