KU LEUVEN

FACULTY OF BIOSCIENCE ENGINEERING
FACULTY OF ENGINEERING SCIENCE
FACULTY OF MEDICINE
FACULTY OF SCIENCE

Investigating and exploiting combinatorial diversity in Nature's drug factory

Antoine A. Ruzette

Promotor: Prof. Joleen Masschelein

Supervisor: Angus Weir

Department: Department of Molecular Biotechnology of Plants and Micro-organisms

Division: Masschelein Lab - Laboratory for Biomolecular Discovery and Engineering

---

Dissertation presented in fulfillment of the requirements for the degree of Master of Science in Bioinformatics at the Katholieke Universiteit Leuven.

Wednesday 3rd August, 2022

Declaration

This dissertation is part of the examination and has not been corrected after defense for eventual errors. Use as a reference is permitted subject to written approval of the promotor stated on the front page.

Antoine A. Ruzette, Wednesday 3rd August, 2022

Acknowledgements

The last two years have been the most meaningful of my education journey. I eventually discovered what daily drives and stimulates my curiosity: bioinformatics academic research in a vibrant and passionate environment.

First and foremost, I would want to express my gratitude to Prof. Joleen Masschelein for her unwavering support. She encouraged my research project and academic ambitions, but most significantly, she provided me with the ideal master's thesis environment. That environment convinced me to pursue academic research in bioinformatics. I thank her for her guidance.

Furthermore, without Angus Weir, Postdoctoral Researcher at the Laboratory for Biomolecular Discovery and Engineering, my experience in the laboratory would not have been the same. He mentored me in the finest way possible on a regular basis. I thank him for sharing his expertise with no limits, teaching me the fundamentals of bacterial genome engineering, and for improving this master's thesis through vibrant discussions.

The Laboratory for Biomolecular Discovery and Engineering at the VIB - KU Leuven is a unique environment filled with ambitious, smart and driven people. Hans, Ruben, Dries, Alejandro, Sophie, I am grateful for your kindness.

I also want to thank my parents for always being attentive and supporting, despite the fact that they do not understand a single word about bioinformatics.

Finally, I thank Emily Meeus and Angus Weir for proof-reading this paper.

Abstract

Natural products are an immensely diversified family of bacterial secondary metabolites harbouring biologically relevant capabilities, of which polyketides and non-ribosomal peptides are samples, and whose production is catalysed by polyketide synthases and non-ribosomal peptide synthases (PKSs and NRPSs).

The current study, conducted by the Laboratory of Biomolecular Discovery at the Katholieke Universiteit Leuven and the Flanders Institute for Biotechnology, aimed to characterize the synthetic capacities of two PKSs found in the bacterial genomes of *Caballeronia udeis* and *Massilia flava*. Methods from bacterial genome engineering and bioinformatics were combined to accomplish this.
On the genome engineering side, the catalytic activities of the *trans*-AT PKSs from *C. udeis* and *M. flava* were attempted to be disrupted. The metabolite synthesized by the PKSs from both bacteria would have been highlighted by comparing the metabolomes produced by intact and knocked-out PKS genomes. The use of pGPI as plasmids for bacterial tri-parental conjugation proved unsuccessful. Another plasmid, pSF100, is now being studied in order to conduct conjugation in a bi-parental design.
On the computational biology side, KS domains were utilized as core structure predictors in order to uncover the conserved domain patterns across polyketide synthases. To that purpose, the *trans*-AT PKS Annotation and Comparison Tool (transPACT) was tailored to the study's objectives. The development of an automatic specific naming structure for the antiSMASH ClusterBlast database, as well as the implementation of the program on the Linux environment of the Vlaamse Supercomputer (VSC), were among the modifications. Two approaches were developed: the query targeted approach and the genome mining approach.
As a result of the query targeted approach, extensive conservation patterns were identified among PKSs from *Aquimarina* sp. AU58, *Lysobacter enzymogenes*, *Massilia violaceinigra*, *Massilia* sp. Root335, *Caballeronia udeis* and *Massilia flava*, indicating that these bacteria probably have similar synthetic capacities. These closely related PKSs will aid in the identification of synthesised metabolites by expanding the quantity of data previously available.
The uniqueness potential of *C. udeis*, *M. flava*, and related PKSs was confirmed by the genome mining method. Indeed, it was shown that these PKSs cluster within a heterogeneous clade rich in non-conserved KS domains, distinct from previously characterised clades. Moreover, a PKS from the *Chromobacterium vaccinii* XC014 strain was identified to be highly similar, therefore it was included to the group of *C. udeis* and *M. flava* related PKSs.

Finally, the present study witnessed transPACT implemented in the Laboratory for Biomolecular Discovery and used to identify *C. udeis* and *M. flava* related polyketide synthases. Furthermore, attempts to edit *M. flava* and *C. udeis* genomes disregarded tri-parental conjugations using pGPI as plasmid in favor of bi-parental conjugations using pSF100.

Abbreviations

BGC	Biosynthetic Gene Cluster
PKS	Polyketide Synthetase
AT	Acyltransferase
KS	Ketosynthase
TE	Thioesterase
KR	Ketoreductase
ER	Enoylreductase
DH	Dehydratase
ACP	Acyl Carrier Protein
T	Thiolation
MT	Methyltransferase
AL	acyl-ligase
SAM	S-adenosyl-methionine
ECH	enoyl-CoA hydratase
GNAT	GCN5-related N-acetyltransferase
Me	Methyl
CoA	Coenzyme A
P450	Cytochrome P450
DEBS	6-Deoxyerythronolide B Synthase
6-DEB	6-Deoxyerythronolide B
NRP	Non-ribosomal peptides
NRPS	Non-ribosomal peptide synthetase
transPACT	trans-AT PKS Annotation and Comparison Tool
transATor	trans-AT polyketide synthase predictor
FAS	Fatty Acid Synthase
CU	*Caballeronia udeis*
MF	*Massilia flava*
pHMM	profile Hidden Markov Models
NC	Non Conserved

Contents

1 Motivations and Context **1**

2 Literature Review **4**

2.1 Introduction: Bacterial Secondary Metabolism Is Immensely Diverse 5

2.2 Bacterial Biosynthetic Gene Cluster 6

2.2.1 Bacterial Biosynthetic Gene Clusters are the blueprints for the assembly lines 6

2.2.2 The collinearity rule lays a relationship between the catalytic domains order and the subsequent chemical structure 7

2.3 Assembly Lines Enzymology 9

2.3.1 Polyketides Assembly Lines 9

2.3.1.1 Polyketides are produced by modular mega-enzymes, named Polyketide Synthases (PKS) 10

2.3.1.2 The canonical organization of Polyketide Synthases 11

2.3.1.3 Initiation and termination modules perform distinct functions than elongation modules 14

2.3.1.4 Post-PKS tailoring processes 14

2.3.2 Non-Ribosomal Peptides (NRP) Assembly Lines 17

2.3.2.1 The canonical organization of Non-Ribosomal Peptide Synthases 17

2.3.2.2 Auxiliary domains in NRP Assembly Line 19

2.3.2.3 Post-tailoring enzymes in NRPS 20

2.4 *trans*-AT mediated assembly lines 21

2.4.1 *trans*-AT PKSs rely on a single or several standalone enzymes responsible of the selection and the loading of the acyl derivatives 21

2.4.2 *trans*-AT PKS pathways introduce a new layer of diversity within the natural products landscape . . . 21

2.4.3 *trans*-AT PKS ketosynthase (KS) domains as core structure predictors . . 23

2.5 A snapshot of the computational toolbox available for the investigation of *trans*-AT Polyketide Synthases . . . 24

2.5.1 antiSMASH . . . 24

2.5.2 transPACT . . . 24

2.5.2.1 A substrate specificity annotation tool of *trans*-AT ketosynthase (KS) domains . . . 25

2.5.2.2 A phylogenetic comparison tool for *trans*-AT PKSs sharing module blocks . . . 26

2.5.2.3 Performance of transPACT . . . 27

2.5.3 transATor: *trans*-AT PKS Polyketide Predictor . . . 28

3 Investigation of *trans*-AT assembly lines in *Massilia flava* and *Caballeronia udeis* 29

3.1 Methods . . . 30

3.1.1 Bacterial genome engineering . . . 30

3.1.1.1 Cultivation . . . 31

3.1.1.2 Recombinant plasmids construction . . . 31

3.1.1.3 Transformation of chemically competent *E. coli* cells . . . 32

3.1.1.4 Pre-screening, sequencing and selection . . . 33

3.1.1.5 Conjugation . . . 34

3.1.2 Computational Biology . . . 36

3.1.2.1 Data acquisition . . . 36

3.1.2.2 antiSMASH . . . 37

3.1.2.3 transPACT . . . 37

3.1.2.4 transATor . . . 38

3.2 Results . . . 40

3.2.1 Bacterial genome engineering . . . 40

3.2.1.1 Carbon sources investigation 40
3.2.1.2 Construction of recombinant plasmids 40
3.2.1.3 Selection of transformed chemically competent SY327 *E. coli* cells 41
3.2.1.4 Conjugation in a tri-parental design 42
3.2.2 Computational Biology 42
3.2.2.1 Identification of *trans*-AT gene clusters and annotation of catalytic domains using antiSMASH 42
3.2.2.2 *trans*-AT KS substrate specificity annotation using transPACT . 42
3.2.2.3 Dendrogram representation of PKSs using transPACT 46
3.3 Discussions 47
3.3.1 Bacterial conjugation using pSF100 plasmid 47
3.3.2 Conservation of module blocks across *trans*-AT PKSs related to *C. udeis* and *M. flava* 47
3.3.3 Limitations of transPACT 51
3.3.3.1 Open reading frame against genetic order 51
3.3.3.2 Computational complexity 52
3.3.4 Genome mining approach using transPACT 52
3.3.4.1 Adaptations to a genome mining approach 52
3.3.4.2 Limitations 53
3.3.4.3 Conserved module blocks across PKSs identified in the antiSMASH ClusterBlast database 53
3.4 Data Accession 55

4 Concluding Remarks **56**

Bibliography **59**

Appendices **64**

Chapter 1

Motivations and Context

In 1845, Charles Darwin described his journey through indigenous lands in his extraordinary book *On the Origin of Species* [1]. During his five-year expedition aboard the HMS Beagle, he would elegantly lay the foundation of modern evolutionary biology. Almost two centuries later, in the Laboratory for Biomolecular Discovery and Engineering at the KU Leuven and at the Flanders Institute for Biotechnology, the indigenous lands that scientists seek to understand are microbial genomes. The microbial landscape produces a broad panel of chemicals. This structural and functional diversity arises from countless decades of evolutionary processes such as genetic mutation, gene flow, genetic drift, horizontal gene transfer, speciation, adaptation or natural selection [2]. The medium and the analytical methods used nowadays are different. However, the fundamental question remains the same. Can we comprehend and exploit the diversity found in nature?

There still exists a myriad of uncharacterised (i.e. orphan) or partially characterised microbial genomes. In this context, Prof. Joleen Masschelein took the lead in 2020 of a laboratory dedicated to the understanding and engineering of bacterial genomes with a drug discovery end goal. Since I began my bioengineering journey, I have always been interested in the discovery of novel compounds that possess beneficial assets for the society. The Laboratory for Biomolecular Discovery and Engineering at the KU Leuven and at Flanders Institute of Biotechnologies was an evident continuity in that direction.

Furthermore, the microscopic world is as diverse as passionate to my eyes. This idea has been a major driver in my choice of research topic. I have not been disappointed. For the past decades, the discovery of drugs and relevant molecules underwent profound changes. One might argue a switch from a chemical therapeutic revolution to a biological therapeutic revolution in the midst of the 1960s. The potential for bacteria to occupy the role of drug factories has now been proven. Actually, the first breakthrough dates back from the beginning of the 1920s when scientists engineered bacteria to include the gene for the human insulin protein [3]. Since then, scientists searching for natural products have released more than 20 pharmaceuticals on the markets [4], including immunosuppressants such as rapamycin [5], hypocholesterolemics such as lovastatin [6], anticancer agents such as doxorubicin [7] and antimicrobials such as erythromycin [8]. The applications of natural products have revolutionized human society. They are relevant in numerous fields including pharmaceuticals, agriculture, cosmetic, animal health and nutrition.

Darwin drew hypothesis from observations. The latter observational framework is one that I identified and to which I still identify nowadays. We observe, perturb and analyse microbial metabolisms to better understand the synthetic capabilities of their genomes. On a daily basis, the idea of better understanding the unknown feels familiar. An atmosphere where an envy for discovery is at its heart is what drives me. I confirm that the Laboratory for Biomolecular Discovery and Engineering is one of the kind.

The challenge in drug discovery is two-folded. The challenge in drug discovery is twofold. To begin, novel biosynthetic pathways and their synthetic capabilities are attempted to be found using bacterial genome engineering and bioinformatics technologies. Second, engineering the bacterial genome is now more than ever at the forefront of the next scientific discoveries in natural products synthesis. The ultimate objective is to modulate biosynthetic assembly lines in the same manner that one may play with LEGO bricks in order to fulfill customised and specific biosynthetic capabilities. Because of their modular architecture, Polyketides Synthases (PKSs) and Non-Ribosomal Peptide Synthases (NRPSs) appear to be a particularly good starting point for the synthetic tuning of natural products.

Practically, the present Master's Thesis compiles together a literature review on natural products assembly line enzymology with a focus on the mechanism of trans-acyltransferase (*trans*-AT) PKS and the observations made during a 3-months research period in Prof. Masschelein's laboratory. The first part discusses the current groundwork that has been accumulated over the years on PKS assembly lines, NRPS assembly lines and *trans-AT* PKS biosynthetic pathways. The content of the current bioinformatics toolbox available for natural product exploration was also reviewed. The second part acts as a logbook of my research on *Caballeronia udeis* and *Massilia flava*, carried out with Prof. Joleen Masschelein and Postdoctoral researcher Angus Weir.

Chapter 2

Literature Review

2.1 Introduction: Bacterial Secondary Metabolism Is Immensely Diverse

Metabolism in bacteria can be split into two categories depending on the stage of bacterial growth, use and function of compound generated. These are defined as primary and secondary metabolism. The primary metabolism of organisms is the set of metabolic processes serving the requisite functions of life. Such a requisite function is aerobic respiration in obligate aerobic bacteria - without that function unaltered, bacteria do not sustain. It is an ultimate necessity for bacteria to accomplish their bacterial lives: grow. One could consider that the primary metabolism is universal among microbes. While this is broadly true, it is an over-simplification. However, we observe that microbes do more than simply survive. They evolve, diversify, compete and conquer. Thus an additional metabolism must exist. This secondary metabolism enables the microbes to gain evolutionary advantages over other species and/or over individuals from the same species. It mainly contrasts with the primary one in a way that it does not serve essential functions in the life of microorganisms. The altered synthesis of a secondary metabolites will not directly harness the survival of microbes in an ideal environment. As aforementioned, microbes prosper in their environment when their growth rate is sufficient, but growing faster than your competitor is not the only strategy that bacteria have developed. The production of an arsenal of secondary metabolites arise as an additional strategy to conquer their environments.

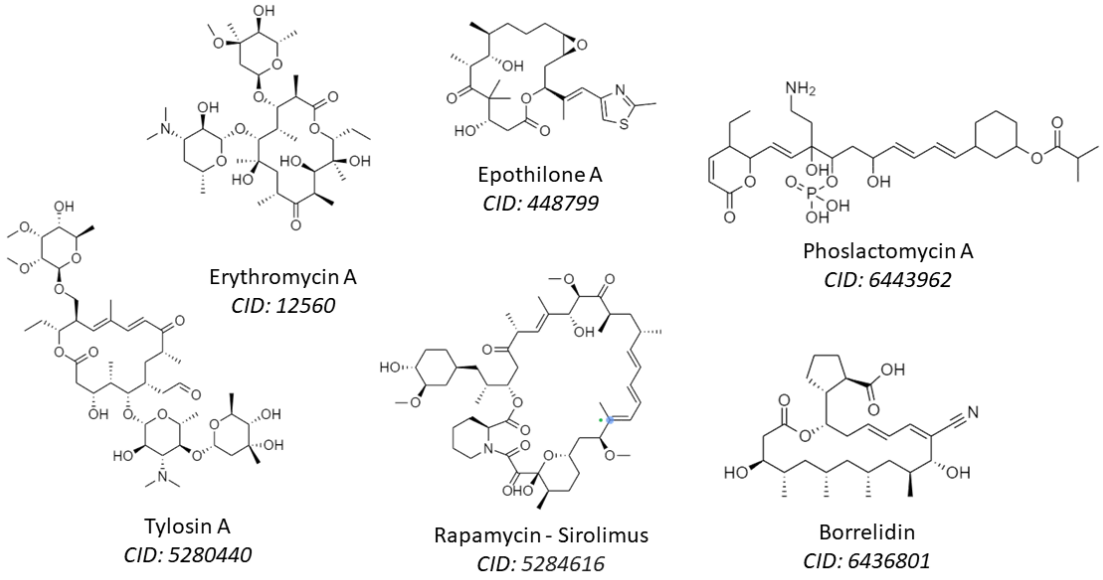


Figure 2.1: **A snapshot of the diversity of complex bioactive polyketides.** PubChem CIDs were accessed on June 21, 2022.

Natural products, as a synonym for secondary metabolites, are widely diverse. Since the building blocks and substrates used in the secondary metabolism are rather limited [9], as they are derived from a few primary metabolites, we can then explain such a biomolecular diversity by the combinatorial character of biosynthetic pathways. Indeed, from a combination of few simple building blocks, assembly lines are capable of producing an immensely diverse arsenal of natural products. This structural and functional diversity of secondary metabolites emerges from minor changes in the number of incorporated carbon atoms and their arrangement in the skeleton chain. From the incorporation of new atoms such as nitrogen, oxygen, chlorine or sulfur to a various of chemical modifications such as cyclization, methylation, oxidation and reduction,

the landscape of natural products is immensely diverse. A snapshot of this chemical diversity is illustrated by diverse complex polyketides on Figure 2.1.

In the following sections, we review a fraction of that diversity with two categories of microbial secondary metabolites, that is polyketides (PK) and non-ribosomal peptides (NRP). Within the scope of this research, the emphasis is on Polyketide Synthases (PKSs).

2.2 Bacterial Biosynthetic Gene Cluster

Before diving in the biochemistry of assembly lines enzymology for the cases of polyketides and non-ribosomal peptides, we first need to understand the atypical architecture of bacterial genomes. During the past decades, the sequencing of thousands of genes encoding for enzymes catalyzing the biosynthesis of polyketides and non-ribosomal peptides unmasked a peculiar bacterial genomic architecture structured in physical clusters [10].

2.2.1 Bacterial Biosynthetic Gene Clusters are the blueprints for the assembly lines

The 6-Deoxyerythronolide B (6-dEB) biosynthetic pathway is a particularly well characterised *cis*-AT Polyketide Synthase produced by *Saccharopolyspora erythraea* [11]. It will be used as a general illustration for canonical assembly line enzymology as depicted on Figure 2.2. *cis*-AT PKS corresponds to the standard view, as opposed to *trans*-AT PKS that were discovered later. The distinction between the two will become more obvious further down the road (See section 2.4). Post assembly line modifications of 6-dEB lead to the synthesis of erythromycin A, a molecule discovered in 1952 that possesses antibiotic properties. It is used to treat several infections in the respiratory tract, skin infections, chlamydia infections, pelvic inflammatory disease, and syphilis. Initially observed in the genome of *Escherichia Coli*, the conserved or persistent genes responsible for the biosynthesis of secondary metabolites are organized into physical clusters. This observation has then been extended to many bacterial genomes [12]. The genes responsible for the synthesis of erythromycin A form a cluster within the genome of *Aeromicrobium erythreum*. Such clusters have been named Biosynthetic Gene Clusters (BGCs) as they act as the blueprints for biosynthesis. A single cluster is therefore responsible for the synthesis of a single and specific metabolite. Gene clustering patterns are observed at the genome level. However, patterns harbouring a modularity aspect can be observed at various levels within bacterial genomes, illustrated with the erythromycin A PKS.

1. Genome level

   All genes responsible for the biosynthesis of erythromycin A are clustered within a BGC in the genome of *Aeromicrobium erythreum*. The size of the BGC is 61,845 base pair long[1]. The BGC contains the core biosynthetic genes but also additional biosynthetic genes and others functional genes. Figure 2.2 depicts solely core biosynthetic genes, with the exception of one extra biosynthetic gene that encodes for the thioesterase (TE) domain, which is responsible for the release of the synthesised molecule from the assembly line. The biochemistry of catalytic modules is emphasized in Section 2.3
   Thus, the genome of bacteria can be partially seen as a succession of genetic clusters, with each of them responsible for the synthesis of a specific compound.

[1] Accesed on MiBIG under the accesion number BGC0000055, on May the 5^{th}, 2022.

2. PKS level

   The biosynthetic gene cluster consists of open reading frames with a modular architecture, with each module classified by its ability to perform one acetate extension and additional processing. Each open reading frame accounts for a multimodular synthase catalyzing the biosynthesis of a specific part of the final compound, namely 6-deoxyerythronolide B (6-dEB). For the 6-dEB case, these synthases have been named DEBS1, DEBS2 and DEBS3. It should be noted that the open reading frame is different than the physical order in the genome, as illustrated by the coloured arrows on Figure 2.2.

3. Module level

   We continue our whistle stop tour within the bacterial genome, now reaching a module level. Each synthase is subsequently divided into modules that are responsible for a specific enzymatic function. For instance, the loading module, as its name suggests, is responsible for the loading of the initial substrate onto the assembly line.

4. Domain level

   Finally, each module is constructed from functional domains. Domains are the assembly line's LEGO bricks because they relate to a highly specific catalytic activity. Over the years, the scientific community has been able to classify them based on their function. It is now possible to predict the function of these sequences using alignment tools. It should be noted that not all domains have been discovered, and that parts of sequences remain uncharacterized, leaving room for future scientific investigations.

This multi-level clustered architecture confers a highly modular aspect to polyketides assembly lines. The latter becomes particularly handy when it comes to engineering polyketides assembly lines. Backed by this modular aspect of assembly lines, a correlation between the core molecular structure of the synthesised polyketide and the order of the enzymatic domains within the PKS has been observed. It has been named the collinearity rule.

2.2.2 The collinearity rule lays a relationship between the catalytic domains order and the subsequent chemical structure

The catalytic domain order of PKSs and the sequence of chemical functional groups in polyketide backbones frequently correlate. This rule is not a generalization and should not be assumed. Nonetheless, the collinearity rule is instinctive and lays the foundation for the prediction of the core linear structure of arising polyketides.

Figure 2.2 helps to grasp the idea of the collinearity rule, illustrated by the erythromycin A biosynthesis. It is possible to predict the chemical modifications the incorporated building block will undergo based on the available knowledge about each domain's function in the assembly line. For example, Module 3 (encoded by the EryA II gene) misses some domains such as ketoreductase (KR), enoylreductase (ER) and dehydratase (DH). The grey circle corresponds to the missing KR domain. It follows that none of the two carbonyl groups from the incorporated

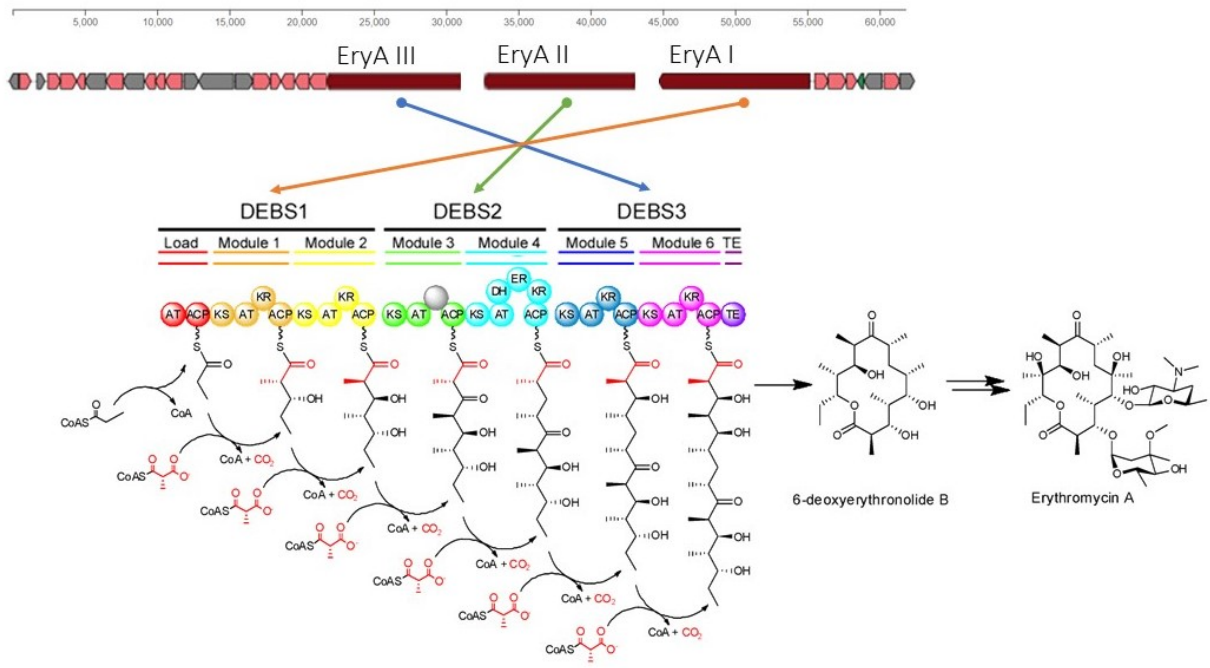


Figure 2.2: **Biosynthesis of erythromycin A by 6-deoxyerythronolide B synthase (DEBS)**, overlaid with the structure of its corresponding biosynthetic gene cluster (MIBiG: BGC0000054) from *Aeromicrobium erythreum*. $EryAI \Rightarrow DEBS1, EryAII \Rightarrow DEBS2$ and $EryAIII \Rightarrow DEBS3$. The grey circle illustrates the lack of KR domain.
Figure reprinted from Kwan et al., 2011 [13].

malonyl-CoA will be reduced to a hydroxyl group. This can be deduced from minimal information such as the DNA sequence of the synthase. Throughout the paper, we will mention several catalytic domains. Representations of each of these domains with their abbreviations are shown on Figure 2.3.

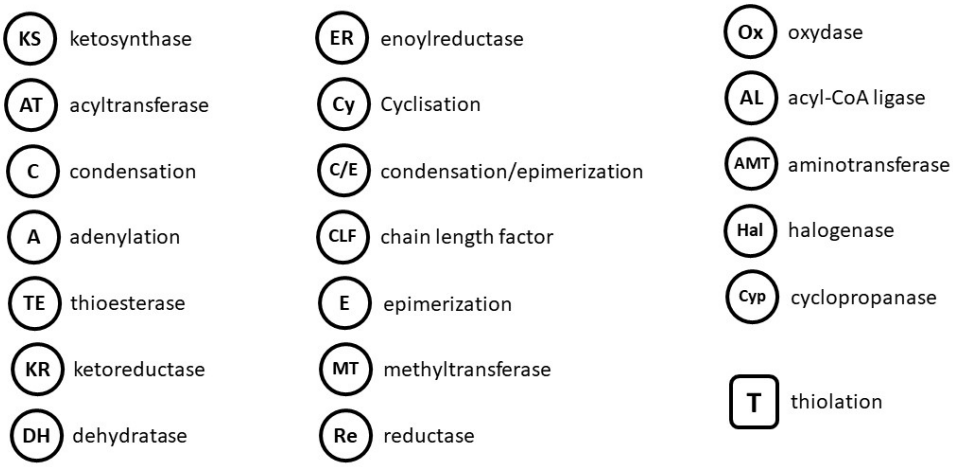


Figure 2.3: **Representations of catalytic domains with their abbreviations.**

2.3 Assembly Lines Enzymology

The general mechanism behind assembly lines of polyketides and non-ribosomal peptides is of the iterative incorporation of monomers into a polymer by repetition of chemical condensation steps, resulting in the elongation of a backbone. A typical assembly line mechanism can be divided into three main steps, the loading, the extension and the release.

1. Loading
   A first substrate, named the starting unit, is loaded on the loading module of the assembly line.

2. Extension
   The starting unit is then extended by another specific substrate in order to begin the formation of a chain. The extension is enabled by the condensation reaction involving those two first substrates. This process of condensation reaction continues over and over until reaching the release module. Within this extension phase, the backbone of the polymer elongates but also undergoes a variety of chemical processes. The latter processes are specific and will shape the final chemical structure of the synthesized natural product.

3. Release
   Once the polymer reaches the end of the assembly line, the release module enables the release of the product from the biosynthetic machinery. Even though the biosynthesis of the product is not entirely complete because the gene cluster also encodes for post-tailoring enzymes, the chain can be considered as finalized at this point.

   Post-PKS modifications are also relevant biochemical processes to reach a bio-active product (discussed in Section 2.3.1.4 for PKSs and in Section 2.3.2.3 for NRPSs).

An interesting vision to see assembly lines as information-transfer catalysts has been formulated by C. Walsh and M. Fischbach in 2006 [14]. The information-transfer concept in biology is typically illustrated in a genetic context, as genetic information passed on via enzymes (i.e. DNA and RNA polymerase) from DNA to mRNA and from mRNA to protein, or against the *central dogma* via reverse transcription [15, 16]. Within the genetic context, enzymes such as RNA and DNA polymerase are seen as the information-transfer catalysts. In the context of natural products assembly, the information-transfer catalysts are mega-enzyme of the synthase type, made of several enzymatic domains. The information passed on are the blueprints for the synthesis of natural products.

2.3.1 Polyketides Assembly Lines

Polyketide biosynthesis will now be focused down to the general mechanism underlying polyketide and non-ribosomal peptide assembly lines. The same logic applies to fatty acid synthesis, which is catalyzed by Fatty Acid Synthases (FASs). The assembly line concept stays unchanged. Recursive chemical condensation steps are used to combine a small set of monomers into a complex polymer backbone [14]. This stepwise mechanism accesses an incredible amount chemical diversity from simple starting materials. This is particularly interesting from an assembly line engineering point of view. The products generated by the manipulation of PKSs though addition, removal and substitutions of domains or modules are so-called unnatural products [17]. The development and the diversification of the (un)natural products portfolio will continue to

thrive in the coming decades, especially with the use of new bioinformatics tools. The monomers incorporated in PKS assembly lines are acyl-CoA thioesters (Figure 2.4) [14]. These monomers are typically derived from the primary metabolism, as discussed in the introduction.

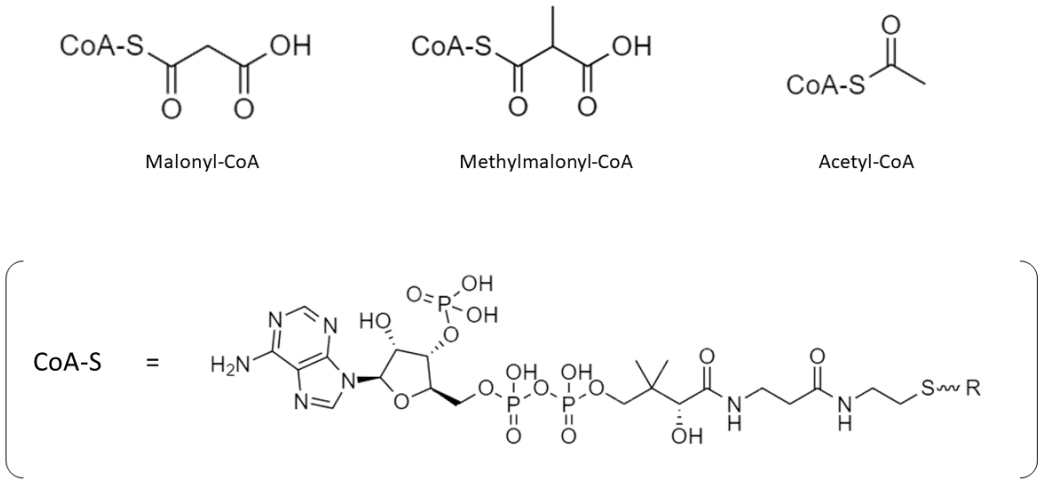


Figure 2.4: **Monomers incorporated during the chain initiation and elongation catalyzed by polyketide synthases typically derive from acyl-CoA thioesters.**

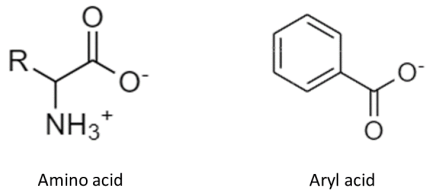


Figure 2.5: **Monomers incorporated during the chain initiation and elongation catalyzed by Non-Ribosomal Peptide Synthases typically derive from amino acids and aryl acids.**

2.3.1.1 Polyketides are produced by modular mega-enzymes, named Polyketide Synthases (PKS)

Polyketide Synthases are modular mega-enzymes varying in size, ranging from the kilodalton to the megadalton order. In extreme case, the size of the synthase can reach up to 5 MDa. The erythromycin A synthase (shown in Figure 2.2) consists of seven modules (i.e. loading module, Module 1-6 and termination module) spread across three peptides of around 200 kilodaltons each, with a total of 10,283 amino acids [18]. As explained in the general case of assembly lines, PKSs are organised in modules and domains, usually distributed over large peptides. In this section, we will focus on *cis*-AT Polyketide Synthases. The catalytic domains and the ACP (T) are connected in a *cis* fashion to generate modules in such organizations. Thus, there exists

an AT domain embedded in each extension module. In the next section, we will extend this literature review to the *trans*-AT systems. Type I systems encompass both *cis*- and *trans*-AT assembly line, as it refers to the modular architecture of PKSs and NRPSs. Type II and type III assembly lines are not included in the scope of the present literature review.

Each module is then composed of enzymatic domains. The core biosynthetic synthesis of PKS involves the step-wise elongation of the polyketide carbon backbones, enabled by core enzymatic domains named acyltransferase (AT), ketosynthase (KS) and acyl-carrier proteins (ACP or T). In *cis*-AT PKS, the combination of these three domains is the minimal architecture of an extension module.

2.3.1.2 The canonical organization of Polyketide Synthases

The common step essential to the biosynthesis of all polyketides is the condensation step enabling the elongation of the growing polyketides. Initiation and termination modules will be discussed subsequently as they differ in function thus in organization.

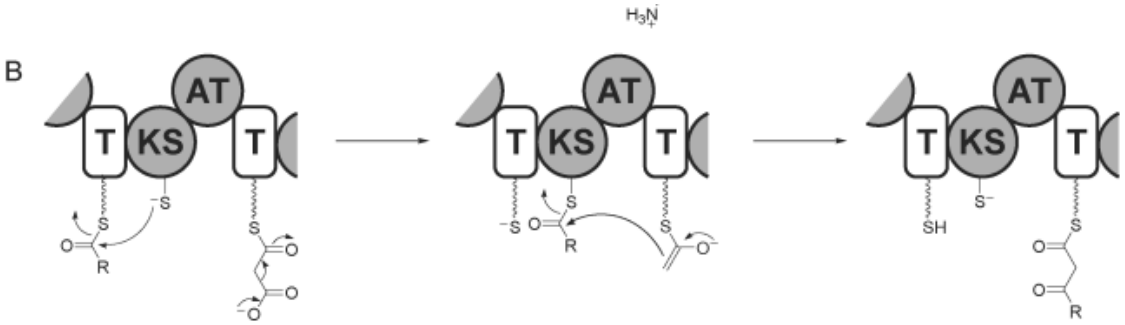


Figure 2.6: **Condensations are catalyzed by the combined action of ketosynthase, acyltransferase and thiolation domains during polyketide biosynthesis.**
Figure reprinted from Fischbach *et al.*, 2006 [14].

Core domains

Type I modules must have at least three specific domains in order to perform an extension step, which corresponds to the incorporation of an acyl unit within the polyketide chain.

1. Acyltransferase - AT

   This 50 kDa domain, abbreviated as AT, is responsible for selecting and loading carboxylic extender units derived from their CoA thioesters on the biosynthetic assembly line. AT domains function as specificity gatekeepers by picking either malonyl-CoA or methylmalonyl-CoA as substrates, as shown in Figure 2.4. They transfer the acyl group of the C_3 (malonyl-CoA) or C_4 (methylmalonyl-CoA) substrate to the pantetheinyl arm's thiolate terminus of the thiolation domain. This transfer from the AT domain to the T domain (or the ACP domain in the case of PKSs) is depicted in Figure 2.7a. It is a net transthiolation reaction, which is an energy neutral transfer of an acyl group, as described by Fischbach *et al.* in their review published in 2006 [14].

2. Thiolation domain - ACP or T

   The acyl carrier protein (ACP) is a four-helix bundle that measures 8 to 10 kilodaltons

and has a phosphopantetheine arm that serves as a tethering point for the growing polyketide chain and the soon-to-be-incorporated extender unit [14]. The thiolation domain is responsible for providing a thiol arm[2] to which substrates and intermediate acyl chains are covalently tethered during chain elongation events [14]. Vance *et al.* depicted in 2008 the thiol arm as a 'swinging arm' [20]. Its role is to keep the chemical protagonists of the biosynthesis attached to the assembly line domains. This 'swinging arm' is pictured as a wiggly line on the figures e.g. Figure 2.7. Furthermore, thiolation domains like ACP can be viewed as protein way stations *videlicet* intermediate stopping points in the process of polyketide biosynthesis [21]. In literature, thiolation domain has been referred to as either ACP or T in PKSs. In the present paper, T will exclusively be used to refer to thiolation domains.

3. Ketosynthase - KS

   The KS domain is a 50 kDa domain that catalyzes the formation of C-C bonds, that is a Claisen-like condensation reaction [22]. Figure 2.8b illustrates the biochemistry of the mechanism of action of ketosynthases. First, an upstream (T_{n-1}) acyl group is transthiolated from an ACP domain to the active site (-SH) of a conserved cysteine on the KS domain. The (methyl)malonyl-CoA tethered to the downstream thiolation domain is then decarboxylated by the KS domain (T_n). As a result, a β-ketoacyl [3] is formed, which is tethered to the downstream T_n domain by an S-C bond. The resultant substrate will subsequently be used for the next Claisen condensation-driven extension step. Following that, each iteration adds two carbons to the polyketide backbone when the integrated extender unit is malonyl-CoA and three carbons when the incorporated unit is methylmalonyl-CoA. At each condensation iteration, the expanding polyketide chain is translocated from an upstream T_{n-1} domain to the next upstream T_n domain, much like the behavior of an assembly line.

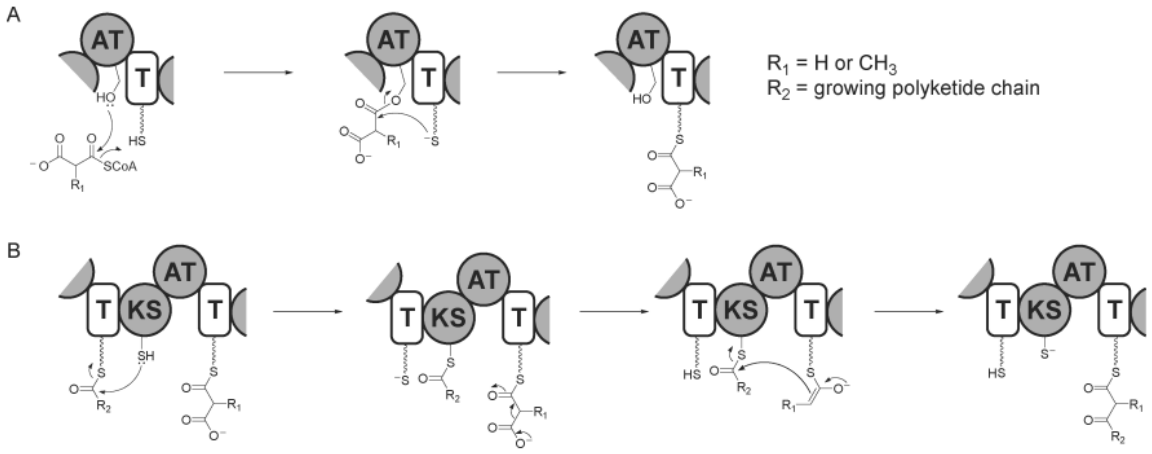


Figure 2.7: **Acyltransferase and ketosynthase domains activities catalyzed by PKSs.** (A) AT domains catalyze the selection and loading of carboxylic extender units on the assembly line, then their transfer to the downstream T_n domain. (B) KS domains catalyze the formation of C-C bonds, that is a Claisen-like condensation reaction, as a result of (methyl)-malonyl-S-T_n decarboxylation.
Figure reprinted from Fischbach *et al.*, 2006 [14].

[2] A thiol or a thiol derivative is any organosulfur compound of the form R-SH, where R represents an alkyl or any other organic group, as recently defined by Aisha H. Al-Moubaraki *et al.*, 2021 [19].

[3] It should be noted that this intermediate is sometimes referred as a β-carbonyl. [23]

Auxiliary domains of importance

Three additional catalytic domains are particularly important in the assembly line mechanism of polyketides, even if they are not invariably present as they are in Fatty Acids Synthases (FASs). PKSs are closely phylogenetically related to FASs. The present section discusses a major example of tailoring on assembly lines. This phenomenon is partially responsible for the outstanding diversity characterizing the polyketides landscape. They add up a layer of potential structural combinations. Their functions have been well established through classical FAS enzymology research, they proceed in the following order: $KR \longrightarrow DH \longrightarrow ER$.

Once the condensation step has been performed by the action of the aforementioned $KS - AT - ACP$ trio, the resulting condensation product (i.e. β-ketoacyl-S-T) may endure a series of reduction steps. These tailoring domains modify the polyketide growing chain, potentially leading to a fully reduced backbone at the β-carbon. Without the actions of the $KR - DH - ER$ domains, a C=O carbon would be incorporated in the polyketide growing chain.

1. Ketoreductase - KR

   The ketoreductase domain catalyzes the reduction of the β-ketoacyl-S-T intermediate formed in the preceding condensation iteration. As its name suggests, ketone reductase (or ketoreductase for short) is a NADPH-dependent oxidoreductase that catalyzes the reduction of a ketone into a secondary alcohol [24]. Thus, it reduces the β-ketoacyl-S-T intermediate into a β-hydroxyacyl-S-T intermediate. The first reactional step in Figure 2.8 depicts the action of the KR domain on the β-carbon, the carbonyl (=O) functional group is transformed into a hydroxyl (-OH) functional group.

2. Dehydratase - DH

   Additionally, the β-hydroxyacyl-S-T intermediate may be dehydrated to yield a α,β-enoyl-S-T intermediate. It basically catalyses the removal of a hydrogen atom from the hydroxyl functional group (as shown in the second step of Figure 2.8).

3. Enoylreductase - ER

   Finally, the alkene (i.e. α,β-enoyl-S-T) intermediate can be further reduced to form a saturated acyl-S-T under the action of the enoylreductase (ER) domain as illustrated by the last reaction of Figure 2.8.

It should be noted that the incomplete action of the latter enzymatic trio is also possible and would preserve intermediate forms due to the absence of one of the domains. As a result, the chemical structures that can result from the enzymatic activities of $KR-DH-ER$ can be either fully reduced forms or partially reduced forms such as α,β-enoyl, β-hydroxyacyl or β-ketoacyl moieties.
The actions of the $KR-DH-ER$ domains complete the polyketide backbone elongation process initiated by the $AT - KS$ tandem. The full reduction of the β-ketone yields a β-methyl group.

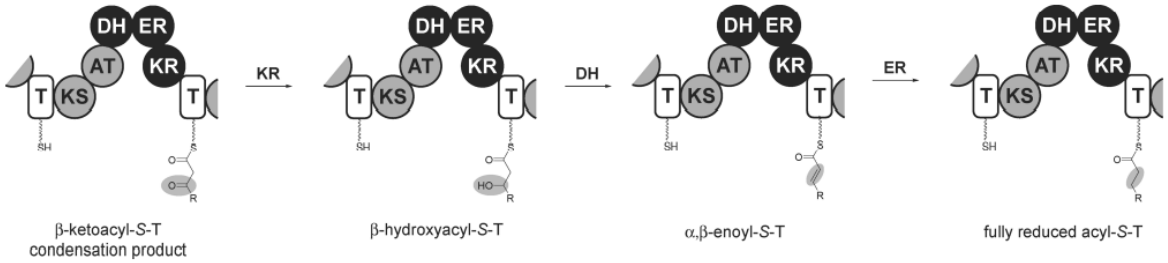


Figure 2.8: **β-carbon reduction cascade in PKSs catalyzed by ketoreductases, dehydratases and enoylreductases.** The KR domain catalyzes the reduction of the β-ketoacyl-S-T_n substrate into its β-hydroxyacyl-S-T form. The DH domains then may serve as a catalyst to dehydrate the earlier reduction product, yielding a α,β-enoyl-S-T substrate. The ER domain's activity may result in a further reduced substrate in the form of a completely saturated acyl-S-T. Reduced variants will be created dependent on the extent of the reduction.
Figure reprinted from Fischbach *et al.*, 2008 [14].

2.3.1.3 Initiation and termination modules perform distinct functions than elongation modules

The condensation step is not required in the initiation module, as its sole purpose is to load the starter unit onto the assembly line. In such a module, the activity of a KS module is therefore not necessary. The ketosynthase domain can exist within the module but be catalytically inactive, or it could be missing. In the former case, the outcome of the decarboxylation of the initial methylmalonyl-S-T_1 is different than previously seen. The starting acyl group on the assembly line will be a derivative of an acetyl- or a proprionyl-CoA. The decarboxylation still releases a CO_2 molecule in the process [14]. In the latter situation, an acyl-CoA, such as acetyl-CoA, is immediately loaded into the assembly line's initial thiolation domain (ACP) by the starting AT domain. As a result, the assembly line is now ready to begin its sequence of condensation steps.

When the assembly line reaches its completion, the termination or release module catalyzes the release of the polyketide full-length acyl chain from the final ACP domain. The release module contains a 35 kDa specific domain, called thioesterase (TE), harboring an CH_2OH active site to which the full-length chain can bind to ultimately form an acyl-O-TE intermediate [25]. Eventually, the catalyzed hydrolysis of the intermediate causes the full-length backbone to be released linearly from its covalent tethering with the terminal ACP domain [26].
TE domains can also have a cyclization effect, thus producing a final acyl chain harboring chemical cyclic structure.

2.3.1.4 Post-PKS tailoring processes

The assembly line mechanism seen so far produces a full-length scaffold, that is a soon-to-be polyketide. Usually, this acyl chain is not yet biologically active. Tailoring on assembly line already took place with the actions of auxiliary domains such as KR, DH and ER. To generate a structure with significant biological activity, post-assembly-line modifications are typically necessary. Post-tailoring enzymes catalyze this process. These enzymes' genetic blueprints are not included in the polyketide synthase genes. They do, however, share the same Biosynthetic Gene Cluster (BGC), close to the PKS core domains.
This post-processing of the final carbon scaffold arising from the assembly line adds up an additional layer of structural and functional diversity in polyketide biosynthesis [27]. In the present section, we will review the major post-assembly-line tailoring events namely glycosylation, oxidation and methylation. As each tailoring reaction is typically specific for a certain synthase, we hereby rely on the Erythromycin synthase case as a representative example of *cis*-AT PKS. In

2002, Rohr *et al.* [28] published a review about post-tailoring enzymes covering researches from 1985 to 2002. In 2010, Salas et al. [29] published another review on the subject, covering the knowledge accumulated from 2002 to 2009. These two publications laid the way for the current non-exhaustive review to be written.

Glycosylation

Glycosylation is the modification of an organic compound by addition of a carbohydrate cycle [30] catalysed by glycosyltransferases. It occurs widely on NRP and both on *cis*-AT and *trans*-AT polyketides. Three types of glycolysation may happen depending on the type of atoms on which the addition occurs: O-glycosylation, N-glycosylation and C-glycosylation. It should be noted that O-glycosylation is more common that the nitrogen- and carbon-driven ones [25]. As previously illustrated, the first step in the synthesis of the *cis*-AT polyketide erythromycin A is the synthesis of the 6-dEB intermediate (Figure 2.9). In order to go from the intermediate to the final bio-active form, that is erythromycin A. The 6-dEB molecule undergoes two O-glycosylations. Ultimately, the erythromycin synthase yields a final polyketide containing two cyclohexyl functional groups added during post-assembly-line tailoring steps.

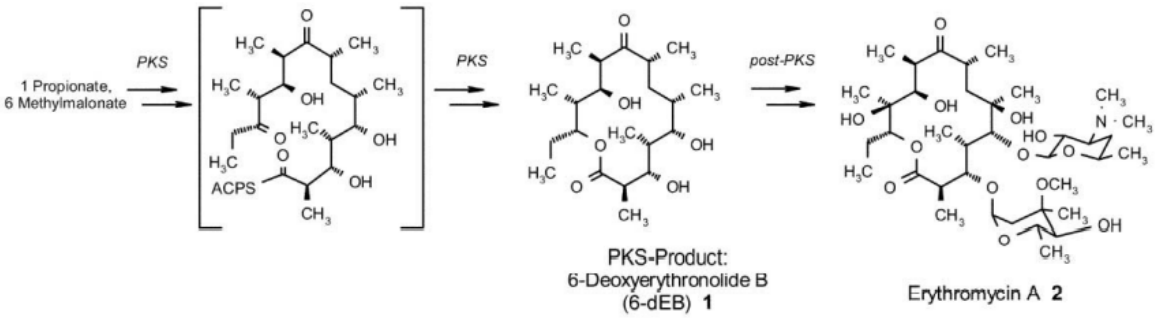


Figure 2.9: **Formation of erythromycin A from a 6-dEB polyketide intermediate results from post-PKS tailoring processes**. 6-deoxyerythronolide B (6-dEB) is a biologically inactive intermediate that must be tailored post-PKS to become active. In the context of erythromycin, two cyclohexyl functional groups are added to produce erythromycin A, the active form.
Figure reprinted from Jurgen *et al.*, 2002 [28].

The erythromycin A PKS is known to be the interplay of three post-assembly-line tailoring phenomena, namely glycosylation, methylation and oxidation [29].

Oxidation

Oxygenases are enzymes catalyzing the oxidation of substrates. Oxidation reactions are diverse. In 2010, Olano *et al.* [29] listed hydroxylations, epoxidations, anthrone oxidations and oxidative rearrangements as examples of common oxidation reactions. For the sake of using the erythromycin A synthase as an illustration, we will focus on hydroxylation reactions. According to the definition of Zhu *et al.* [31] in 2017, hydroxylation is the most common reaction type in phase I metabolism[4] and usually produces a chemically stable and more polar hydroxylated metabolite than the inoculated drug. Indeed, some natural products arising from PKS are hydrophobic. For example, the well-characterised case of erythromycin A is a relevant illustration of a hydrophobic PK scaffold that undergoes hydroxylation to increase its hydrophilicity [14]. The

[4]The phase I metabolism involves the metabolic activation of foreign compounds, typically seen as the addition of polar groups on the inoculated drug.

hydroxylation of the nascent erythromycin A is catalyzed by a heme-utilizing the cytochrome P450 [32].

Methylation

Methylation alters the structure of a substrate by adding a methyl functional group to its structure, catalyzed by enzymes of the methyltransferase group. S-adenosyl-methionine (SAM) is used as a methyl group donor by methyltransferases. Different forms of methylation mechanisms can result from functional addition on various atoms such as carbon, oxygen, sulfur, and nitrogen. The substrate's lipophilic behavior is enhanced by O- and N-methylation.

The EryG module catalyzes the methylation of the hydroxyl group situated on one of the cyclohexyl groups added during previous post-PKS activities in the erythromycin A pathway [29]. The comparison of erythromycin A and C hexoses demonstrates the addition of a methyl group to generate the final erythromycin A structure (Figure 2.10, highlighted by black boxes).

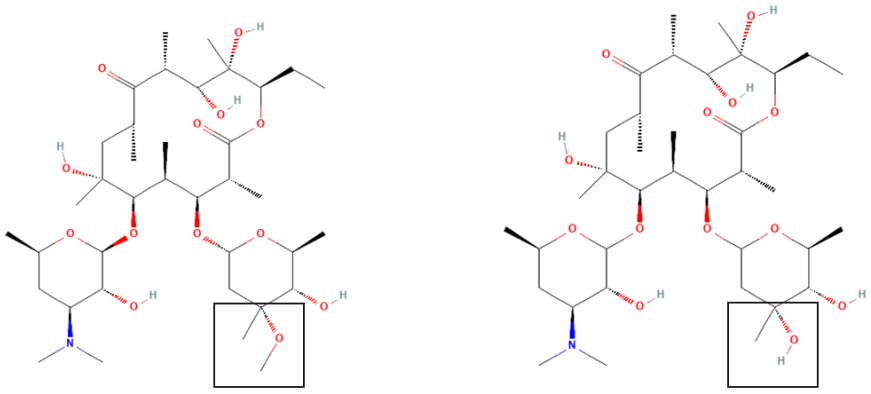


Figure 2.10: **Comparison of the molecular structure of Erythromycin A and C**. The methylation post-assembly-line tailoring process that 6-dEB undergoes to produce erythromycin A is highlighted by the black boxes. A lack of methylation post-PKS tailoring yields erythromycin C.
Both molecular structures were accessed on May 17, 2022 under the PubChem CID 12560 and 441095, respectively.

Multiple other post-tailoring events exist. For further information about cyclization and aromatization catalyzed by oxygenases, one should refer to the review written by Ji *et al.* [33] in 2010. Acylation also widely occurs as a post-assembly line tailoring event. González-Sabín *et al.* [34] wrote in 2011 a review about acylation events in natural products chemistry catalyzed by hydrolases and acyltransferases. Additionally, halogenation of the release chain is common. A regular halogenation event is fluorination, leading to the synthesis of organofluorines. Walker *et al.* wrote a review about the subject in 2014 [35]. Finally, aromatic cycles are also common features in polyketide scaffolds. The review written by Zhan *et al.* [36] in 2009 provides extensive information about the biosynthesis of bacterial aromatic polyketides.

2.3.2 Non-Ribosomal Peptides (NRP) Assembly Lines

2.3.2.1 The canonical organization of Non-Ribosomal Peptide Synthases

A second class of relevant natural products are Non-Ribosomal Peptides (NRPs). As their name suggests, their synthesis differs from the dogmatic RNA translation process handled by ribosomes. During this ribosome-independent synthesis, large multi-modular enzymes called Non-Ribosomal Peptide Synthases (NRPS) catalyse the production of NRPs. The diversity in the incorporated monomers goes far beyond the twenty essential amino acids [37]. The extender units can be proteinogenic or nonproteinogenic amino acids or even other carboxylic groups. Actually, homology models (i.e. sequence alignments) enable the prediction of amino acid substrate for each module [38]. The NRPS assembly line logic is similar to the *cis*-AT PKS one.
As for PKSs, NRPSs are composed of modules responsible for, among others, the initiation, elongation and release of the growing linear chain. We will review the typical machinery and logic for the three types of modules. We will emphasize on the differences and the similarities between NRPSs and PKSs. Figure 2.11 depicts the three core domains that we will encounter in the NRP extension paradigm.

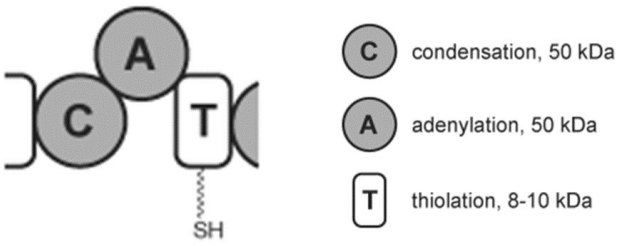


Figure 2.11: **Core domains involved in NRP assembly lines.** The joint activity of condensation (C), adenylation (A), and thiolation (T) domains is requisite for NRPS backbone elongation. The condensation domain is tasked with the creation of amide bonds. The adenylation domain catalyzes amino or aryl acid recruitment onto the NRPS assembly line as well as the acyl transfer to the neighbouring thiolation domain. Thiolation domains, like in PKS, are in charge for covalently tethering substrates to the biosynthetic assembly line.
Figure reprinted from Fischbach *et al.*, 2006 [14].

Initiation module of NRPS

The initiation module is typically made of two domains, responsible for adenylation (A) and thiolation (T). Both domains have similar functions to the AT and ACP domain from PKSs, respectively. The adenylation domain, as ACP in PKSs, acts as a gatekeeper for substrate specificity [39].It selects, activates, and loads the carboxylic acid substrate, which is primarily amino acids, onto the NRPS's subsequent peptidyl carrier protein (PCP) domain. PCP corresponds to the thiolation (T) domain. For the initiation case, the first amino acid is covalently installed on the first thiolation domain, noted T_1.

Elongation module of NRPS

As soon as the initiation of the assembly line is completed, the aminoacyl chain is ready to be extended. To do so, three domains are required, two catalytic ones and a thiolation one. The A domain catalyzes the adenylation (also known as AMPylation) of the soon-to-be-incorporated amino acid. Adenylation is the process in which the amino acid is fused with an adenosine monophosphate (AMP) molecule to yield an aminoacyl-AMP intermediate. Practically, the adenylation (A) domain catalyzes the activation of the carboxyl functional group which triggers the attachment of the amino acid with an adenosine triphosphate (ATP) molecule (Figure 2.12a).

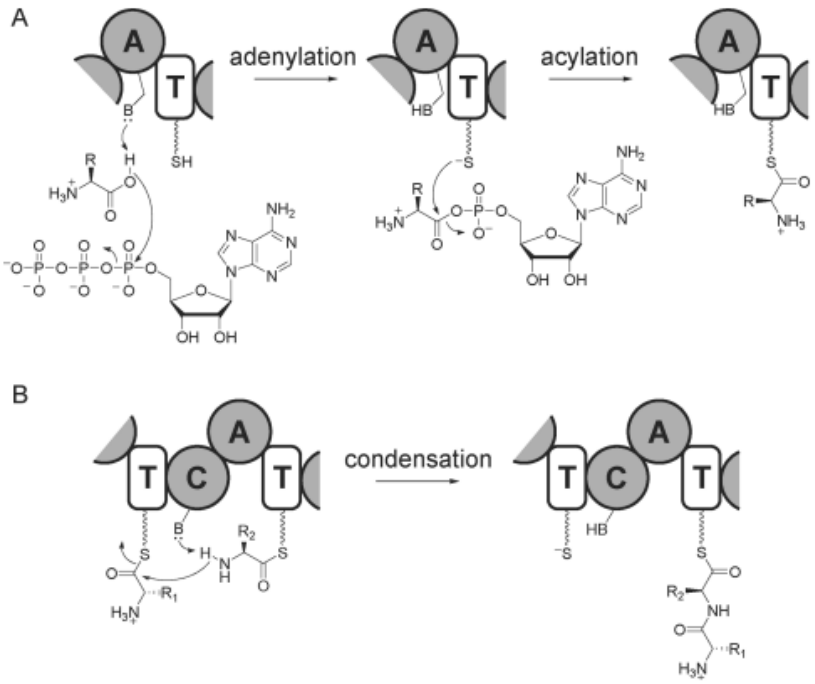


Figure 2.12: **Adenylation and condensation domains activities catalyzed by NRPSs.** (A) Adenylation recruits amino acids, which are then acylated and thus attached to the subsequent thiolation domain. (B) The C domain catalyzes the creation of C-N bonds.
Figure reprinted from Fischbach *et al.*, 2006 [14].

Subsequently, the aminoacyl group derived from the amino acid is tethered to the adjacent thiolation (T) domain, also called a peptidyl carrier protein (PCP) domain in NRPS. This reaction is called an acylation and is still catalyzed by the A domain.
The condensation (C) domain of NRPS, like the KS domain in the PKS framework, catalyzes bond formation. A C-N bond forms between the upstream peptidyl-S-T_{n-1} and the downstream aminoacyl-S-T_n in NRPS. The former is referred to as an electrophile, whereas the latter is referred to as a nucleophile [14].

Release or termination module of NRPS

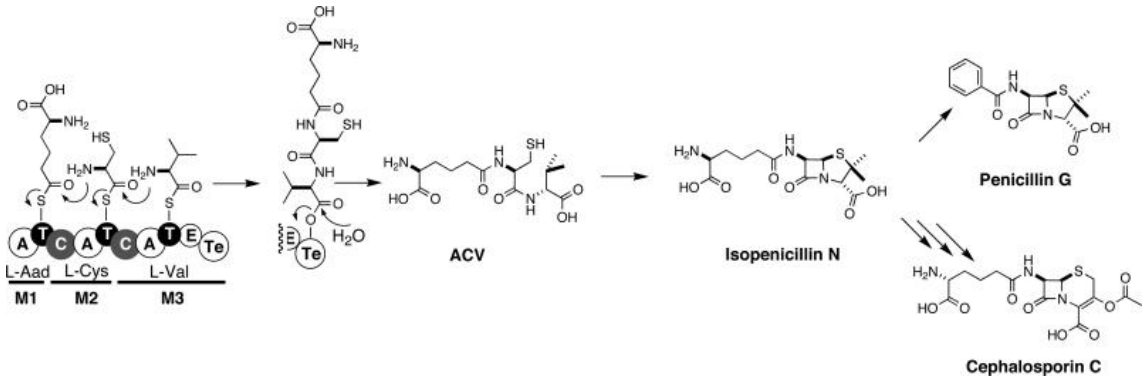


Figure 2.13: **The NRPS ACV synthetase catalyzes the assembly of the ACV tripeptide intermediate.** The release of ACV occurs in a linear manner. Isopenicillin N is a crucial intermediate in the synthesis of penicillin G.
Figure reprinted from Felnagle *et al.*, 2008 [40].

As in PKSs, the termination module in NRPSs also ends with a thioesterase (TE) domain. This domain enables the release of the full-length NRP in a linear or a cyclic way [41]. An illustration of a linear release by a NRPS TE domain is provided on Figure 2.13. Indeed, the released chain is not cycled upon liberation. L-δ-(R-aminoadipoyl)-L-cysteinyl-D-valine (ACV) is the immediate precursor of isopenicillin N, playing a key role is the final biosynthesis of penicillin [42]. It should be observed that the release of ACV from the TE domain of ACV synthase (composed of three catalytic modules namely M1-3) happens in a linear manner.

2.3.2.2 Auxiliary domains in NRP Assembly Line

In the present section, we will review two key auxiliary domains in NRPS assembly lines namely Epimerization (E and C/E) and Cyclization (Cy).

Epimerization

In 1989, Christopher T. Walsh [43] wrote a review in which he compiled observations about D-centers contained in the nascent polypeptide chain during peptidoglycan synthesis. Since then, we know that the presence of D-amino acid residues are a common characteristic of NRP scaffold [14]. In microbial producer cell, D-oriented amino acids are lacking from the pool of building blocks. As a matter of fact, they are produced from L-amino acids using the action of epimerase and racemase enzymes [44]. Both enzymes can catalyze the inter-conversion of carbon centers between their L- and D- enantiomers. L- and D- prefix denote the direction, anticlockwise or clockwise, in which a solution of the molecule rotates a plane-polarized light.
The main mechanism used by NRPS assembly lines implies a 50 kilodaltons epimerization (E) domain. This domain usually acts independently in Gram-positive bacteria and is occasionally conjugated with the condensation (C) domain in Gram-negative bacteria. Both mechanisms are depicted on Figure 2.14. In the preponderant mechanism, the E and C domains act sequentially. Firstly, the L-peptidyl-S-T intermediate is epimerized, catalyzed by the E domain, to form a D-peptidyl-S-T intermediate. Then, the downstream C_{n+1} domain, that is specific for D-donor, performs the incorporation step of the D-amino acid residue (Figure 2.14a). In the occasional organization, the epimerization and the condensation domains work in a "fused" manner. There exists a single domain responsible for both catalytic activities. In term of chirality, this "dual-function" domain as named by Fischbach *et al.* [14], is specific for D-donor and L-acceptor.
Even if the occasional mechanism differs from the preponderant one, the outcome of both mechanisms remain the same (Figure 2.14b).

Cyclization

In addition to the aforementioned chirality-specific condensation (C) domains, there exist C domains that possess a cyclodehydration activity, named Cy. They operate in a three-folded manner. The mechanism starts with the Cy-catalyzed formation of an amine bound between the downstream and upstream intermediates. The latter corresponds to an elongation step. Then, a Cy-catalyzed cyclization takes place during which a five-membered heterocyclic compound is synthesized. A final Cy-catalyzed dehydration step yields a 2-thiazoline (Figure 2.15, X = S), that is a heterocycle containing an imine functional group and a sulfur atom on its ring.

The present review on NRPS and Type I PKS assembly line is not exhaustive. Diversity in domains is an intrinsic characteristic of these synthases. The most common ones have been described here, however, as depicted on Figure 2.3, CLF, MT, Re, Ox, AL, AMT, Hal and Cyp domains are all used in PKS and NRPS biosynthesis but are not covered in this review. To a larger extend, some domains await the scientific community to decipher their characteristics.

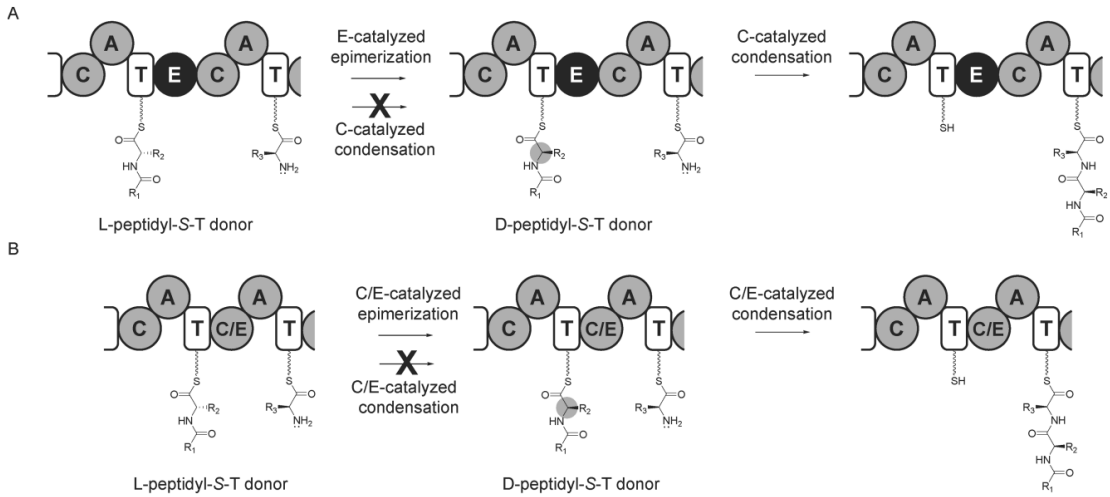


Figure 2.14: **Amino acid epimerization strategies in NRP assemby lines.** (A) The preponderant strategy corresponds to the epimerization of amino acids catalyzed by the E domain from an E -$^{D}C_{L}$ didomain. (B) In other NRPSs, single bifunctional $^{D}C_{L}$/E domains catalyze the epimerization of the L-intermediate to its D-form, which is subsequently selected for condensation.
C/E stands for condensation/epimerization. $^{D}C_{L}$ stands for a condensation domain taking D-intermediate as donor and L-intermediate as output.
Figure reprinted from Fischbach *et al.*, 2006 [14].

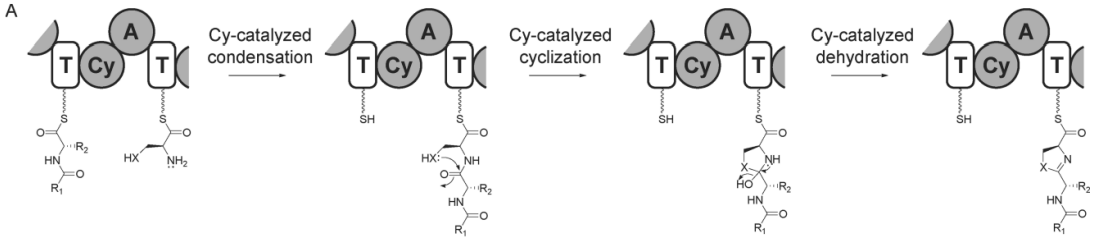


Figure 2.15: **Cy-catalyzed formation of heterocycles in NRP backbones.** The nucleophilic attack of the side chain of cysteine (X = S), serine (X = O), or threonine (X = O) on the previously incorporated carbonyl is catalyzed by Cy domains, followed by the dehydration of the resulting product to form oxazoline or thiazoline functional groups.
Figure reprinted from Fischbach *et al.*, 2006 [14].

2.3.2.3 Post-tailoring enzymes in NRPS

The functions catalyzed by post-tailoring enzymes are widely diverse in NRPS. This topic goes beyond the focus of this master's thesis paper. Thus, it will not be covered. However, one should refer to Walsh *et al.*, 2001 [45] for a extensive review on post-tailoring enzymes in NRPS.

2.4 *trans*-AT mediated assembly lines

Until now, the assembly lines machinery's that have been described function in a *cis*-AT way. We refer to these PKSs as *cis*-AT PKSs. However, there exists another type of PKS assembly lines, named *trans*-AT PKSs. Their differences are mainly driven by a different mechanism of selection and loading catalyzed by AT domains. *cis*-AT PKSs incorporate an AT domain into each elongation module, whereas *trans*-AT PKSs employs freestanding AT domains that are not directly integrated into the canonical PKSs architecture. The present section reviews the *trans*-AT mechanism, its implications and the additional layer of complexity it adds to an already complex biosynthetic organization.

Initially, trans-AT PKS had been disregarded by the scientific community due to their sparseness in typically studied microbes. However, they have now been characterized as enzymes catalyzing the synthesis of a major class of natural products. Indeed, in 2013, O'Brien *et al.* mentioned that trans-AT PKSs account for almost 38% of all bacterial modular PKSs [46, 47]. Additionally, the lack of a direct correlation between the core structure of the synthesized polyketides and the assembly line architecture caused a drawback in the instinctive understanding of *trans*-AT PKSs. Thus, the idea behind this frivolous collinearity rule is not applicable. Since then, *trans*-AT PKSs became a trending topic for their diverse non-canonical architecture, their modularity and their yet-to-discover catalytic features.

It was previously mentioned that *cis*-AT PKSs evolved from Fatty Acid Synthases (FASs). *trans*-AT PKSs, on the other hand, have emerged independently of the canonical design. Nguyen *et al.* suggested in 2008 that *trans*-AT PKSs tend to evolve by horizontal gene transfer and recombination of PKS genetic fragments [48].

2.4.1 *trans*-AT PKSs rely on a single or several standalone enzymes responsible of the selection and the loading of the acyl derivatives

Modules of *trans*-AT PKSs lacks an acyltransferase (AT) domain. As their name suggests, *trans*-AT PKSs rely on a single or a few free-standing AT domains at each elongation step acting in *trans*. The standalone AT catalytic domains are typically encoded by genes clustered in the same biosynthetic gene cluster as the core domains. A comparison of a fictional synthesis between *trans*-AT and *cis*-AT PKSs is depicted on Figure 2.16. The latter highlights the difference of acyltransferase mechanism in both types of assembly line. AT domains, as explained previously, catalyze the selection and loading of the extender's Coenzyme-A thioesters derivatives. The domain acts as a specificity gatekeeper. In both contexts, the function of the AT domain remains the same even if the enzymatic mechanism differs. A single malonyl-specific AT often interacts with the majority of the PKS modules of the *trans*-AT assembly line, resulting in the bulk of the integrated building blocks being malonyl-derived. After chain elongation, these units are further functionalized by reduction, methylation, and other less prevalent modifications [47].

2.4.2 *trans*-AT PKS pathways introduce a new layer of diversity within the natural products landscape

We previously mentioned the canonical *cis*-AT $KS-AT-ACP$ architecture of elongation modules, often fully or partially completed with the catalytic actions of $KR-DH-ER$. In *trans*-AT PKSs, this canonical organization is not present anymore.
In the present section, we will review several mechanisms adding up a layer of structural di-

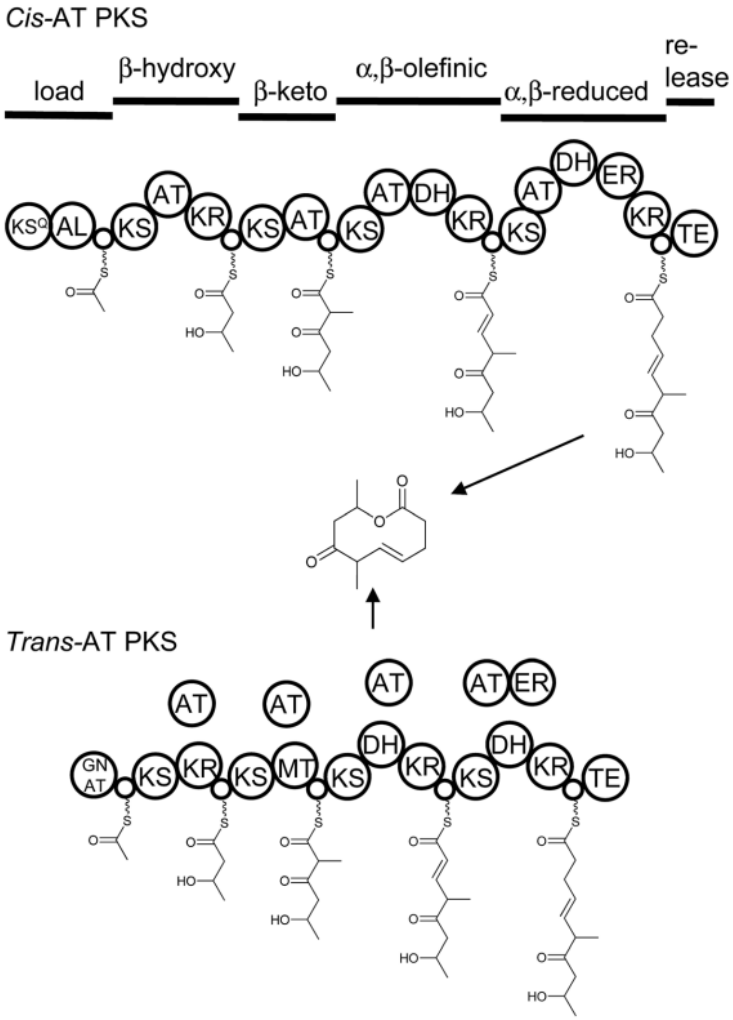


Figure 2.16: **Comparison of the biosynthetic architecture between *cis*-AT PKSs and *trans*-AT PKSs for a fictional polyketide**. ACP domains are shown by tiny unmarked circles. At the top are functional annotations of the resulting intermediate.
Figure reprinted from Helfrich *et al.*, 2016 [47].

versity to the existing natural products landscape. *trans*-AT PKSs are particularly interesting as they deviate from the canonical organization of *cis*-AT PKSs. In 2009, Jörn Piel wrote a review [49] which identified more than 50 canonical module variants. Among these variants are included modules with unusual domains order or unique domains, non-elongating but still modifying modules, domains having a catalytic action across several modules, modules that are split on two peptides thus are apparently missing and non-AT domains functioning in a *trans* manner [49, 47]. A snapshot of the functional and structural diversity is depicted in Appendix A. We will not go into the details of each module singularity. Extensive explanations about several *trans*-AT module singularities can be found in *Biosynthesis of polyketides by trans-AT polyketide synthases* [47], a review written by Helfrich *et al* covering the researches on the topic for the 2009-2015 period.

Another source of diversity in *trans*-AT PKSs is the incorporation of non-canonical extender units, leading to the synthesis of aberrant polyketides. Furthermore, *trans*-AT domains are the playground of NRPS-PKS hybrid assembly lines, ultimately giving rise to an entirely new

class of PK-NRP hybrid structure. Finally, it should be noted that there still exists a load of uncharacterised catalytic domains.

2.4.3 *trans*-AT PKS ketosynthase (KS) domains as core structure predictors

The chemical structure of the compounds arising from an assembly line working in *cis* can be roughly predicted by investigating the order of the catalytic domains. However, this is not the case for *trans* assembly because of their non-canonical architecture. Thus another strategy has been developed based on phylogenetic studies of the ketosynthase (KS) domains. On one hand, it was demonstrated that *cis*-AT PKS KS domains phylogenetically cluster together according to their natural-product specificity. On the other hand, *trans*-AT PKS KS domains phylogenetically relate according to their function. They are said to be specific for the substrate they accept. Indeed, KS domains that phylogenetically cluster together were shown to accept similar intermediates for chain extension. Furthermore, phylogenetically related KS domains were discovered to accept chain intermediates with comparable chemical patterns at the α, β, and γ chain positions [47]. Thus, each KS domain is specific for a certain moiety in *trans*-AT PKSs [48]. By knowing this, the procedure to roughly predict the core structure of polyketides can be outlined as below:

1. Homology search of the KS DNA sequences under study against a reference database i.e. manually annotated KS sequences for their substrate specificity.
2. The phylogenetic cluster a KS sequence belongs to predicts its KS specificity, that is the type of intermediate the KS domain will perform the elongation step on.
3. By knowing the specificity of each KS domain and the open reading frame order of the assembly line, the rough chemical structure of the polyketide chain can be deduced.

Based on the work of Nguyen *et al.* [48] in 2008, a systemic classification of the *trans*-AT KS domains has been initiated. Therefore, a huge diversity of KS specificities had been confirmed. An approximate classification is presented in a recent paper written by Helfrich *et al.* [50]. The ketosynthase (KS) substrate specificity can be separated in height classes: amino acids, β-hydroxyl groups, double bounds, *E*-configured double bounds, *Z*-configured double bounds, non-elongating KS (KS^0), starters and others. Each class correlates with a polyketide motif. This classification can be further refined in more specific sub-classes, amounting to around 100 possible KS specificities.

The investigation of the KS domains is at the heart of several bioinformatics software used in the exploration of natural products such as PKSs. A variety of current bioinformatics tools rely on the prediction of the substrate specificity of KS domains to understand the nascent polyketide structure. One tool in question, transPACT [50] uses phylogenetic placement on a reference phylogeny to predict the substrate specificity of KS sequences. Then, it outputs a dendrogram representation of the query PKSs. Additionally, the transATor [51] tool attempts to predict the rough polyketide structure from the KS specificities within their *trans*-AT PKSs. These two software will be further discussed in the next section, alongside with their architectures, their implications and their limitations.

2.5 A snapshot of the computational toolbox available for the investigation of *trans*-AT Polyketide Synthases

After having reviewed the biochemistry of assembly line enzymology, we now head to the computational aspects of it. Bioinformatics work in the field of natural products has been initiated by the work of the laboratories of Piel, Helfrich and Medena. In the present section, we emphasise on three software that is antiSMASH, transPACT and transATor. The first one is most popular tool in the natural products landscape. The two others are niche tools that have been developed specifically for the investigation of *trans*-AT PKSs. For the past decade, we hopefully encompassed the continuous development of several bioinformatics tools. More than ever, the need for efficient, reliable and accurate software is required in the field of natural products discovery.

2.5.1 antiSMASH

antiSMASH stands for antibiotics & Secondary Metabolite Analysis Shell. It was developed in 2011 by *Medema et al.* [52] in order to provide an efficient and reliable tool to predict biosynthetic gene clusters. As the number of bacterial sequences was booming, a platform for genome mining of natural products or secondary metabolites was demanded. Nowadays, the antiSMASH database v3 contains about 3,000,000 domain sequences.
The antiSMASH workflow identifies gene clusters encoding for enzymes catalyzing the synthesis of natural products from their genome sequences, using profile Hidden Markov Models (pHMMs). The details of pHMMs go beyond the scope of this literature review [5]. One should keep in mind that pHHMs capture the profile of signature genes. Query sequences can then be aligned to the profile in order to identify signature genes. Ideally, the input sequences are annotated nucleotide sequences. However, if no annotations are available, antiSMASH automatically generates gene annotation using the Glimmer3 algorithm [54]. Glimmer3 enables the prediction of coding regions using Markov Models.

Then, based on the gene cluster identification, antiSMASH performs four different analysis. In the scope of this paper, we emphasise on the domain analysis of NRPSs and PKSs. Using pHMMs once again, antiSMASH attempts to annotate the catalytic domains included in the genetic information of query sequences. Finally, the outcome of the predictions are visualized on a single web-page.

2.5.2 transPACT

transPACT stands for *trans*-AT PKS Annotation and Comparison Tool. The development of transPACT began in 2015 but was only recently published by Helfrich *et al.* [50] in 2021. Two applications are supported by the transPACT platform:

1. The prediction of the substrate specificity of KS sequences
2. The identification of conserved modules blocks among the PKSs under study

The transPACT platform has been extensively used in the context of the research topic of this paper, as discussed in the Chapter 3. The architecture and the applications of the software will

[5] Extensive information about profile HMMs are available in a review written by Sean R Eddy in 2011 [53].

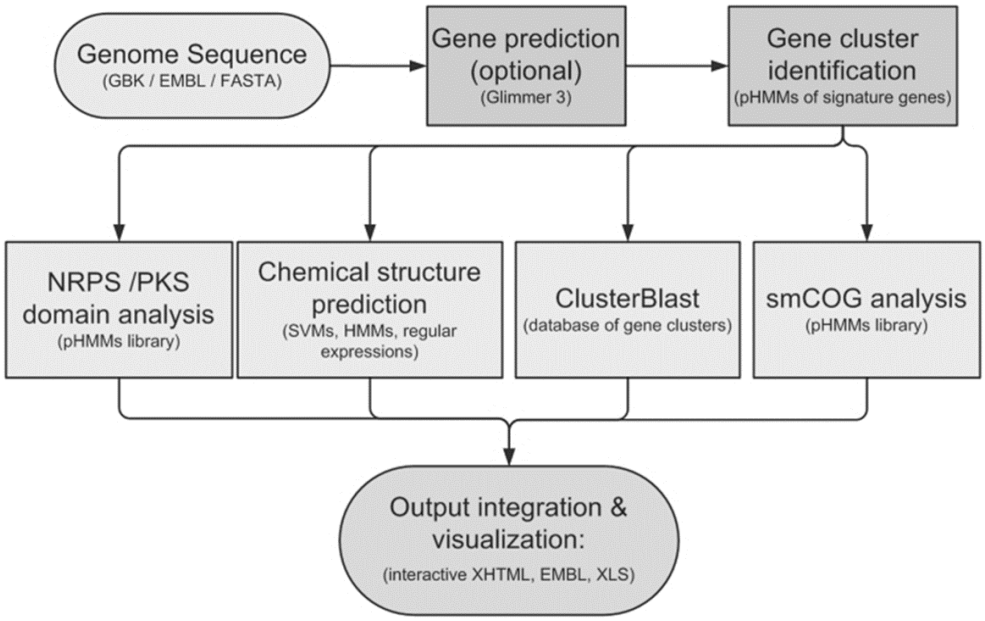


Figure 2.17: **Workflow of the antiSMASH comprehensive analysis tool for secondary metabolites.** The antiSMASH comprehensive tool: https://antismash.secondarymetabolites.org/ and the antiSMASH database: https://antismash-db.secondarymetabolites.org/.
Figure reprinted and adapted from Medema *et al.*, 2011 [52].

be detailed in this literature. In the next section, its integration within a high performance computer[6], its adaptations to a global genome mining approach and its limitations will be discussed.

2.5.2.1 A substrate specificity annotation tool of *trans*-AT ketosynthase (KS) domains

The first application of the transPACT platform is the annotation of KS sequences with their substrate specificities. We previously mentioned in section 2.4.3 that KS sequences rule the specificity of the incorporated substrates in *trans*-AT assembly lines. Thus, the automated prediction of such substrate specificity annotations is a major driver in the development of novel bioinformatics tools.

First, a reference phylogeny must be constructed and given as input to the search algorithm. An illustration of this step is shown in Figure 2.18a. Helfrich *et al.* [50] generated a maximum-likelihood reference tree using a set of sequences from 647 KS domains, dispersed over 49 well characterized *trans*-AT PKSs reported by Helfrich and Piel, 2016 [47]. The function of each clade composed of iso-functional KS sequences was manually annotated based on previous literature. Then, the pplacer algorithm classifies each query sequence based on its phylogenetic relationships on reference tree. The pplacer software package implements traditional likelihood-based phylogenetic classification in a linear computational time [55]. This method can therefore be scaled for large data sets such as in a genome mining context. Finally, the substrate specificity of each query KS sequence is inferred, if the KS domain is monophyletic, from its iso-functional clade membership.

[6]The Vlaamse Super Computer (VSC), https://www.vscentrum.be/

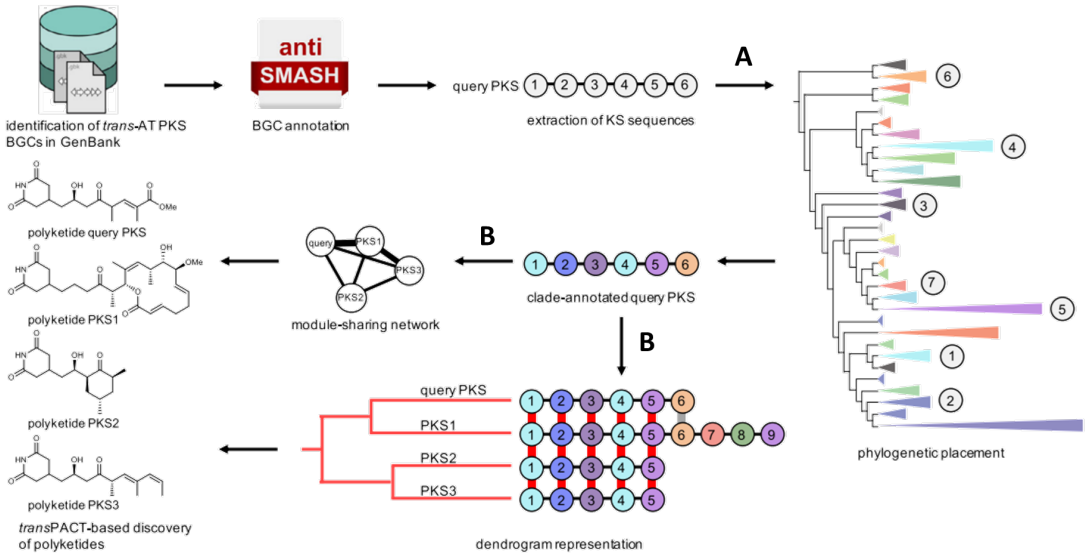


Figure 2.18: **Workflow of the transPACT software.** (A) This step corresponds to the phylogenetic placement of KS sequences to a reference phylogeny in order to predict their substrate activities. (B) This step corresponds to phylogenetic comparison of the PKSs sharing module or domain blocks using network and dendrogram representations.
Figure reprinted and adapted from Helfrich *et al.*, 2021 [50].

2.5.2.2 A phylogenetic comparison tool for *trans*-AT PKSs sharing module blocks

Now that all KS sequences have been tagged with their functions, we may visualize the similarity relationship between the PKSs under study. This step is sketched in Figure 2.18b. The transPACT comparison tool enables a network and a dendrogram representation. In the network visualization, the nodes correspond to the PKSs and the length of the edges is proportional to the number of shared PKS KS domains, also referred as a shared module block. A module is considered to be shared with another PKS if it possesses the same KS substrate specificity. In the dendrogram visualization, the architectural relationships are outlined according to a distance measure calculated as a combination of the following items:

1. Jaccard Index of shared KS domains

   In the present context, the Jaccard Index is a similarity metric determined as the proportion of common KS domains between PKS_1 and PKS_2 compared to the total number of KS domains in PKS_1 or PKS_2. It can be calculated as follow [56]:

   $$J_{PKS_{1,2}} = \frac{|N(PKS_1) \cap N(PKS_2)|}{|N(PKS_1) \cup N(PKS_2)|}$$

   where $N(PKS_x)$ corresponds to the number of KS domains in the x-indexed PKS.

2. Differences in the copy number of shared KS domains

   Bacterial genomes may lodge a single copy of a gene, or several copies of it. This variation arises from a duplication or a deletion event [57]. The difference in the copy number of shared KS domains is a driver in the calculation of the distance separating two PKSs. This is calculated using the Duality Diagram Similarity (DSS) Index, expressing the difference in sequence per domain. If both PKSs contain the same KS domain and their sequences

are similar, the difference is set to zero. However, in a pair of PKSs, the difference is set to one if a KS domain is only contained in one of the two PKSs.

3. Synteny conservation

   Synteny is hereby referred as the conservation of the module order between PKSs. Naturally, PKSs sharing the same order for a number of KS domains should be considered more closely related in a dendrogram representation. Two metrics are used to gauge the synteny conservation between PKSs, namely the Goodman-Kruskal gamma index and the Tanimoto Adjancy Index.
   The Goodman and Kruskal gamma index measures to what extend the order of a series of KS domains is similar to the order of another series of KS domains. It is a metric for ordinal association and is computed as follow:

$$GK = \frac{1+\gamma}{2}$$

   where $\gamma = \frac{|N_R - N_S|}{|N_R + N_S|}$, $N_S \equiv$ number of pairs of KS domains ranked in the same order for two PKSs, $N_R \equiv$ number of pairs of KS domains ranked in the reversed order for two PKSs.
   The Tanimoto Similarity corresponds to the rate of the number of common elements between two series over the total number of elements in the two series [58]. However, in transPACT, it has been implemented to compute the similarity of adjacent KS domains between two PKSs in such a way that each element of the series is a pair of adjacent KS domains. The Tanimoto Adjacency is then calculated as follow:

$$AI = \frac{|pairs(PKS1) \cap pairs(PKS2)|}{|pairs(PKS1) \cup pairs(PKS2)|}$$

The combination of these three drivers for similarity dictates the estimation of the distance matrix. Each driver is assigned a weight by the user to compute the similarity score between two PKSs as follow:

$$S = (Jaccardw * Jaccard) + (DDSw * DDS) + (GKw * GK) + (AI * AIw)$$

The distance is then simply calculated as below:

$$D = 1 - S$$

By iterating over the PKSs pairs, the distance matrix is constructed. Finally, it is straightforward to generate a dendrogram representation from a distance matrix in which each leaf corresponds to a PKS and each branch length to its corresponding distance.

2.5.2.3 Performance of transPACT

The classification performance of transPACT was assessed on a test set made of 12 well-characterized *trans*-AT PKSs. They encode the following polyketides: Mycalamide A, Peloruside A, Pateamine A, Gladiofungin A, Lobatamide A, Alpiniamide A, Lacunalides, Lagriamide, Macrobrevin, Ripostatin, Pyxipyrrolone A and Phthoxazolin. The substrate specificity of 81.2% of the 133 KS domains contained in the 12 aforementioned polyketides was correctly predicted.

It should be noted that the domains for which transPACT was not able to assign any substrate specificity were not considered as misclassifications. Practically, they are annotated as part of the "not assigned" class. If one would reckon these as mistakes, the classification accuracy would drop to 70.1%.

2.5.3 transATor: *trans*-AT PKS Polyketide Predictor

The transATor software, developed by the Piel Lab *et al.* [51], is a molecular structure *de-novo* prediction tool for *trans*-AT polyketide synthase products. It stands for *trans*-AT PKS polyketide predictor.

The specificity of KS domains for their incoming substrate is predicted in a similar fashion as the transPACT software, that is using the phylogenetic placement of query KS sequences on a cladogram of *trans*-AT KS domains of reference. However, the set of characterized *trans*-AT PKSs used in transPACT is not the same as in transATor. Thus, clades classification may differ from one software to another.
The core structure is initially generated based on the predicted KS substrate specificities. Then, further modifications are translated on the nascent polyketide according to the catalytic characteristics of the domains present in *trans*-AT the assembly line.

Chapter 3

Investigation of *trans*-AT assembly lines in *Massilia flava* and *Caballeronia udeis*

3.1 Methods

The presently described master's thesis blended methodologies from bacterial genome engineering and computational biology.

3.1.1 Bacterial genome engineering

Ultimately, we want to identify the chemical compounds biosynthesized by the two assembly lines encoded by PKSs from *C. udeis* and *M. flava*. To do so, the experimental workflow outlined in Figure 3.1 has been designed. The mass spectra differences between the metabolomes tied to unaltered bacterial genomes and to genetically-engineered (i.e. altered) bacterial genomes should highlight the lack of certain biosynthetic capabilities. On one side, we investigated the cultivation of both bacteria on different poor medium. Each poor medium corresponded to a single carbohydrates source, namely glucose, fructose, glycerol and ribose. Cultivation without any carbohydrates source was also carried on as a control. The metabolome produced by this bacterial cultivate represents the unaltered one. On the other hand, the genomes of *C. udeis* and *M. flava* have been modified to disrupt the gene encoding for the PKSs under study. The metabolome produced by this bacterial cultivate therefore represents the altered one. The below-described methods have been performed for both bacteria, *C. udeis* and *M. flava*.

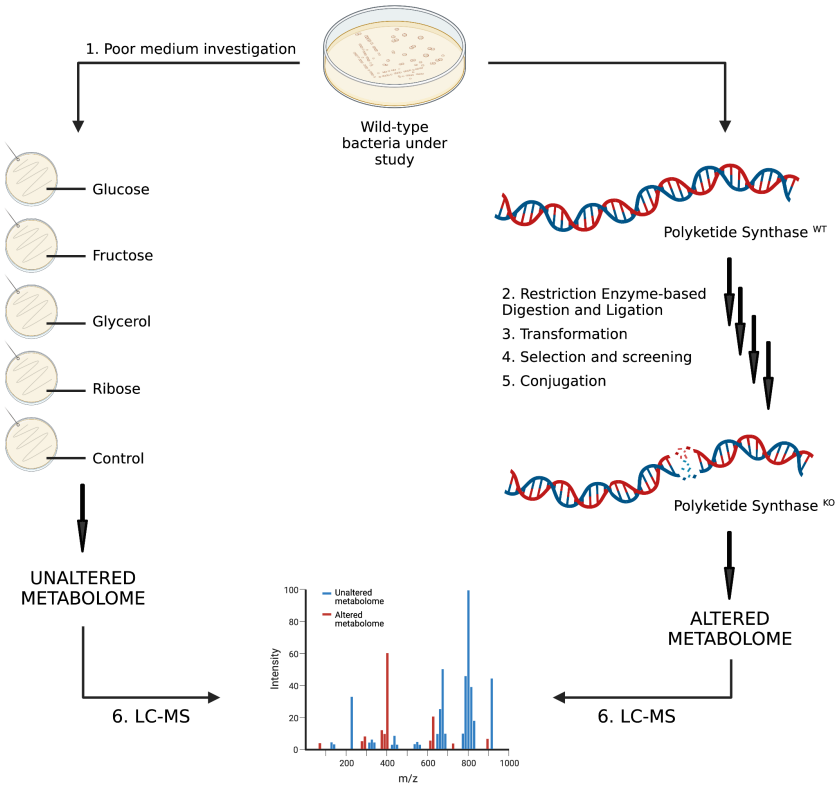


Figure 3.1: **Workflow of the genetic engineering investigation of the assembly lines under study from *Caballeronia udeis* and *Massilia flava***. On the left side is illustrated the bacterial cultivation on five different poor media. On the right side is depicted the disruption of targeted genes in order to disable the operation of the assembly lines. Created with BioRender.com.

3.1.1.1 Cultivation

Caballeronia udeis

C. udeis is a Gram-negative aerobic soil bacterium from the family Burkholderiaceae and order Burkholderiales, originally sampled in Kansas, USA in 2000 [59]. Since the bacterium is mesophilic, it optimally grows between 20°C and 45°C. Bacteria have been cultivated at 28°C on Trypticase Soy Agar (TSA) as a solid rich medium at pH 6-7.

Massilia flava

M. flava is Gram-negative aerobic soil bacterium from the family Oxalobacteraceae and order Burkholderiales, originally sampled in Ningxia Province, China [60] in 2011. The bacterium is mesophilic. Bacteria have been cultivated at 28°C on R2A Agar. It is a solid low nutrient medium suited for slow-growing species.

The carbon source that suits best each bacterium was investigated by comparing the growth on different carbon sources. Bacteria were cultivated on BSM medium completed with ribose, glucose, fructose or glycerol. Normal growth was monitored by a control corresponding to unaltered BSM medium.

3.1.1.2 Recombinant plasmids construction

The first step of the bacterial genome engineering corresponds to the construction of recombinant DNA from pGPI plasmids and bacterial genes targeted for disruption. Approximately 1000 base pairs (bp) long bacterial DNA inserts, located in GO485_00250 for *M. flava* and in AWB69_2900 for *C. udeis*, are selected for PCR amplification.

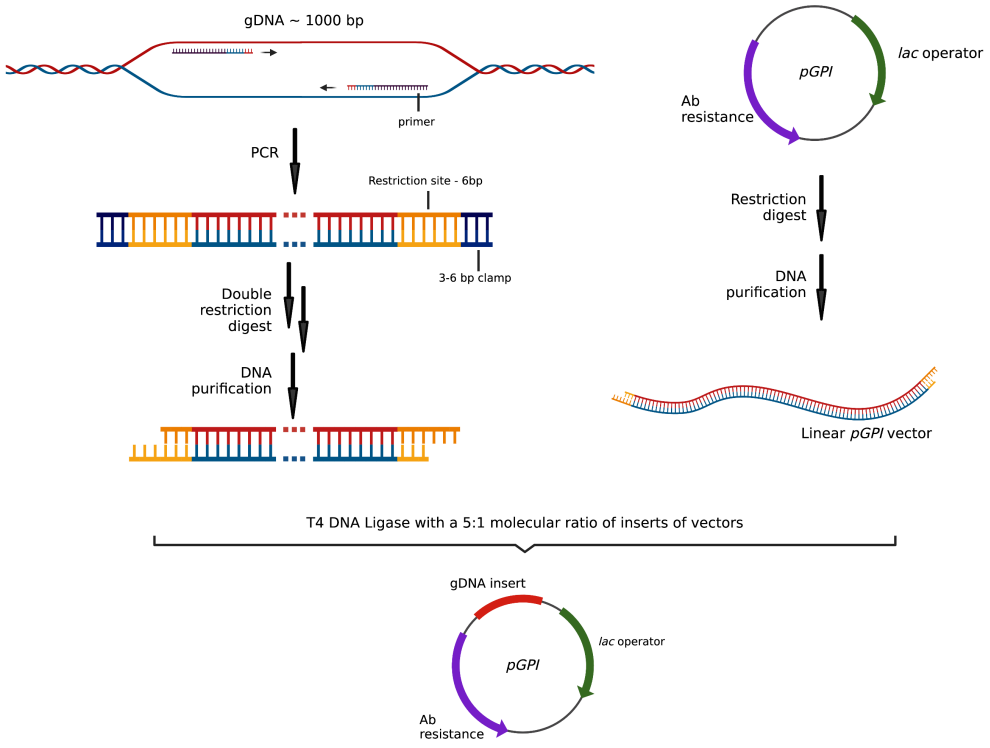


Figure 3.2: **Experimental workflow for the construction of recombinant pGPI plasmids**, inserted with bacterial DNA inserts. The gene under disruption is GO485_00250 for *M. flava* and AWB69_2900 for *C. udeis*.
Created with BioRender.com.

PCR primers were composed of a 18 bp unique recognition sequence, a 6bp restriction site for XbaI or EcoRI enzymes and a 3-6 bp clamp. The yielded 1000 bp gDNA inserts possessed blunt ends. For further ligation, blunt ends were converted into sticky ends through a double digestion. The DNA inserts were simultaneously digested with XbaI and EcoRI at their corresponding restriction sites. Optimal reaction conditions for the double digest were calculated by the DoubleDigest Calculator by ThermoFischer Scientific[1].

In parallel, pGPI plasmids underwent two single digestions with, respectively, XbaI and EcoRI restriction enzymes. The restriction sites for both restriction enzymes are located in the *lac* operator, which hereby function as a multiple cloning site. The digestion yielded linear pGPI plasmids with sticky ends.

As the sticky ends of the gDNA inserts match the sticky ends of the linear pGPI plasmids, the ligation of both DNA fragments happened, catalyzed by T4 DNA Ligase. To ensure an efficient DNA ligation, the DNA mixture contained a 5:1 molecular ratio of inserts to plasmids. By measuring the mass concentration (using the NanoDrop Spectrophotometer) of both the inserts and the plasmids mixtures and by knowing the length (in bp) ratio of both DNA fragments, we verified that this ratio was respected.

3.1.1.3 Transformation of chemically competent *E. coli* cells

The ligation mixture arising from the previous step contained both native and recombinant pGPI plasmids. Empty plasmids were also observed as a result of primer self-annealing. In order to

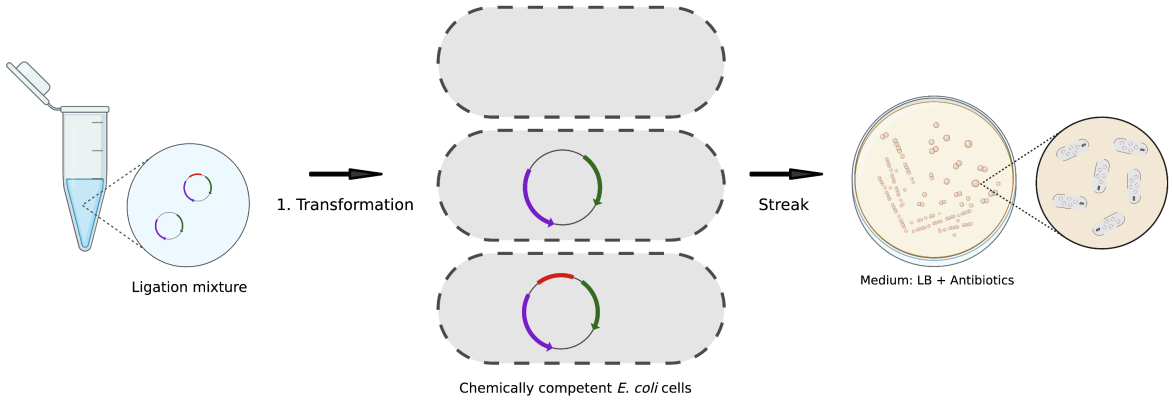


Figure 3.3: **Experimental workflow for the transformation** of chemically competent *E. coli* cells (i.e. SY327 strain) using the ligation mixture that contains native and recombinant pGPI plasmids. The red gene corresponds to the gDNA insert from *C. udeis* or *M. flava*. The antibiotic is Trimethoprim.
Created with BioRender.com.

maintain the plasmids, they need to be introduced into bacterial host cells, which are chemically competent *E. coli* cells. In such cells, the entry of plasmids is facilitated at a high temperature of 42°C due to a porous cell membrane. Competent cells that were previously prepared by members of the laboratory were used. This step, called transformation, is illustrated in Figure 3.3. Thus, three transformation outcomes are expected. When the transformation was successful, the introduced plasmid can either be an empty or a recombinant one. When the transformation was not

[1]https://www.thermofisher.com/be/en/home/brands/thermo-scientific/molecular-biology/thermo-scientific-restriction-modifying-enzymes/restriction-enzymes-thermo-scientific/double-digest-calculator-thermo-scientific.html, accessed on June 06, 2022.

successful, the chemically competent *E. coli* cells did not contain any exogenous genetic material.

Then, the transformed competent *E. coli* cells were streaked on LB + Trimethoprim (Tp) plates. Given their antibiotics resistance conferred by their plasmids, only the transformed bacterial cells survived. Since we were only interested in the cells transformed using the recombinant plasmids, a selection step was undertaken to discard the cells transformed using the native plasmids.

3.1.1.4 Pre-screening, sequencing and selection

This pre-screening, sequencing and selection step is illustrated on Figure 3.4.
From the previously streaked plates, 32 bacterial colonies were suspended into 100 μL wells of a 96 wells plate. After two hours and for each suspension, a 2.5 μL aliquot was sampled for colony PCR. In parallel, the remaining 97.5 μL of each suspension was completed with a glycerol solution. Each glycerol stock was then labelled and kept at -20°C for further usage. Within each well of the plate, a colony PCR was undertaken. The competent *E. coli* cells were lysed during the initial heating step of the PCR reaction. Primers were designed in a backbone-

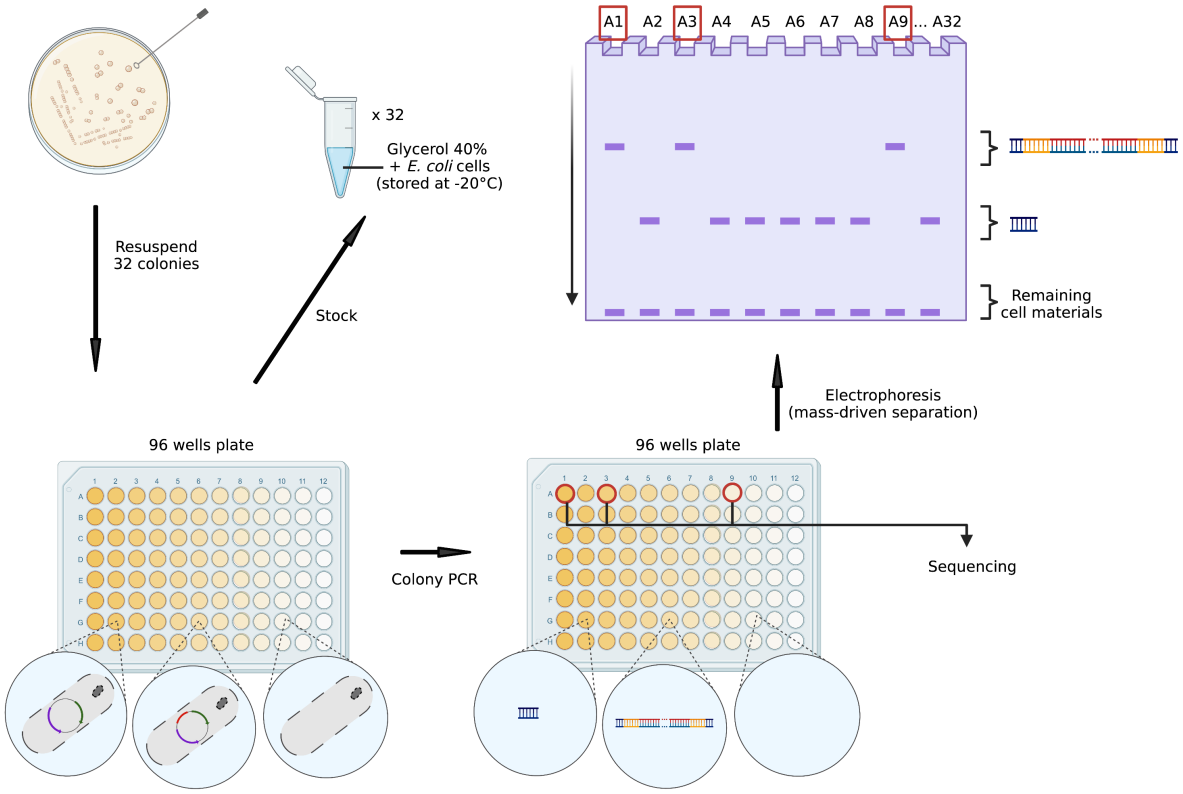


Figure 3.4: **Experimental workflow of the pre-screening, sequencing and selection step.** Red boxes and circles correspond to the selected wells, that is the ones containing recombinant plasmids.
Created with BioRender.com.

specific fashion that annealed to the flanks of the DNA insert. The amplified genetic material present in each well was then transferred into electrophoresis wells. Each DNA material then travels through the agarose gel according to its mass. Therefore, the electrophoresis columns exhibiting a band that reflects a higher mass correspond to wells containing the recombinant plasmids. While the columns exhibiting a band that reflects a lower mass correspond to the wells containing the native plasmids. This was further confirmed by comparing to a ladder of

known fragment length. By matching the electrophoresis columns to the plate wells, we were able to send the recombinant plasmids for sequencing (depicted by red boxes and circles on Figure 3.4). This electrophoresis acted as a pre-screening step in order to reduce the amount of genetic material to sequence. However, sequencing acted as the screening step to ultimately discriminate the wells containing the native plasmids from the ones containing the recombinant plasmids.

Finally, we matched the relevant wells to our glycerol stocks. We thus selected the relevant glycerol stocks for further applications, that is the ones containing the transformed competent *E. coli* cells using bacterial gDNA inserts from PKSs genes from *C. udeis* or *M. flava*.

3.1.1.5 Conjugation

With the plasmids in hand, efforts turned to inserting the homologous region into the wild type genomes using conjugation. This was attempted using a tri-parental model, whereby three protagonists are required, namely HBI01 *E. coli* cells as helpers, transformed SY327 *E. coli* competent cells as donors and *C. udeis* or *M. flava* cells as recipients. Actually, two conjugations occur in such a set-up, as visually outlined by the two steps in Figure 3.5. However, experimentally, both conjugations happened on a single filter paper sat on a plate. The first

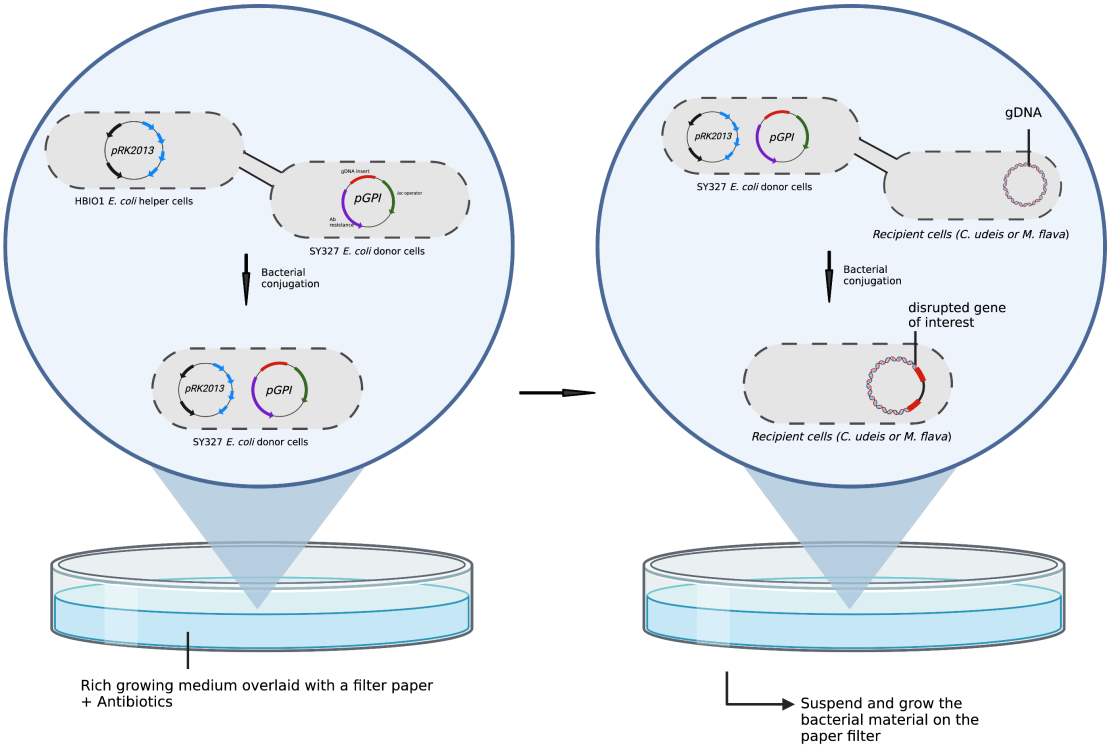


Figure 3.5: **Experimental workflow of the conjugation step.** The conjugation step is two-folded, corresponding to a tri-parental mating design.
Created with BioRender.com.

conjugation i.e. genetic transfer occurred between the helper and the donor cells. Since SY327 *E. coli* competent cells are not able to conjugate and transfer DNA, helper cells were required. The pRK2013 conjugative plasmid from the helper strain was transferred to the donor cells. Donor cells containing both pRK2013 and transformed pGPI plasmids were given conjugation

capabilities.

The second conjugation step occurred between the donor cells and the recipient cells. With the recipient cells containing the recombinant plasmids, homologous recombination between the gDNA insert and the PKSs genes takes place. The latter causes the insertion of an approx. 5000 bp DNA fragment into a biosynthetic gene. It should therefore disrupt the correct translation of this gene and ultimately disrupt the enzymatic activities of the whole polyketide synthase.

Experimentally, the filter paper allows all bacteria to grow on non-selective solid media. The antibiotic was removed from the original medium of both *E. coli* cells mixture by spinning it down and re-suspending the cells in LB medium without antibiotics. This step was repeated twice. Then, in an eppendorf tube, recipient cells, HBIO1 *E. coli* helper cells and SY327 *E. coli* donor cells were mixed together in a 2:1:1 volume ratio. 100 μL of this mixture was sampled and spread onto a sterile filter sat on a plate containing R2A or TSA without antibiotics.
For the conjugation of *M. flava*, bacteria were cultivated on R2A + 400 $\mu g/\mu L$ trimethoprim at 28°C for 24 hours. For the conjugation of *C. udeis*, bacteria were cultivated on TSA + 150 $\mu g/\mu L$ trimethoprim + 10 $\mu g/\mu L$ triclosan at 28°C for 24 hours.

The day after, the results of scraping the filter papers were re-suspended in 1 mL of *Milli-Q* water and cultivated on R2A or TSA with trimethoprim. The 1 mL suspensions were split in three rich medium + antibiotics cultivations as follow: 10, 100 and 800 μL.

3.1.2 Computational Biology

For this master's thesis, the bioinformatics efforts were focused on the understanding, the integration and the application of the *trans*PACT software for the investigation of *trans*-AT biosynthetic pathways.

3.1.2.1 Data acquisition

Caballeronia udeis

The reference genome of the LMG27134 *Caballeronia udeis*[2] strain is accessible under the RefSeq:GCF_001544555.2 accession number. In 2016, the Laboratory for Microbiology of the UGent submitted a whole genome shotgun sequencing of *C. udeis* under the FCOK02000000 WGS accesion number. On June 13, 2022, the reference genome consisted of 242 contigs amounting to a total of 10,051,569 base pairs, from FCOK02000001 to FCOK02000242. The given Locus Tag Prefix is AWB69.

Massilia flava

The reference genome of the DSM26639 *Massilia flava*[3] strain is accessible on the RefSeq Database under the RefSeq:GCF_009789595.1 accession number. In 2019, a group from the University of Tuebingen submitted the complete genome assembly of *M. flava*. On June 13, 2022, the reference genome consisted of 6,025 genes amounting to a total of 6,922,031 base pairs. The given Locus Tag Prefix is GO485.

The *trans*-AT biosynthetic gene clusters were identified from the whole genome sequences using antiSMASH v5.1.2. Then, Angus Weir, Postdoctoral researcher in the Laboratory for Biomolecular Discovery, reconstructed the assembly lines from the whole genome sequence and from the antiSMASH v5.1.2 PKS domain predictions of both bacteria. The latter are shown on Figure 3.6.

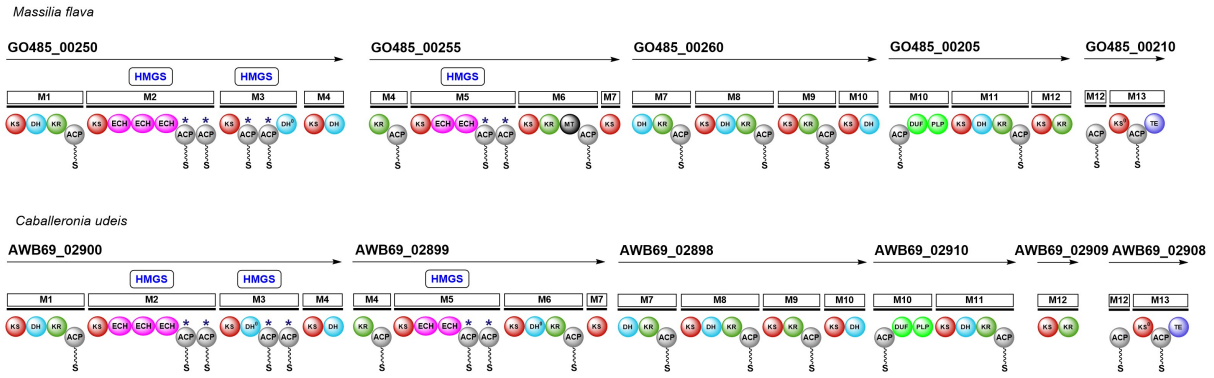


Figure 3.6: **Predictions of the biosynthetic pathways encoded by *trans*-AT PKSs in *C. udeis* and *M. flava*.** HMGS: 3-hydroxy-3-methylglutaryl synthase (branching enzyme). The rest of the domain acronyms can be translated using Figure 2.3.
Figure created by Angus Weir.

In addition to the two aforementioned gene clusters under study, the KS sequences of seven related *trans*-AT PKSs and of the oximidine pathway were also fetched as input data for transPACT. They can be found under the following Taxonomic names and genome accession numbers (ac-

[2]Synonym: *Burkholderia udeis*, NCBI Taxonomy ID: 123866
[3]NCBI Taxonomy ID: 871742

cessed lastly on June 18, 2022) in NCBI GenBank. It should be noted that the letters in brackets correspond to their declared locus tag.

- *Aquimarina* sp. AU58, GCA_900312745.1 (DK441)
- *Burkholderia pseudomallei* MSHR4299, GCA_000774465.1 (XB90)
- *Burkholderia pseudomallei* strain BDE, GCA_000757095.1 (DO63)
- *Lysobacter enzymogenes*, GCA_002355975.1 (BLU84)
- *Massilia* sp. Root335, GCA_001425685.1 (ASC93)
- *Massilia violaceinigra* B2, GCA_002752675.1 (CR152)
- *Sporocytophaga myxococcoides*, GCA_000426725.1 (K350)

These were found serendipitously with cluster BLAST. From their antiSMASH gene clusters and domains prediction, their putative biosynthetic pathways have been drafted (see Appendix C). Overall, none of the seven PKSs under study have been published yet.

Finally, the most recent antiSMASH ClusterBlast database[4] was employed in the transPACT genome mining approach.

3.1.2.2 antiSMASH

The 5.1.2 version of antiSMASH was used, with a detection strictness set to relaxed. The whole genome sequences of both bacteria, under NZ_FCOK00000000.2 for *C. udeis* and under CP046904 for *M. flava*, were used as inputs.

3.1.2.3 transPACT

The *trans*-AT PKS Annotation and Comparison Tool (*trans*PACT) has already been introduced in the literature review. A general introduction has been outlined. The present section describes the specific application of transPACT for this research topic, as illustrated in Figure 3.7. As transPACT is embedded under a Linux platform, the computational efforts were performed under a Linux Operating System. The Vlaamse Super Computer (VSC)[5] was used. A Linux platform running on an Oracle VM VirtualBox was also used under a Mac OS.

The software takes as input a FASTA file containing the protein sequences of KS domains. The query KS sequences were merged to the KS sequences of reference in a single FASTA file. It should be noted that the query sequences correspond to the KS sequences of the *trans*-AT PKSs under study. The KS sequences of reference are manually annotated sequences used as template for phylogenetic placement. The set of KS sequences was then aligned using MUSCLE. Additionally, the headers of the KS sequences were parsed (see Figure 3.7a) in such a way that they meet a consistent and minimal naming structure in the form of *<PKS name>|<KS number>*. The choice of such a naming structure will be further discussed in Section 3.3. These aligned and renamed KS sequences then undergo two data processing pipelines.

On the one hand, each KS sequence was annotated with its substrate specificity by the transPACT annotation tool (Figure 3.7b). The latter yielded an annotation file matching each KS name with its phylogenetic clade and a clade description. Each clade corresponds to a specific substrate type (e.g. molecules harbouring *(E)*-double-bonds). Then, only the KS names and their assigned specificity were kept and sent to a TSV (Tab Separated Values) file. The latter was subsequently

[4]accessible at https://dl.secondarymetabolites.org/releases/clusterblast/clusterblast$_2$0201113.*tar.xz*
[5]https://www.vscentrum.be/, accessed on May 17, 2022.

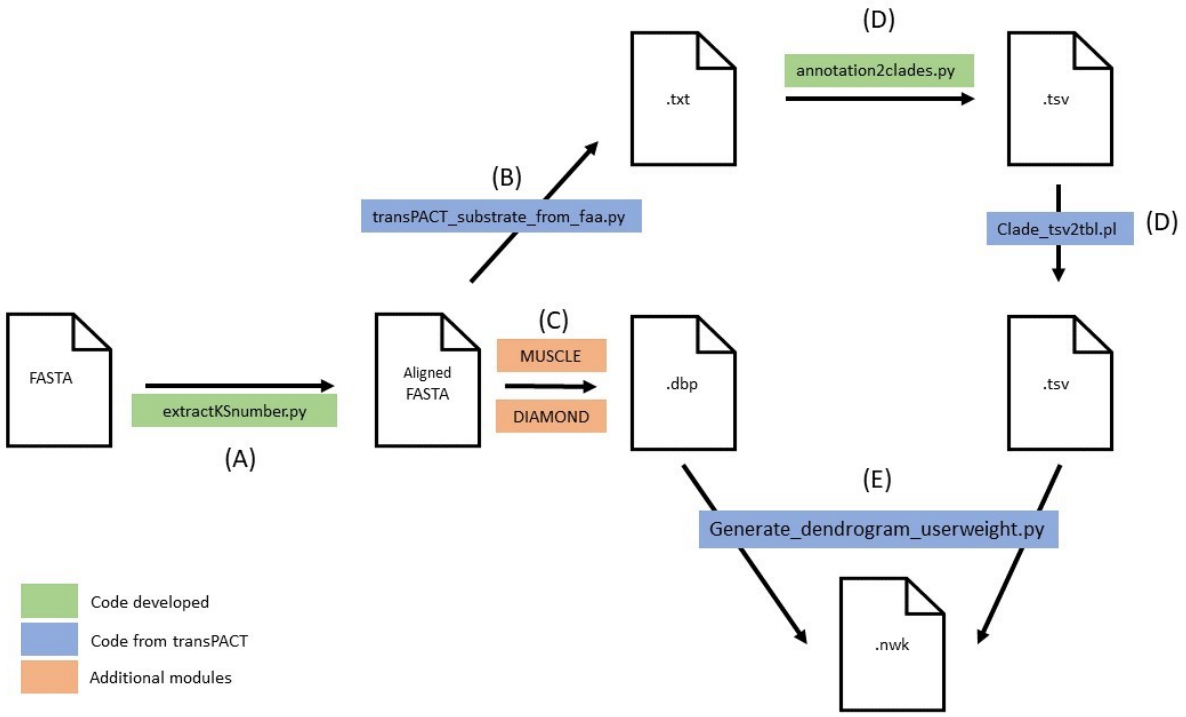


Figure 3.7: **Files continuum of the transPACT software.** (A) Alignment of the KS sequences and extraction of the KS numbers. (B) Annotation of KS sequences with their substrate specificity. (C) Creation of all vs. all diamond table. (D) Processing of the annotation file. (E) Generation of dendrogram visualization.

converted by a Perl script from a long data format to a wide data format. Practically, the final TSV (Tab Separated Values) file contains ordered KS numbers as columns, PKS names as rows and substrate specificity annotations as values (Figure 3.7d).

On the other hand, an all vs. all table was created (Figure 3.7c). First, the KS sequences were aligned using MUSCLE. Then, a database was generated from the aligned KS sequences. Finally, DIAMOND searched through this database using the same KS sequences as query. By searching each protein sequence against the whole set of protein sequence, DIAMOND identified the best hits among the set of sequences.

Finally, similarities between PKSs under study were visualized using a dendrogram representation. The corresponding script (Figure 3.7e) took the all vs. all sequences search and the final annotation files as inputs. It yielded a Newick tree file that can be inspected using e.g. ITOL[6]. Additionally, a Perl script generated the ITOL domain annotations representing the substrate classes in a color code (see Appendix B).

3.1.2.4 transATor

On June 19, 2022, the transATor web application was depreciated. Thus, the stand-alone version was ran locally using the provided Docker container[7] as explained on the corresponding GitHub web-page[8].

[6]https://itol.embl.de/, accessed on June 17, 2022.
[7]https://www.docker.com/products/docker-desktop/, accessed on June 19, 2022.
[8]https://github.com/pcm32/transator-container, accessed on June 19, 2022.

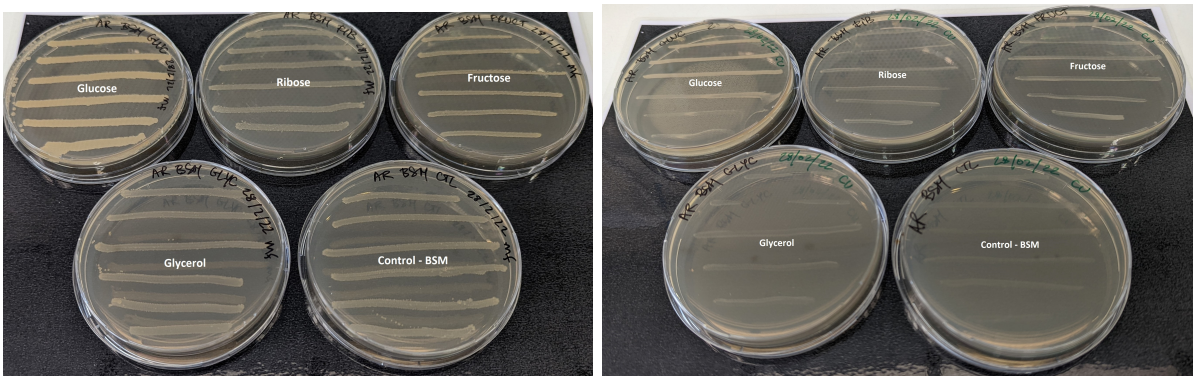
The protein sequence FASTA files of the open reading frame-ordered KS domains of the *trans*-AT PKSs under study were fetched as input. The rough molecular structure and KS substrate annotations were returned as output.

3.2 Results

3.2.1 Bacterial genome engineering

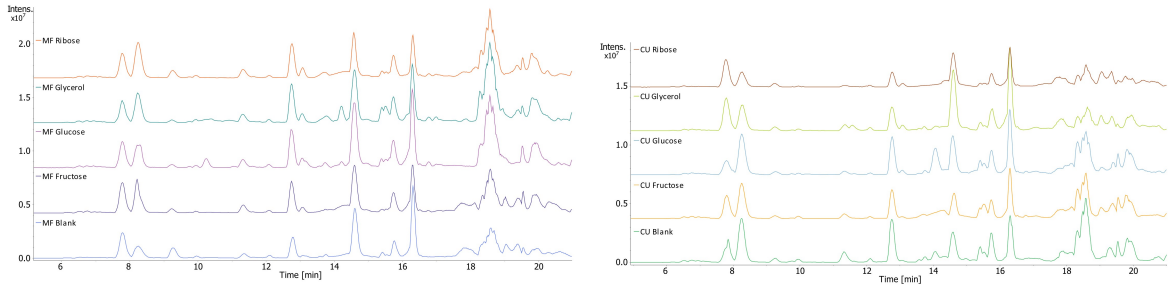
3.2.1.1 Carbon sources investigation

Plates using BSM medium supplemented with glucose as a carbon source show the most vigorous bacterial growth, as shown in the upper left corners of Figure 3.8. Under stressful conditions, glucose is a common carbon source for bacteria growth. Furthermore, the metabolomes of both bacteria were analyzed using LC-MS for each carbon source. As seen in Figure 3.9, the spectra were stacked for visualization convenience. Given a bacteria, the metabolomic patterns are globally similar across carbon sources. Some peaks, such as MF glucose at 10 minutes of retention time, are exclusive to a specific carbon source.
Because the metabolites under investigation are unknown, it is hard to determine the optimal carbon source with certainty. Nonetheless, this process was required to gather knowledge on the metabolomic patterns of both bacteria.



(a) *Massilia flava* (b) *Caballeronia udeis*

Figure 3.8: **Growth of both *M. flava* and *C. udeis* on BSM plates completed with four different carbon sources.** From top left to bottom right, glucose, ribose, fructose, glycerol and control. The control consists of no addition of carbon sources, that is BSM medium.



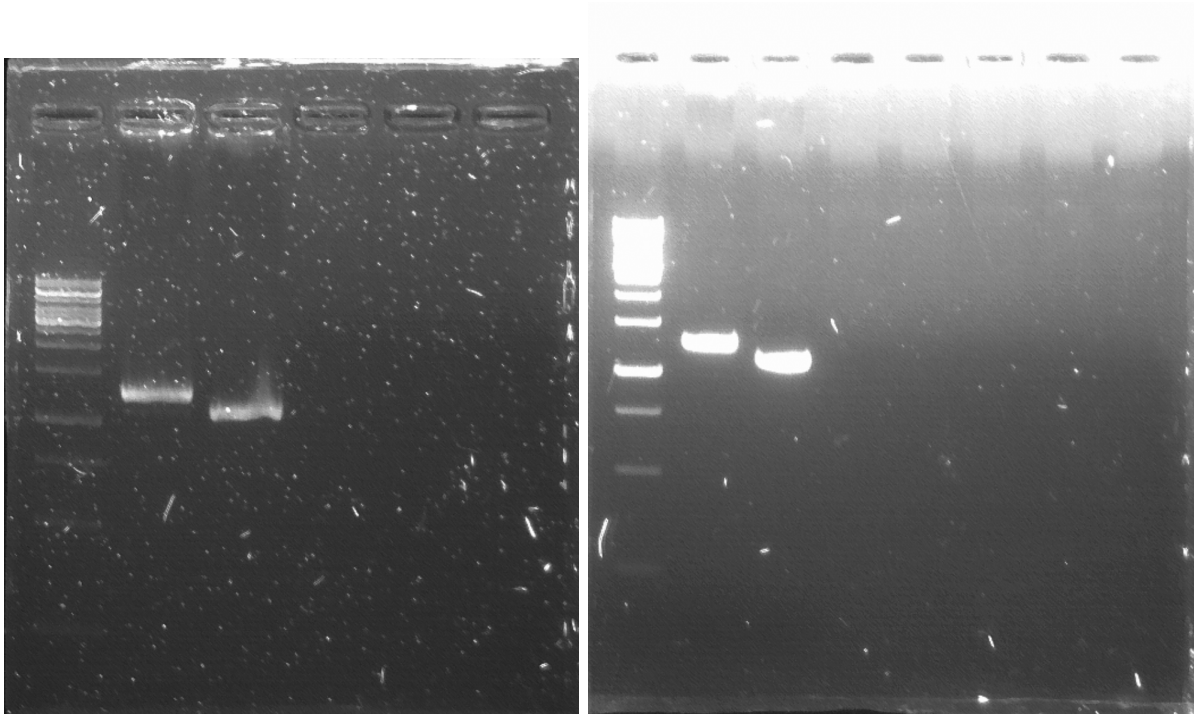
(a) *Massilia flava* (b) *Caballeronia udeis*

Figure 3.9: **Stacked LCMS spectra of both M. flava and C. udeis on BSM plates completed with four different carbon sources.**

3.2.1.2 Construction of recombinant plasmids

The plasmid mixture was PCR amplified to validate the intake of the genes of interest, GO485_00250 and AWB69_2900 from *M. flava* and *C. udeis*, respectively. Figure 3.10 shows that a fraction of the plasmids recombined with the genes of interest while another fraction did not. A size difference of roughly 1000 base pairs can be seen, which corresponds to the size of the genes of

interest. As a result, the production of recombinant plasmids was deemed successful.

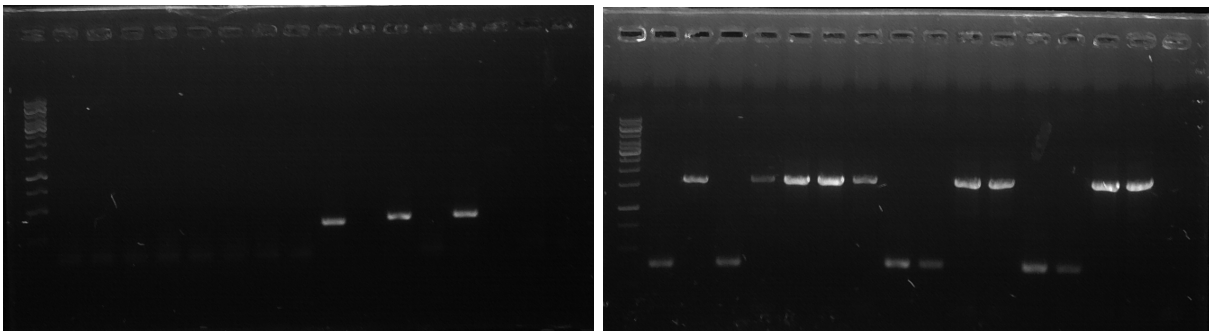


(a) *Massilia flava* (b) *Caballeronia udeis*

Figure 3.10: **PCR outputs confirming the correct construction of recombinant plasmids.** *C. udeis*'s AWB69_2900 gene and *M. flava*'s GO485_00250 gene were inserted in the original genome of pGPI plasmids. The first column outlines a ladder of known length. The second column reflects the size of recombinant plasmid's genomes while the third column represents the size of original plasmid's genome.

3.2.1.3 Selection of transformed chemically competent SY327 *E. coli* cells

The transformation mixture was poured into 16 wells with either empty competent *E. coli* cells or competent *E. coli* cells transformed with recombinant or empty plasmids. Three of the 16 wells contained competent *E. coli* cells that had been transformed with recombinant plasmids for *M. flava*. For *C. udeis*, nine wells were identified as containing chemically competent *E. coli* cells that had been successfully transformed. The heavier recombinant plasmids are detected by mass-driven electrophoresis, as shown in Figure 3.11 for both bacteria.



(a) *Massilia flava* (b) *Caballeronia udeis*

Figure 3.11: **Colony PCR outputs determining the chemically competent *E. coli* cells that have taken up recombinant plasmids** The first column outlines a ladder of known length.

The presence of recombinant plasmids in chemically competent E. coli cells was confirmed by sequencing the genetic material from successfully tagged glycerol stocks. The successfully transformed competent E. coli cells were then employed for the following phase, tri-parental conjugation.

3.2.1.4 Conjugation in a tri-parental design

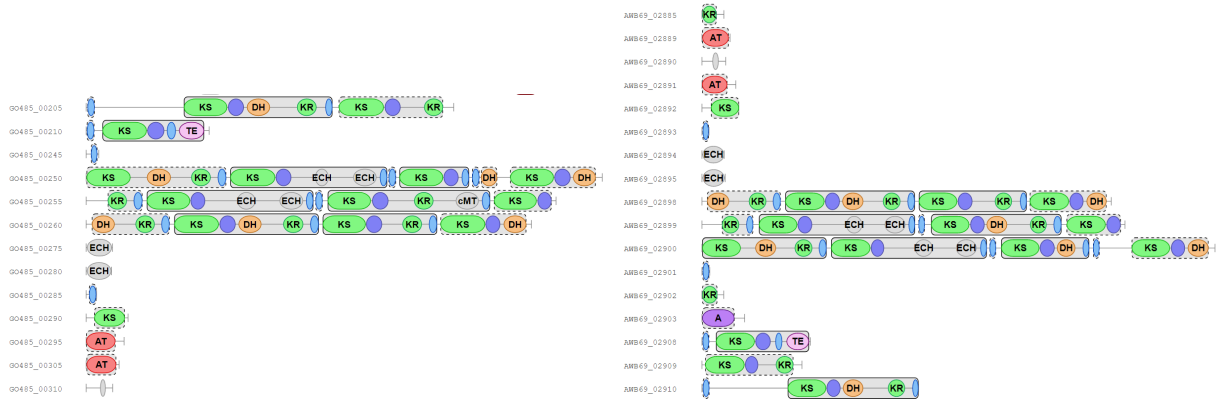
Unfortunately, the conjugation in a tri-parental design did not succeed. The plating of suspensions obtained by scraping the conjugation filter paper revealed no growth. Section 3.3 will go into the reasons for, limits of, and enhancements to tri-parental conjugation utilizing pGPI as a plasmid.

3.2.2 Computational Biology

3.2.2.1 Identification of *trans*-AT gene clusters and annotation of catalytic domains using antiSMASH

Massilia flava
Thirteen gene clusters were identified using a relaxed detection strictness. Within these, antiSMASH identified a *trans*-AT PKS gene cluster in the first region of the *M. flava* genome, located at 27,364 - 139,854 nt. for a total of 112,491 nt. It then predicted the encoding gene, the location and the catalytic activity of the PKS domains as shown in Figure 3.12a.

Caballeronia udeis
Nine gene clusters were discovered using a relaxed detection criterion. Within these, antiSMASH identified a *trans*-AT PKS gene cluster in the FCOK02000016 contig, located at 7,578 - 121,793 nt. for a total of 114,216 nt. It then predicted the encoding gene, the location and the catalytic activity of the PKS domains as shown in Figure 3.12b.



(a) *Massilia flava* (b) *Caballeronia udeis*

Figure 3.12: **antiSMASH identification of PKS/NRPS domains within predicted gene clusters responsible for *trans-AT* PKSs in the *Massilia flava* and *Caballeronia udeis* genomes.** ECH stands for enoyl-CoA hydratase. Blue circles represent *trans*-AT docking locations. Blue ellipses represent PKS PP domains, that is the phosphopantetheine attachment site of ACPs.

3.2.2.2 *trans*-AT KS substrate specificity annotation using transPACT

The results of the transPACT annotation tool for the query PKSs under study are displayed in Table 3.1. For each *trans*-AT KS domain, its corresponding clade number and clade description is given. The clade description gives extensive information about the incoming substrate type

in the assembly line. For instance, a non-elongating KS domain correspond to a KS domain imputed from its elongating catalytic activity.

Table 3.1: **Annotation of the KS substrates specificity using transPACT software.** PBC = *Pseudomonas baetica*, GO485 = *Massilia flava*, AWB69 = *Caballeronia udeis*, DK441 = *Aquimarina* sp. AU58, XB90 = *Burkholderia pseudomallei* MSHR4299, DO63 = *Burkholderia pseudomallei* strain BDE, BLU84 = *Lysobacter enzymogenes*, ASC93 = *Massilia* sp. Root335, CR152 = *Massilia violaccinigra* B2 and K350 = *Sporocytophaga myxococcoides*
A clade not conserved corresponds to a KS specificity not catalogued in the reference phylogeny.

***trans*-AT KS domains**	**Clade number**	**Clade description**
PBC_KS1_01859	Clade_33	non_elongating
PBC_KS2_01861	Clade_36	non_elongating_reduced/bOH
PBC_KS3_01862	Clade_87	exometh/red_bMe
PBC_KS4_01863	clade_not_conserved	NA
PBC_KS5_01863	Clade_23	pyran/furan
PBC_KS6_01863	Clade_56	b_L_OH
PBC_KS7_01863	Clade_63	mainly_eDB
PBC_KS8_01864	Clade_3	bimod_bOH
PBC_KS9_01865	Clade_62	eDB
PBC_KS10_01866	Clade_3	bimod_bOH
PBC_KS11_01867	Clade_63	mainly_eDB
PBC_KS12_01867	Clade_71	b_keto
PBC_KS13_01867	clade_not_conserved	NA
GO485_KS11_00205	Clade_66	double_bonds
GO485_KS12_00205	Clade_31	eDB
GO485_KS13_00210	Clade_5	shifted_double_bonds
GO485_KS1_00250	Clade_2	amino acids
GO485_KS2_00250	Clade_62	eDB
GO485_KS3_00250	Clade_87	exometh/red_bMe
GO485_KS4_00250	Clade_85	aMe_bOH
GO485_KS5_00255	Clade_62	eDB
GO485_KS6_00255	Clade_87	exometh/red_bMe
GO485_KS7_00255	Clade_80	aMe_b_L_OH
GO485_KS8_00260	Clade_62	eDB
GO485_KS9_00260	Clade_66	double_bonds
GO485_KS10_00260	Clade_50	b_OH
AWB69_KS8_02898	Clade_62	eDB
AWB69_KS9_02898	Clade_65	double_bonds
AWB69_KS10_02898	Clade_50	b_OH
AWB69_KS5_02899	Clade_62	eDB
AWB69_KS6_02899	Clade_87	exometh/red_bMe
AWB69_KS7_02899	Clade_111	b_DOH
AWB69_KS1_02900	Clade_2	amino acids
AWB69_KS2_02900	Clade_62	eDB
AWB69_KS3_02900	Clade_87	exometh/red_bMe
AWB69_KS4_02900	Clade_89	b_keto

AWB69_KS13_02908	clade_not_conserved	NA
AWB69_KS12_02909	clade_not_conserved	NA
AWB69_KS11_02910	Clade_68	double_bonds
DK441_KS1_RS14040	Clade_8	GNAT_Starter
DK441_KS2_RS14040	Clade_87	exometh/red_bMe
DK441_KS3_RS14040	Clade_56	b_L_OH
DK441_KS4_RS14040	Clade_80	aMe_b_L_OH
DK441_KS5_RS14045	Clade_62	eDB
DK441_KS6_RS14045	Clade_87	exometh/red_bMe
DK441_KS7_RS14050	Clade_62	eDB
DK441_KS8_RS14050	Clade_87	exometh/red_bMe
DK441_KS9_RS14050	clade_not_conserved	NA
DK441_KS10_RS14060	Clade_62	eDB
DK441_KS11_RS14060	Clade_68	double_bonds
DK441_KS12_RS14060	Clade_50	b_OH
DK441_KS13_RS14005	clade_not_conserved	NA
DK441_KS14_RS14005	Clade_31	eDB
DK441_KS15_RS14015	clade_not_conserved	NA
XB90_KS1_RS21815	Clade_8	GNAT_Starter
XB90_KS2_RS21815	Clade_60	non_elongating_b_OH
XB90_KS3_RS21815	Clade_113	a_Me_b_OH
XB90_KS4_RS21810	Clade_82	b_MeeDB
XB90_KS5_RS21810	Clade_62	eDB
XB90_KS6_RS21810	Clade_82	b_MeeDB
XB90_KS7_RS21810	Clade_80	aMe_b_L_OH
XB90_KS8_RS21805	Clade_1	amino acids
XB90_KS9_RS21805	Clade_12	reduced/shifted_double_bonds
XB90_KS10_RS21800	Clade_60	non_elongating_b_OH
DO63_KS1_RS20800	Clade_1	amino acids
DO63_KS2_RS20800	Clade_12	reduced/shifted_double_bonds
DO63_KS3_RS20785	Clade_62	eDB
DO63_KS4_RS20785	Clade_82	b_MeeDB
DO63_KS5_RS20785	Clade_80	aMe_b_L_OH
DO63_KS6_RS20775	Clade_8	GNAT_Starter
DO63_KS7_RS20775	Clade_60	non_elongating_b_OH
DO63_KS8_RS20775	Clade_113	a_Me_b_OH
DO63_KS9_RS20770	Clade_82	b_MeeDB
DO63_KS10_RS20935	Clade_60	non_elongating_b_OH
BLU84_KS1_RS26565	Clade_2	amino acids
BLU84_KS2_RS26565	Clade_62	eDB
BLU84_KS3_RS26565	Clade_87	exometh/red_bMe
BLU84_KS4_RS26565	Clade_89	b_keto
BLU84_KS5_RS26560	Clade_62	eDB
BLU84_KS6_RS26560	Clade_87	exometh/red_bMe
BLU84_KS7_RS26560	Clade_80	aMe_b_L_OH
BLU84_KS8_RS14405	Clade_62	eDB
BLU84_KS9_RS14405	Clade_66	double_bonds
BLU84_KS10_RS14405	Clade_50	b_OH
BLU84_KS11_RS14470	clade_not_conserved	NA

BLU84_KS12_RS14470	Clade_31	eDB
BLU84_KS13_RS14465	Clade_5	shifted_double_bonds
ASC93_KS1_12760	Clade_111	b_DOH
ASC93_KS2_12760	Clade_47	aMe_eDB
ASC93_KS3_12740	Clade_87	exometh/red_bMe
ASC93_KS4_12740	Clade_82	b_MeeDB
ASC93_KS5_12735	Clade_87	exometh/red_bMe
ASC93_KS6_12735	clade_not_conserved	NA
ASC93_KS7_12730	Clade_62	eDB
ASC93_KS8_12730	Clade_68	double_bonds
ASC93_KS9_12730	Clade_50	b_OH
ASC93_KS10_12785	Clade_43	zDB
CR152_KS1_RS21580	clade_not_conserved	NA
CR152_KS2_RS21580	Clade_62	eDB
CR152_KS3_RS21580	Clade_87	exometh/red_bMe
CR152_KS4_RS21580	clade_not_conserved	NA
CR152_KS5_RS21585	Clade_62	eDB
CR152_KS6_RS21585	Clade_87	exometh/red_bMe
CR152_KS7_RS21585	Clade_111	b_DOH
CR152_KS8_RS21590	Clade_62	eDB
CR152_KS9_RS21590	clade_not_conserved	NA
CR152_KS10_RS21590	Clade_50	b_OH
CR152_KS11_RS21525	clade_not_conserved	NA
CR152_KS12_RS21525	Clade_31	eDB
CR152_KS13_RS21530	Clade_41	acetyl_Starter
K350_KS1_RS29580	Clade_10	aromatic_Starter
K350_KS2_RS29580	Clade_47	aMe_eDB
K350_KS3_RS32630	Clade_87	exometh/red_bMe
K350_KS4_RS32430	Clade_62	eDB
K350_KS5_RS32430	Clade_87	exometh/red_bMe
K350_KS6_RS32430	clade_not_conserved	NA
K350_KS7_RS0119365	Clade_82	b_MeeDB
K350_KS8_RS0119365	Clade_50	b_OH
K350_KS9_RS0119325	Clade_45	lactate_Starter
K350_KS10_RS0119325	clade_not_conserved	NA
K350_KS11_RS0119325	clade_not_conserved	NA

3.2.2.3 Dendrogram representation of PKSs using transPACT

The results of the transPACT comparison tool are shown on the rectangular phylogeny on Figure 3.13. This dendrogram will be further discussed in the next section.

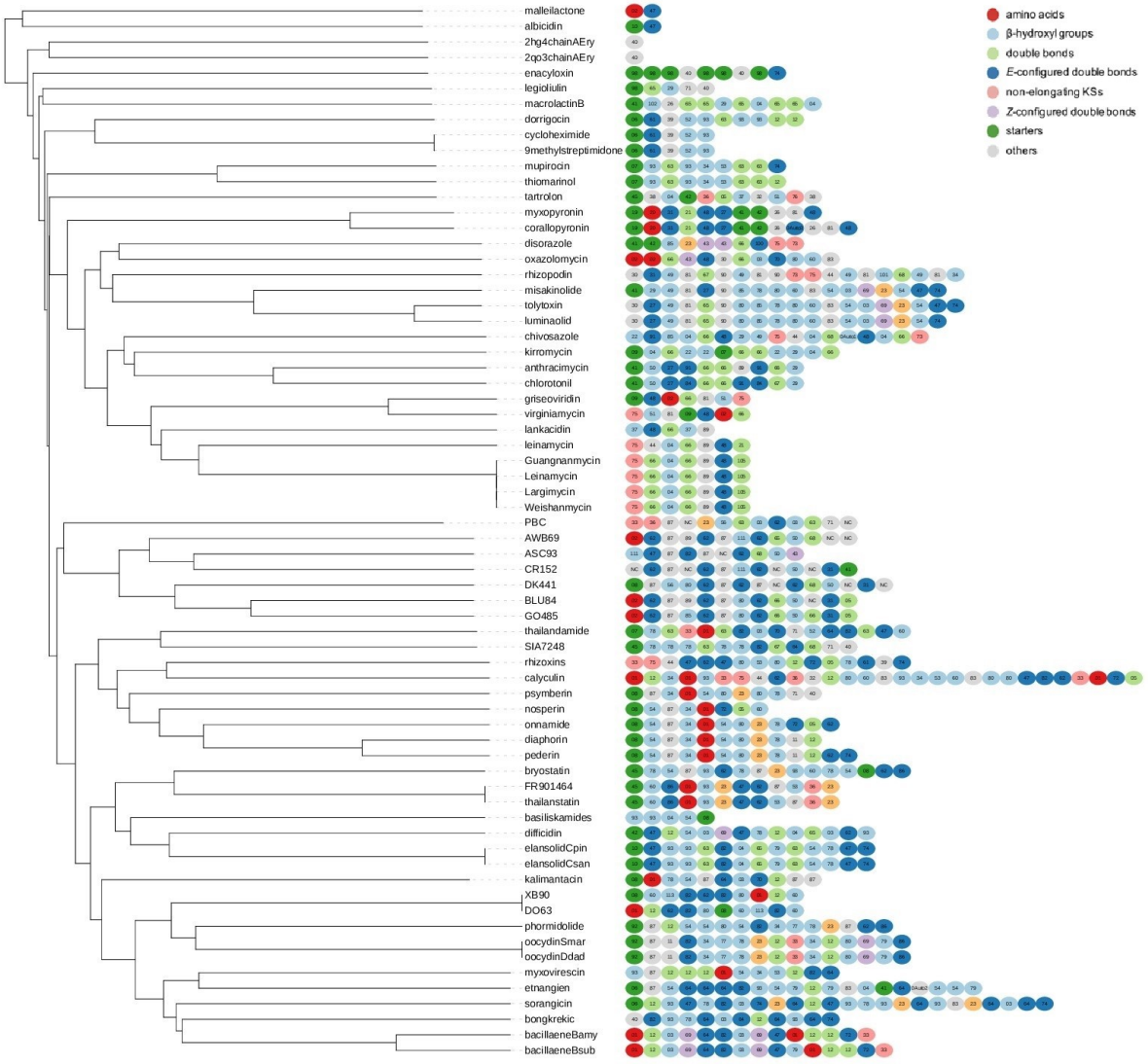


Figure 3.13: **Rectangular dendrogram visualization of conserved PKS motifs across the *trans*-AT PKSs under study placed on the the PKSs of reference.** Generated using the transPACT software as described on Figure 3.7. A fully detailed colors code is given in Appendix C.

3.3 Discussions

In the present section are discussed the aforementioned methods and results.

3.3.1 Bacterial conjugation using pSF100 plasmid

Attempts to accomplish conjugation using pGPI in a tri-parental design were unsuccessful. Initially, pGPI was chosen as a promising plasmid option because it allowed for negative selection using antibiotics such as trimethoprim and triclosan. Because pGPI lacks a replication origin, helper E. coli cells were necessary, explaining the tri-parental design. The laboratory was also used to work with pGPI plasmids.

The bacteria C. udeis and M. flava were found to be particularly slow growing. Three kinds of bacteria competed for resources on plates during the tri-parental conjugation. The latter might be one explanation for conjugation failures. To be more specific, if this explanation was verified, the conjugation would not have failed properly speaking. Given the large number of proliferating E. coli colonies on plates, appropriately transformed *C. udeis* and *M. flava* cells were simply unable to grow.

Prospects for future work are being investigated using pSF100 as plasmids for bacterial genome editing activities. Because pSF100 plasmids have an origin of replication, an additional E. coli helper strain is no longer necessary. As a consequence, conjugation may be done in a bi-parental configuration, reducing the number of *E. coli* cells competing with *C. udeis* and *M. flava* cells on plates. It follows that another antibiotic will be used for negative selection. Furthermore, techniques for monitoring bacterial conjugation using Congo Red Stain are being developed.

3.3.2 Conservation of module blocks across *trans*-AT PKSs related to *C. udeis* and *M. flava*

The transPACT software identifies conserved KS patterns across PKSs by estimating their similarities based on genetic distance, synteny and conservation of KS functional specificity. Moreover, the transATor program predicts the core polyketide backbone arising from the assembly lines under investigation. Figure 3.14 was created using this process. Since both software use different reference trees to annotate KS specificities, the predictions differ. Both predictions will be discussed and compared. At first, the results from transPACT, namely Figure 3.13, will be discussed. Additionally, a list matching each clade number to its description can be found in Appendix B.

As hypothesised with a BLAST search, A high KS domain conservation occurs between *C. udeis*, *M. flava* and related PKSs. It should be noted that both PKSs from *Burkholderia pseudomallei* bacteria do not cluster with the rest. One of the highest conserved domain region is observed between PKSs from *C. udeis* and *M. flava*. The incoming substrate specificity of KS domains is strictly conserved for KS1-6. Then, the specific substrate accepted by KS7 differs between GO485 and AWB69. The seventh KS of the former accepts substrates harbouring α-methylene-β-*L*-hydroxyl group (clade 80) while the one of the latter accepts substrates harbouring β-D-hydroxyl group (clade 111). It follows that the GO485 polyketide backbone is predicted by transATor to host an α-dimethylene-β-hydroxyl group compared to the minimal β-hydroxyl group predicted for the AWB69 polyketide backbone. Moreover, the module block KS8-11 is also strictly conserved across *C. udeis* and *M. flava*. Clade 65, 66 and 68 correspond

to the same annotation, namely double bonds. However, transPACT was not able to annotate the specificity of the two last KS domains from AWB69 PKSs.
Additionally, according to the transPACT annotation tool, the BLU84 PKS from *Lysobacter enzymogenes* is the closest to the GO485 PKS. Indeed, the two assembly lines are widely similar with eleven out of thirteen KS domains functionally conserved. The fourth KS is predicted to accept α-methylene-β-hydroxyl group (clade 85) and β-ketone group (clade 89) for, respectively, BLU84 and GO485.

Interestingly, two module blocks are particularly recurrent across the six PKSs, labelled as KS5-6 and KS8-10. Firstly, KS5 and KS6 were mostly annotated as clade 62 and 87, corresponding, respectively, to domains accepting *E*-configured double bonds and exomethylene or reduced β-methylene groups. Secondly, KS8, KS9 and KS10 were annotated in majority as clade 62, clades 65/66/68 and clade 50. Clade 62 reflects KS domain accepting substrate with *E*-configured double bonds. Then, clade 65, 66 and 68 correspond to KS domains being specific for double bonds groups. Finally, clade 50 gathers KS domains accepting substrates harbouring hydroxyl group on their β-carbon.

The molecular backbones depicted on Figure 3.14 were generated using the transATor prediction tool. The KS specificity annotations predicted by the transATor differ than the ones predicted by the transPACT as they do not implement the same reference tree. The transATor-predicted KS annotations are outlined in Figure 3.15. When comparing the KS annotations from transPACT and transATor for *C. udeis* and *M. flava*, it is clear that even if the predictions are generally consistent, a couple differences between the two software's prediction tools are still present, as listed below.

1. The fourth KS domains are predicted to select β-ketonic (clade 89) substrates by transPACT. However, transATor predicts the selection of substrates presenting α-methylene shifted double bonds or hydroxyl groups.

2. *M. flava*'s seventh KS domain is predicted to accept substrates carrying α-methylene-β-L-OH (clade 80) and α-L-(di)methylene-β-OH according, respectively, to transPACT and transATor. Ultimately, the transATor-predicted core backbone of the polyketide arising from *M. flava*'s *trans*-AT assembly line will harbour a dimethylene group.

3. For the ninth KS domain, transATor predicts an *E*-configured double bonds (eDB) while transPACT simply annotates the incoming substrates as mounted by a double bonds group (clades 65 and 66) with no regards of its configuration for both PKSs.

4. In both PKSs, the transATor-assigned annotation for the tenth KS domain relates to D-oriented β-hydroxyl group. However, *trans*PACT-assigned annotations do not include any specification about the orientation of the β-hydroxyl group (clade 50).

5. Predictions also differ from both software for the eleventh KS domain specificity. For both PKSs, transPACT predicts a double bonds specificity (clade 66 and 68). However, transATor predicts a β-D-hydroxyl group specificity and a β-ketone or double bonds specificity for, respectively, *M. flava* and *C. udeis*. It should be noted that for *C. udeis*, the predictions overlap as incoming substrates mounted with a double bonds motif is accepted by both software.

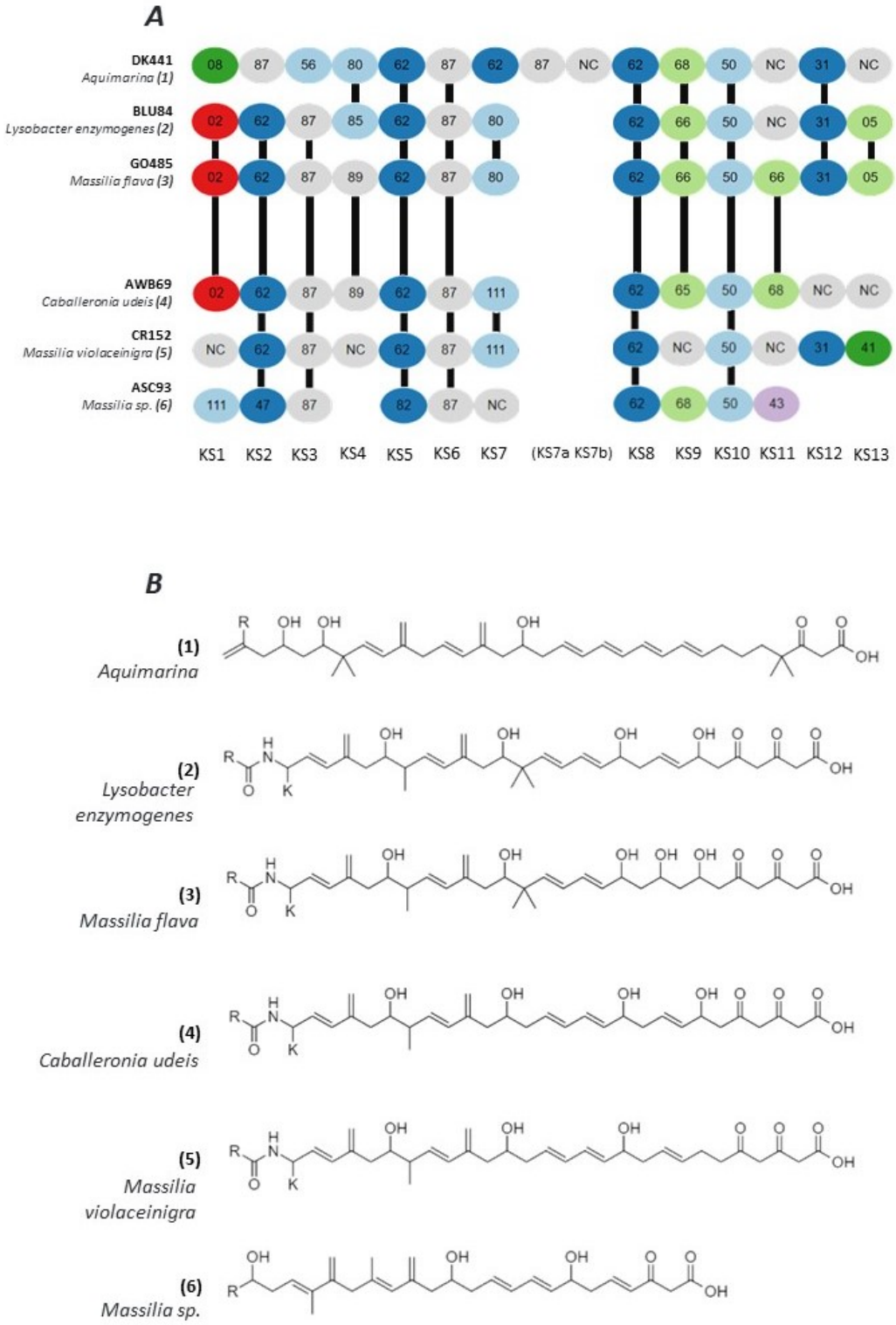


Figure 3.14: **Module blocks similarities between *C. udeis* and *M. flava*-like *trans*-AT PKSs. A.** Horizontal black lines reflects KS substrate specificities conservation. The PKSs from *C. udeis* and M. *flava* are highly similar. The KS10-11-12 block is highly conserved across the six PKSs, ruling the formation of two double bonds and a hydroxyl group. Also, the KS5-6 block is particularly well conserved, ruling the formation of a double bond followed by an exomethylene group. **B.** Core molecular structures were predicted using the transATor software while the dendrogram was generated using the transPACT software thus specificity annotations slightly differ.

6. For the last two KS domains of both assembly lines, transATor is not able to give accurate predictions. Both motif predictions fell in the *various specificities* class. On the other side, transPACT endures the same setback for *C. udeis* PKS, annotating the KS domains as *non conserved* (clade NC). However, it was able to find a significant clade placement (clade 31 and 05) on the reference tree for the *M. flava* PKS's twelfth and thirteenth KS domains.

The annotations provided by transPACT and transATor must coincide to be inferred from one tool to another. If both tools were to implement the same unique reference tree, they might be combined. This master tool would accept FASTA files containing concatenated protein sequences of *trans*-AT PKS BGC as input. Then, on this file, transATor and transPACT could both be applied. A KS domains identification component should be included upstream of transPACT in order to select just the *trans*-AT KS sequences. Profile Hidden Markov Models (pHMMs) can be used to implement the latter. The accompanying materials provide a python script for recognizing *trans*-AT KS domains (see Data Accession). This will be further discussed in the genome mining approach. Alternatively, an analogous method is already integrated in the transATor program's pipeline.

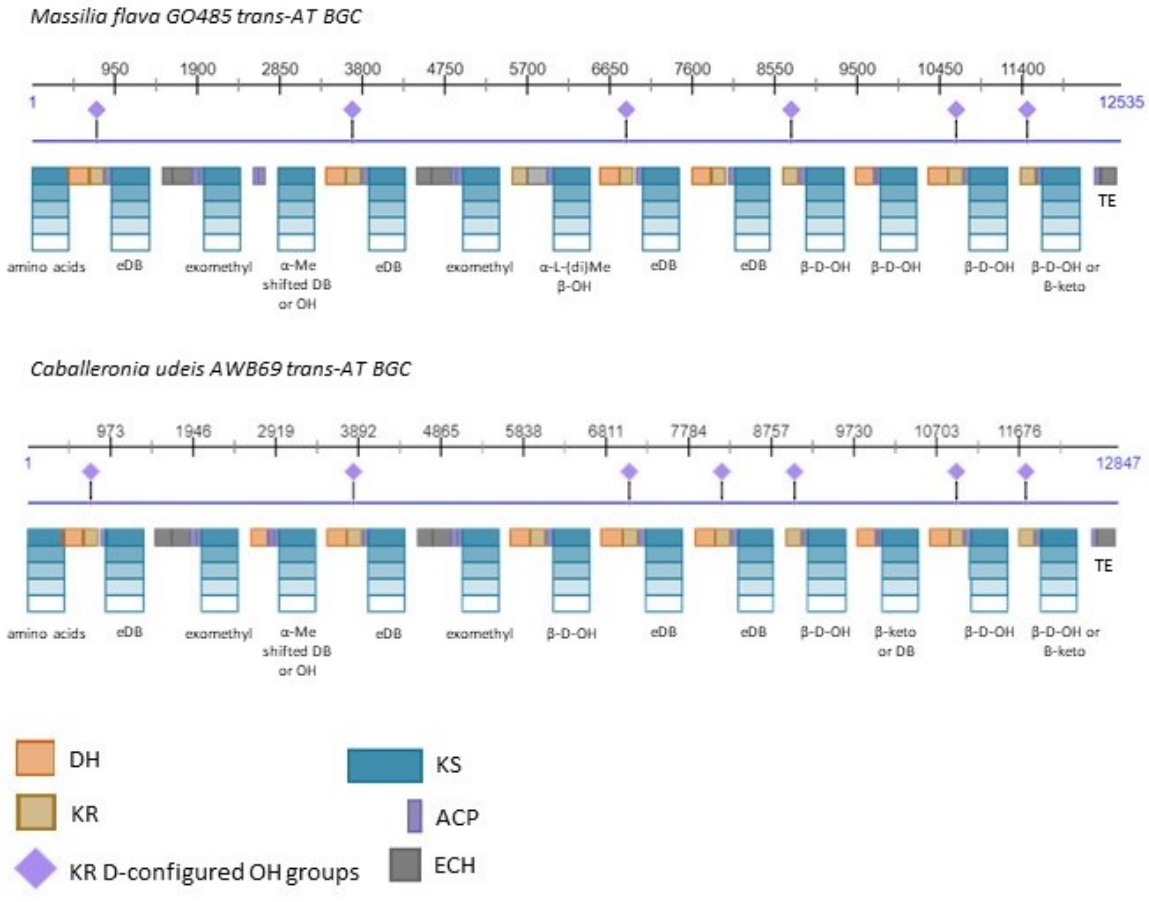


Figure 3.15: **TransATor-predicted KS substrate specificity annotations for *C.udeis* and *M.flava* PKSs.** A color legend is given on the figure. β and α refer to the position of the group on the carbon backbone.

As a corollary, the execution of both software might well be sequential in such a way that transPACT benefits from the outputs of the transATor. As a side note, the action of post tailoring enzymes are not taken into account in both software. Thus, the molecular predictions

are rough structure corresponding to the linear backbones of polyketides arising from assembly lines. They are unlikely to represent the final bio-active polyketides. They still give valuable and relevant information about the functional groups shaping the ultimate molecules. Finally, the complexity of the reference tree impacts the performance of both annotation tools. Currently, transATor maintains a reference tree of 90 clades, whereas transPACT implements a reference tree of 128 functional KS clades. The scientific community is constantly deciphering the functional multiplicity of trans-AT KS domains. As a result, constant revision of the reference tree is essential to keep the annotations as accurate and feasible as possible. The latter is vividly illustrated by the KS domains from *C. udeis* and *M. flava* annotated as *non-conserved* or *various specificities* by both tools. It implies that the present reference trees solely take into consideration characterized KS substrate specificities. Because unknown functionalities are inherent in uncharacterized PKSs, this constitutes a significant setback in the use of those algorithms in natural product discovery. These bioinformatics methods are particularly useful in elucidating the overall mechanism leading to polyketides' core backbones. Subsequently, they will emphasise the idiosyncrasies of uncharacterized PKSs, but will not allow annotations of such features. Essentially, a KS substrate specificity can only be appropriately labelled if the reference tree of KS sequences has a corresponding specificity.

3.3.3 Limitations of transPACT

The main limitations of the transPACT tool are the encoding of KS labels into an open reading frame order and its computational complexity.

3.3.3.1 Open reading frame against genetic order

The KS sequences fed into transPACT must be encoded in a consistent order. There are two orders to consider: the open reading frame order and the genetic order. On the one hand, the open reading frame framework encodes KS sequences according to gene expression order. The genetic framework, on the other hand, encodes KS sequences according to their underlying genetic arrangement. *Massilia flava*'s *trans*-AT PKS demonstrates the distinctions between both orders of encoding KS sequences. The antiSMASH-annotated domains are shown in their underlying genetic arrangements in Figure 3.12. In such a case, a gene's location in the biosynthetic gene cluster under study determines its order. Thus the ordering labels assigned to KS sequences would range from KS1 to KS13, starting with the first KS domain of the GO485_02898 gene and concluding with the last KS domain of the GO485_02910 gene.
The sequence of gene expression, however, does not necessarily correspond to the underlying genetic architecture. Indeed, a first gene that comes before a second gene in the same biosynthetic gene cluster is not particularly expressed or translated before the second gene. Figure 3.6 depicts the reconstruction of the open reading frame order. It denotes the actual sequence in which each domain achieves catalytic activity.

The open reading frame order is favored for interpretation over the genetic order. KS sequences in the antiSMASH ClusterBlast database, although, are encoded according to their underlying genetic arrangement. The reconstruction of the open reading frame order is currently handled manually. Remarkably, the sixth version of antiSMASH now predicts the open reading frame sequence of domains [61]. Still, KS sequences ordering labels must be manually encoded. As a result, a fully automated method for transitioning from an underlying genetic framework to an open reading frame framework would be a significant improvement to the transPACT software. Furthermore, it would represent a significant advance in the software's high-throughput utilization. TransPACT currently requires a substantial effort to manually reconstruct the open

reading frame sequence and then encode the KS sequences order in a FASTA file for more than 10 PKSs. Recent versions of antiSMASH i.e. v6 and onward allow for the development of such an automated order conversion method.

The current transPACT version's KS sequences of reference were manually organized by its developers based on their open reading frame layout. The application of transPACT in the present master's thesis is known as query targeted. Indeed, a few PKSs under investigation are fetched into transPACT following manual encoding of their KS sequence headers in accordance with their open reading frame arrangements. Genome mining, on the other hand, refers to the large-scale application of transPACT. In a subsequent section, the cornerstone of a *trans*-AT genome mining strategy implementing transPACT is explained.

Finally, for the transPACT software to be used in large scale and versatile set-ups, a reduced computational complexity, a standardized naming system for KS numbers and a consistent method to reconstruct the open reading frame order out of the genetic order of KS sequences must be developed.

3.3.3.2 Computational complexity

The annotation of KS domain specificities was benchmarked at 8 seconds per KS sequence on the Vlaamse supercumputer, similar to the benchmark of seven seconds computed by Helfrich *et al.* [50]. The computational complexity of transPACT is linear on the number of input KS sequence. Linearity makes transPACT a promising tool but the benchmarked time remains long.

3.3.4 Genome mining approach using transPACT

3.3.4.1 Adaptations to a genome mining approach

TransPACT was previously utilized in a query-targeted approach. The investigation concentrated on a few key PKSs. A genome mining strategy is presently being discussed. The use of bioinformatics algorithms for the discovery of novel natural product biosynthetic pathways is referred to as genome mining [62]. With the exception of two drivers, the approach utilized in a genome mining setup mimics the query targeted one, namely the large-scale identification of *trans*-AT PKSs and the adaption of the KS numbers labeling strategy to the naming structure of the antiSMASH ClusterBlast database.

Because the entire ClusterBlast database provides the input data, the initial step is to identify *trans*-AT PKSs in the database. *Trans*-AT KS domains were identified using a python script provided by M. Medema implementing Hidden Markov Models derived from the SMART software[9]. Trans-AT KS domains were identified using the PKS.hmm model. AT domains were identified using the AT.hmm model. Then, identified trans-AT KS domains were filtered out if they lied in the close proximity of an identified AT domain. Out of the approximately 3,000,000 BGC-encoded protein sequences available in the ClusterBlast database, 17,027 were identified as trans-AT KS domains across 5,543 trans-AT PKSs distributed over 3,712 microbial genomes.

Because the current master's thesis focuses on trans-AT PKSs with around 13 KS domains, as in *C. udeis* and *M. flava*, only PKSs with 10 to 16 KS domains were chosen for future examination. As a result, 511 of the 5,543 HMM-identified *trans*-AT PKSs were chosen. It also

[9]The Simple Modular Architecture Research Tool (SMART) software analyses protein domain architecture in order to annotate domain families [63].

allowed the annotation tool's computing time to be considerably decreased, while concentrating efforts on PKSs of interest. It should be emphasized that the majority of the detected PKSs had two or less KS domains, with just 24.1 percent of the database corresponding to pathways with three or more KS domains.

Furthermore, BGC-encoded protein sequences contained in the ClusterBlast database are named according to the following structure:

> *<Genome ID>|<Pathway positional ID>|<Module positional ID>|<Strand sense>|<Prefix Locus Tag>|<Description>|<NCBI Refference Sequence ID>_ <KS number within a same module>*

In the query targeted approach, KS numbers were listed among assembly lines, but in the present large scale approach, they are listed among modules. For illustration, in a two-module PKS with three KS domains in each module, the KS domains would be specified as KS1, KS2, and KS3 inside each module. As a result, the ClusterBlast PKSs had to be adjusted in order to transition from module-wide to assembly line-wide KS number labelling. A Python script developed for the current master's thesis was used to do this. It is accessible by following the data accession guidelines listed in the Data Accession section.

3.3.4.2 Limitations

The primary drawback of the genome mining approach is the antiSMASH ClusterBlast database's genetic order encoding of KS domains. The genetic order encoding does not adequately reflect the actual open reading frame expression of genes, as was described earlier in the Open reading frame vs genetic order section, and is thus not preferred for interpretation.

In contrast, the PKSs under investigation, including those from *M. flava* and *C. udeis*, were encoded in an open reading frame order. As a result, comparing conserved module blocks from two distinct encoding schemes is difficult. The next section discusses Figure 3.16.

3.3.4.3 Conserved module blocks across PKSs identified in the antiSMASH ClusterBlast database

TransPACT was utilized in a genome mining setup to find conserved module blocks in ten to sixteen KS domains long PKSs from the antiSMASH ClusterBlast database. The results, depicted in Figure 3.16, will be discussed, with a focus on C. udeis, M. flava, and related PKSs.

Overall, the present study revealed three well-defined PKS clades that catalyse characterised natural products, namely bacillaene, macrolactin H, and difficidin. This is consistent with the database-wide conservation study of assembly lines across bacteria reported in 2021 by Helfrich *et al.* [50].

The genome mining approach also guided the finding of an additional assembly line substantially related to *C. udeis* and *M. flava* PKSs. Nine of the thirteen KS domains that make up the related PKSs are shared by *C. vaccinii*'s PKS, as seen in Figure 3.17. It should be emphasized that for the sake of comparison with *C. udeis*, *M. flava* and related PKSs, the order of the KS domains for both *C. vaccinii* strains was manually switched from a genetic order to an open reading frame order. Interestingly, this newly discovered genome contains a *trans*-AT PKS that is still uncharacterized. It serves as more proof that the PKSs under investigation

catalyze assembly lines manufacturing a metabolite that has not yet been assigned a specific identity. Furthermore, *C. udeis*, *M. flava* and related PKSs cluster within a heterogeneous clade, together with a variety of uncharacterised PKSs. Non-conserved KS domains - designated as NCs in Figures 3.16 and 3.17 - are strikingly abundant in this heterogeneous clade. In fact, such KS domain specificities could not be characterized by the reference tree utilized in transPACT's annotation tool. Uncharacterised KS domain specificities, biosynthesis pathways, and metabolites offer opportunities for future study in an environment of discovery.
Ultimately, there is considerable evidence that the PKSs under investigation encode for an assembly line producing an unique metabolite that still remains uncharacterized.

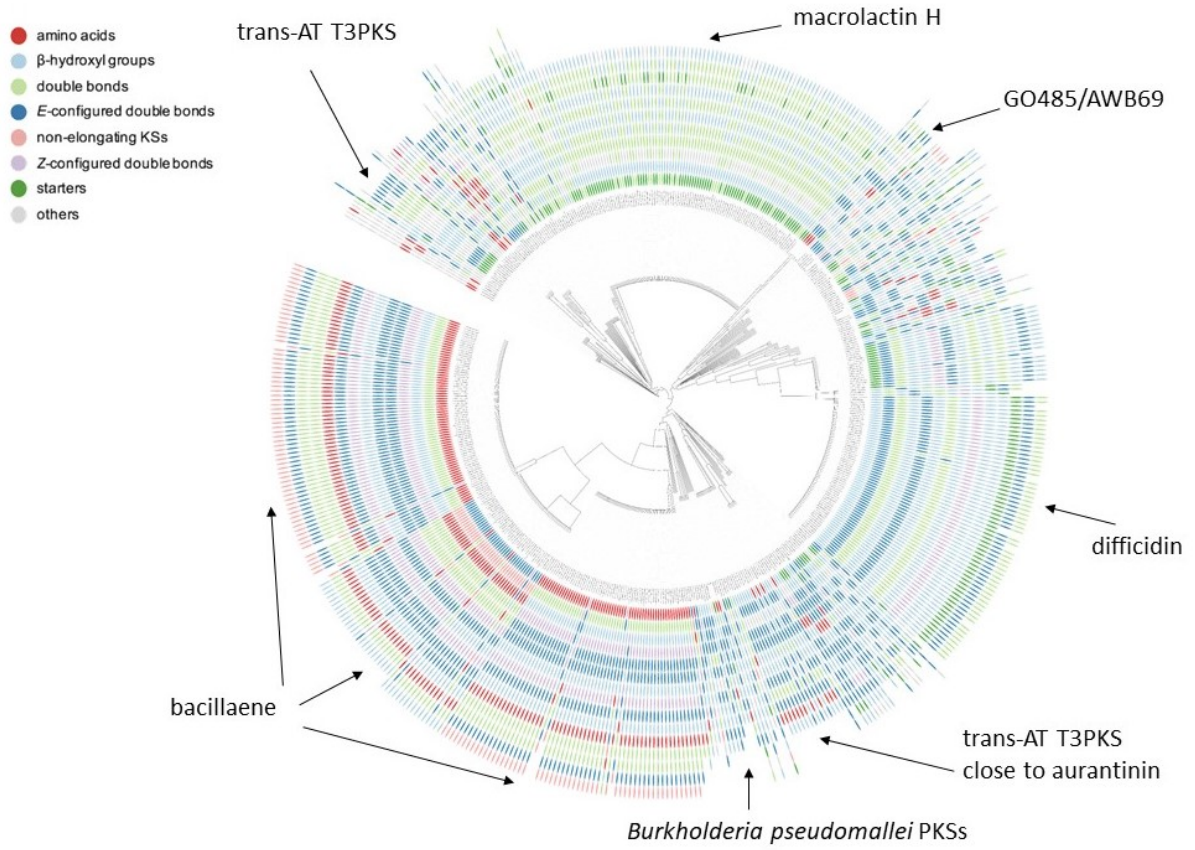


Figure 3.16: **Circular dendrogram representation of conserved PKS motifs across 10 to 16 KS domains long *trans*-AT PKSs from the antiSMASH ClusterBlast database, overlaid with PKSs under study.** The figure was generated using the transPACT software as described on Figure 3.7. A fully detailed colors code is given in Appendix C. Circular and rectangular dendrogram visualizations can be accessed on https://itol.embl.de/export/852011852072360831658172247.

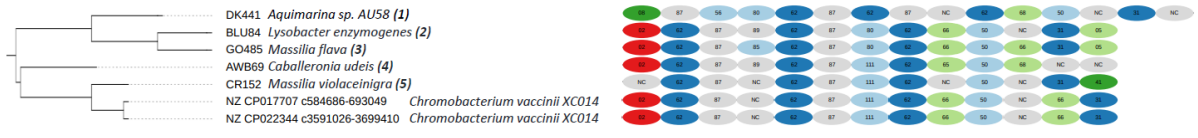


Figure 3.17: **Module blocks conservation between C. udeis, M. flava and related *trans*-AT PKSs, identified following the genome mining approach.** An additional PKS from *Chromobacterium vaccinii* XC014 was identified to be highly related to *C. udeis* and *M. flava* PKSs, that can be accessed under NZ_CP017707 on NCBI RefSeq database.

3.4 Data Accession

Data on PKSs from C. udeis and M. flava are restricted thus are accessible upon request to joleen.masschelein and antoinealexism.ruzettechevalier@student.kuleuven.be in a private repository. Python scripts developed in the context of the present master's thesis are freely available in the following repository: https://github.com/AntoineRuzy/BDE_transPACT.

The query targeted approach's dendrogram representations are publicly available on the following ITOL sharing link: https://itol.embl.de/export/85201119494240216558271 99.
The genome mining approach's dendrogram representations are publicly available on the following ITOL sharing link: https://itol.embl.de/export/85201185207360831658172247.

A comprehensive and user-friendly instructional for transPACT, viewable in your preferred web browser, can be found on the aforementioned public repository.

Chapter 4

Concluding Remarks

The central objective of this master's thesis research project was to discover the metabolites synthesized by polyketide synthases from C. udeis and M. flava. The combination of bacterial genome engineering and bioinformatics tools allowed key progress toward this goal. A productive partnership resulted from providing bioinformatics support to the team from the Laboratory for Biomolecular Discovery at the VIB - KU Leuven.

On the level of bacterial genome engineering, conjugation attempts using pGPI as a plasmid in a tri-parental design were unsuccessful. The slow growing *C. udeis* or *M. flava* cells compete for plate resources with the plentiful *E. coli* cells, which is a plausible reason for this failure. It was concluded that tri-parental conjugation, as well as the usage of pGPI as plasmid, were not suitable for disrupting the *trans*-AT PKS genes of *C. udeis* and *M. flava*. Future research will therefore focus on the usage of pSF100 as plasmid in a bi-parental conjugation set-up in an effort to limit the competition between *E. coli* cells and bacterial cells under study. Another future development axis is the implementation of conjugation monitoring methods using Congo Red stain to evaluate the conjugation success.

On the level of computational biology, the *trans*-AT Polyketide Annotation and Comparison Tool (transPACT) tool was implemented and the team from the laboratory was trained. For the team to employ transPACT in future research initiatives, a comprehensive and user-friendly instructional was created. TransPACT annotated the KS domains specificities by phylogenetic placement on a reference tree. Conservation patterns of module blocks across PKSs were then visualized by dendrogram representations. Initially, PKSs that are very similar to *C. udeis* and *M. flava* were identified surreptitiously by BLAST search. TransPACT, which was first employed in query-targeted way, corroborated BLAST's observations. It also provided information on the conserved module blocks across *M. flava*, *C. udeis*, and related PKSs. Rough polyketide structures arising from the assembly lines encoded by C. udeis and M. flava were predicted by the transATor software. The latter employs an approach for annotating KS domains specificities that is comparable to the former.
Furthermore, transPACT was applied to the antiSMASH ClusterBlast database following a genome mining approach. The lack of conservation between *C. udeis* and *M. flava*'s PKSs and PKSs of the same length supported the research's promising aspects. Indeed, PKSs under study did not cluster with clades of characterised PKSs, indicating that the research is geared toward the identification of novel natural products. Interestingly, another related PKS from *Chromobacterium vaccinii* XC014 was discovered, providing an additional improvement in data availability. An appealing future project would be the development of software that uses profile hidden markov models (pHMMs) to convert PKSs from a genetic order to an open reading frame order. The sixth edition of antiSMASH currently provides such an application, however the newly encoded KS ordering labels are not yet provided. At the moment, just an image output may be downloaded. Furthermore, incorporating transPACT into the antiSMASH workflow would introduce challenges previously described in the software's limitations, which, if addressed, would result in a substantial and widely accessible application.

The laboratory will continue the current research. The next step is to successfully disrupt the *trans*-AT PKSs of *C. udeis* and *M. flava* for metabolomic comparisons. The rough polyketide structures predicted by bioinformatics will subsequently be refined by LC-MS analysis. If the discovered polyketides are shown to be biologically relevant, their synthesis and further optimizations will be the next exciting research project.

On a personal level, this master's thesis has been challenging, rewarding and meaningful. This would not have been possible without the team at the laboratory, particularly Joleen and Angus.

Bibliography

[1] Charles Darwin. *On the origin of species*. Macmillan Collector's Library, London, England, January 2017.

[2] Kostas Kampourakis. *Evolutionary Processes*, page 121–148. Understanding Life. Cambridge University Press, 2 edition, 2020.

[3] Dimitrios T Karamitsos. The story of insulin discovery. *Diabetes Res Clin Pract*, 93 Suppl 1:S2–8, August 2011.

[4] Matthew Cummings, Rainer Breitling, and Eriko Takano. Steps towards the synthetic biology of polyketide biosynthesis. *FEMS Microbiol Lett*, 351(2):116–125, January 2014.

[5] Sung Ryeol Park, Young Ji Yoo, Yeon-Hee Ban, and Yeo Joon Yoon. Biosynthesis of rapamycin and its regulation: past achievements and recent progress. *J Antibiot (Tokyo)*, 63(8):434–441, June 2010.

[6] Suzanne M Ma and Yi Tang. Biochemical characterization of the minimal polyketide synthase domains in the lovastatin nonaketide synthase LovB. *FEBS J*, 274(11):2854–2864, April 2007.

[7] Ajithkumar Vasanthakumar, Karuppasamy Kattusamy, and Ranjan Prasad. Regulation of daunorubicin biosynthesis in streptomyces peucetius - feed forward and feedback transcriptional control. *J Basic Microbiol*, 53(8):636–644, February 2013.

[8] Kira J Weissman and Peter F Leadlay. Combinatorial biosynthesis of reduced polyketides. *Nat Rev Microbiol*, 3(12):925–936, December 2005.

[9] V.S. Malik. Microbial secondary metabolism. *Trends in Biochemical Sciences*, 5(3):68–72, 1980.

[10] Brandon T Pfannenstiel and Nancy P Keller. On top of biosynthetic gene clusters: How epigenetic machinery influences secondary metabolism in fungi. *Biotechnol Adv*, 37(6):107345, February 2019.

[11] V A Mironov, O V Sergienko, I N Nastasiak, and V N Danilenko. [biogenesis and regulation of biosynthesis of erythromycins in saccharopolyspora erythraea: a review]. *Prikl Biokhim Mikrobiol*, 40(6):613–624, November 2004.

[12] Maria J Martin, Javier Herrero, Alvaro Mateos, and Joaquin Dopazo. Comparing bacterial genomes through conservation profiles. *Genome Res*, 13(5):991–998, April 2003.

[13] David H Kwan and Frank Schulz. The stereochemistry of complex polyketide biosynthesis by modular polyketide synthases. *Molecules*, 16(7):6092–6115, July 2011.

[14] Michael A Fischbach and Christopher T Walsh. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem Rev*, 106(8):3468–3496, August 2006.

[15] J D Watson and F H Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. J.D. watson and F.H.C. crick. published in nature, number 4356 april 25, 1953. *Nature*, 248(5451):765, April 1974.

[16] John M Coffin and Hung Fan. The discovery of reverse transcriptase. *Annu Rev Virol*, 3(1):29–51, July 2016.

[17] Christian Hertweck. Decoding and reprogramming complex polyketide assembly lines: prospects for synthetic biology. *Trends Biochem Sci*, 40(4):189–199, March 2015.

[18] C M Kao, L Katz, and C Khosla. Engineered biosynthesis of a complete macrolactone in a heterologous host. *Science*, 265(5171):509–512, July 1994.

[19] Aisha H. Al-Moubaraki and I.B. Obot. Top of the line corrosion: causes, mechanisms, and mitigation using corrosion inhibitors. *Arabian Journal of Chemistry*, 14(5):103116, 2021.

[20] Hei Sook Sul and Stuart Smith. Chapter 6 - fatty acid synthesis in eukaryotes. In Dennis E. Vance and Jean E. Vance, editors, *Biochemistry of Lipids, Lipoproteins and Membranes (Fifth Edition)*, pages 155–190. Elsevier, San Diego, fifth edition edition, 2008.

[21] Zhe Zhou, Jonathan R Lai, and Christopher T Walsh. Interdomain communication between the thiolation and thioesterase domains of EntF explored by combinatorial mutagenesis and selection. *Chem Biol*, 13(8):869–879, August 2006.

[22] Ewa Maria Musiol-Kroll and Wolfgang Wohlleben. Acyltransferases as tools for polyketide synthase engineering. *Antibiotics*, 7(3), 2018.

[23] Ralph Reid, Misty Piagentini, Eduardo Rodriguez, Gary Ashley, Nina Viswanathan, John Carney, Daniel V. Santi, C. Richard Hutchinson, and Robert McDaniel. A model of structure and catalysis for ketoreductase domains in modular polyketide synthases. *Biochemistry*, 42(1):72–79, 2003. PMID: 12515540.

[24] Yi Jin and Trevor M. Penning. Aldo-keto reductases and bioactivation/detoxication. *Annual Review of Pharmacology and Toxicology*, 47(1):263–292, 2007. PMID: 16970545.

[25] Christopher T Walsh. The chemical versatility of natural-product assembly lines. *Acc Chem Res*, 41(1):4–10, May 2007.

[26] Mark E Horsman, Taylor P A Hari, and Christopher N Boddy. Polyketide synthase and non-ribosomal peptide synthetase thioesterase selectivity: logic gate or a victim of fate? *Nat Prod Rep*, 33(2):183–202, February 2016.

[27] James Staunton and Kira J. Weissman. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.*, 18:380–416, 2001.

[28] Uwe Rix, Carsten Fischer, Lily L. Remsing, and Jürgen Rohr. Modification of post-pks tailoring steps through combinatorial biosynthesis. *Nat. Prod. Rep.*, 19:542–580, 2002.

[29] Carlos Olano, Carmen Méndez, and José A Salas. Post-PKS tailoring steps in natural product-producing actinomycetes from the perspective of combinatorial biosynthesis. *Nat Prod Rep*, 27(4):571–616, April 2010.

[30] P Van den Steen, P M Rudd, R A Dwek, and G Opdenakker. Concepts and principles of o-linked glycosylation. *Crit Rev Biochem Mol Biol*, 33(3):151–208, 1998.

[31] L. Zhu, L. Lu, S. Wang, J. Wu, J. Shi, T. Yan, C. Xie, Q. Li, M. Hu, and Z. Liu. Chapter 11 - oral absorption basics: Pathways and physicochemical and biological factors affecting absorption. In Yihong Qiu, Yisheng Chen, Geoff G.Z. Zhang, Lawrence Yu, and Rao V. Mantri, editors, *Developing Solid Oral Dosage Forms (Second Edition)*, pages 297–329. Academic Press, Boston, second edition edition, 2017.

[32] J M Weber, J O Leung, S J Swanson, K B Idler, and J B McAlpine. An erythromycin derivative produced by targeted gene disruption in saccharopolyspora erythraea. *Science*, 252(5002):114–117, April 1991.

[33] Junjie Ji, Keqiang Fan, Xiaojing Peng, Xiuyun Tian, Meixue Dai, and Keqian Yang. [post-modification oxygenases in the biosynthesis of aromatic polyketides–a review]. *Wei Sheng Wu Xue Bao*, 50(4):444–451, April 2010.

[34] Javier González-Sabín, Roberto Morán-Ramallal, and Francisca Rebolledo. Regioselective enzymatic acylation of complex natural products: expanding molecular diversity. *Chem Soc Rev*, 40(11):5321–5335, June 2011.

[35] Mark C Walker and Michelle C Y Chang. Natural and engineered biosynthesis of fluorinated natural products. *Chem Soc Rev*, 43(18):6527–6536, September 2014.

[36] Jixun Zhan. Biosynthesis of bacterial aromatic polyketides. *Curr Top Med Chem*, 9(17):1958–1610, 2009.

[37] Ségolène Caboche, Valérie Leclère, Maude Pupin, Gregory Kucherov, and Philippe Jacques. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. *J Bacteriol*, 192(19):5143–5150, August 2010.

[38] T Verne Lee, Richard D Johnson, Vickery L Arcus, and J Shaun Lott. Prediction of the substrate for nonribosomal peptide synthetase (NRPS) adenylation domains by virtual screening. *Proteins*, 83(11):2052–2066, September 2015.

[39] Benjamin P Duckworth, Daniel J Wilson, and Courtney C Aldrich. Measurement of nonribosomal peptide synthetase adenylation domain activity using a continuous hydroxylamine release assay. *Methods Mol Biol*, 1401:53–61, 2016.

[40] Elizabeth A Felnagle, Emily E Jackson, Yolande A Chan, Angela M Podevels, Andrew D Berti, Matthew D McMahon, and Michael G Thomas. Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Mol Pharm*, 5(2):191–211, January 2008.

[41] T A Keating and C T Walsh. Initiation, elongation, and termination strategies in polyketide and polypeptide antibiotic biosynthesis. *Curr Opin Chem Biol*, 3(5):598–606, October 1999.

[42] Michael F. Byford, Jack E. Baldwin, Chia-Yang Shiau, and Christopher J. Schofield. The mechanism of acv synthetase. *Chemical Reviews*, 97(7):2631–2650, 1997. PMID: 11851475.

[43] C T Walsh. Enzymes in the d-alanine branch of bacterial cell wall peptidoglycan assembly. *J Biol Chem*, 264(5):2393–2396, February 1989.

[44] Tetsuya Miyamoto and Hiroshi Homma. D-Amino acid metabolism in bacteria. *J Biochem*, 170(1):5–13, September 2021.

[45] Christopher T Walsh, Huawei Chen, Thomas A Keating, Brian K Hubbard, Heather C Losey, Lusong Luo, C.Gary Marshall, Deborah Ann Miller, and Hiten M Patel. Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on nrps assembly lines. *Current Opinion in Chemical Biology*, 5(5):525–534, 2001.

[46] Robert V O'Brien, Ronald W Davis, Chaitan Khosla, and Maureen E Hillenmeyer. Computational identification and analysis of orphan assembly-line polyketide synthases. *J Antibiot (Tokyo)*, 67(1):89–97, December 2013.

[47] Eric J. N. Helfrich and Jörn Piel. Biosynthesis of polyketides by trans-at polyketide synthases. *Nat. Prod. Rep.*, 33:231–316, 2016.

[48] Tuanh Nguyen, Keishi Ishida, Holger Jenke-Kodama, Elke Dittmann, Cristian Gurgui, Thomas Hochmuth, Stefan Taudien, Matthias Platzer, Christian Hertweck, and Jörn Piel. Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat Biotechnol*, 26(2):225–233, January 2008.

[49] Jörn Piel. Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat Prod Rep*, 27(7):996–1047, May 2010.

[50] Eric J N Helfrich, Reiko Ueoka, Marc G Chevrette, Franziska Hemmerling, Xiaowen Lu, Stefan Leopold-Messer, Hannah A Minas, Adrien Y Burch, Steven E Lindow, Jörn Piel, and Marnix H Medema. Evolution of combinatorial diversity in trans-acyltransferase polyketide synthase assembly lines across bacteria. *Nat. Commun.*, 12(1):1422, March 2021.

[51] Eric J N Helfrich, Reiko Ueoka, Alon Dolev, Michael Rust, Roy A Meoded, Agneya Bhushan, Gianmaria Califano, Rodrigo Costa, Muriel Gugger, Christoph Steinbeck, Pablo Moreno, and Jörn Piel. Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nature Chemical Biology*, 15(8):813–821, August 2019.

[52] Marnix H Medema, Kai Blin, Peter Cimermancic, Victor de Jager, Piotr Zakrzewski, Michael A Fischbach, Tilmann Weber, Eriko Takano, and Rainer Breitling. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res*, 39(Web Server issue):W339–46, June 2011.

[53] Sean R Eddy. Accelerated profile HMM searches. *PLoS Comput Biol*, 7(10):e1002195, October 2011.

[54] Arthur L Delcher, Kirsten A Bratke, Edwin C Powers, and Steven L Salzberg. Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics*, 23(6):673–679, January 2007.

[55] Frederick A Matsen, Robin B Kodner, and E Virginia Armbrust. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11:538, October 2010.

[56] Juan I Fuxman Bass, Alos Diallo, Justin Nelson, Juan M Soto, Chad L Myers, and Albertha J M Walhout. Using networks to measure similarity between genes: association index selection. *Nat Methods*, 10(12):1169–1176, December 2013.

[57] Aurélien Macé, Zoltán Kutalik, and Armand Valsesia. Copy number variation. *Methods Mol Biol*, 1793:231–258, 2018.

[58] Salih Tuna and Mahesan Niranjan. Classification with binary gene expressions. *J. Biomed. Sci. Eng.*, 02(06):390–399, 2009.

[59] Anatoly P. Dobritsa and Mansour Samadpour. Transfer of eleven species of the genus burkholderia to the genus paraburkholderia and proposal of caballeronia gen. nov. to accommodate twelve species of the genera burkholderia and paraburkholderia. *International Journal of Systematic and Evolutionary Microbiology*, 66(8):2836–2846, 2016.

[60] Jiewei Wang, Jianli Zhang, Huancheng Pang, Yabo Zhang, Yuyi Li, and Jinping Fan. Massilia flava sp. nov., isolated from soil. *Int J Syst Evol Microbiol*, 62(Pt 3):580–585, April 2011.

[61] Kai Blin, Simon Shaw, Alexander M Kloosterman, Zach Charlop-Powers, Gilles P van Wezel, Marnix H Medema, and Tilmann Weber. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res*, 49(W1):W29–W35, July 2021.

[62] Luisa Albarano, Roberta Esposito, Nadia Ruocco, and Maria Costantini. Genome mining as new challenge in natural products discovery. 18(4), April 2020.

[63] Ivica Letunic, Tobias Doerks, and Peer Bork. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Research*, 43(D1):D257–D260, 10 2014.

[64] Arvind K Chavali and Seung Y Rhee. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. *Brief Bioinform*, 19(5):1022–1034, September 2018.

[65] Willard A. Hareland, Ronald L. Crawford, Peter J. Chapman, and Stanley Dagley. Metabolic function and properties of 4-hydroxyphenylacetic acid 1-hydroxylase from pseudomonas acidovorans. *Journal of Bacteriology*, 121:272 – 285, 1975.