Bridging the gap between genomics- and metabolomics-based perspectives of myxobacterial secondary metabolite production capability

Dissertation

zur Erlangung des Grades

des Doktors der Naturwissenschaften

der Naturwissenschaftlich-Technischen Fakultät

der Universität des Saarlandes.

von

Fabian Panter

Saarbrücken

2018

Tag des Kolloquiums: 01.04.2019

Dekan: Prof. Dr. Guido Kickelbick

Berichterstatter: Prof. Dr. Rolf Müller

          Prof. Dr. Christian Ducho

          Prof. Dr. Tobias Gulder

Vorsitz: Prof. Dr. Uli Kazmaier

Akad. Mitarbeiter: Dr. Judith Becker

Diese Arbeit entstand unter der Anleitung von Prof. Dr. Rolf Müller am Institut für Pharmazeutische Biotechnologie der Naturwissenschaftlich-Technischen Fakultät der Universität des Saarlandes von Dezember 2014 bis November 2018.

Οἶδα οὐκ εἰδώς

Ich weiß, dass ich unwissend bin.

entlehnt aus der Apologie des Sokrates, Volksgericht von Athen 399 v. Chr.

Danksagung

Vorveröffentlichungen aus dieser Dissertation

Teile dieser Arbeit wurden vorab mit Genehmigung der Naturwissenschaftlich-Technischen Fakultäten, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen veröffentlicht oder sind derzeit in Vorbereitung zur Veröffentlichung:

Publikationen

Fabian Panter, Daniel Krug, Sascha Baumann and Rolf Müller*: Self-resistance guided genome mining uncovers new topoisomerase inhibitors from myxobacteria. *Chem. Sci.*, 2018 May 3; **9**(21):4898-4908
DOI: 10.1039/C8SC01325J

Fabian Panter, Daniel Krug and Rolf Müller*: Novel methoxymethacrylate natural products uncovered by statistics-based mining of the Myxococcus fulvus secondary metabolome. *ACS chem. biol.* 2019 Jan 18; **14**(1):88-98
DOI: 10.1021/acschembio.8b00948

Jan Gorges,[1] **Fabian Panter,**[1] Louise Kjaerulff,[1] Thomas Hoffmann, Uli Kazmaier* and Rolf Müller*: Structure, Total Synthesis and Biosynthesis of Chloromyxamides - Myxobacterial Tetrapeptides Featuring an Uncommon 6-Chloromethyl-5-methoxypipecolic Acid Building Bloc. *Angew. Chem. Int. Ed. Engl.* 2018 Oct 22; **57**(43):14270-14275
DOI: 10.1002/anie.201808028

Louise Kjaerulff,[1] Ritesh Raju,[1] **Fabian Panter**, Ullrich Scheid, Ronald Garcia, Jennifer Herrmann, and Rolf Müller*: Pyxipyrrolones: Novel cytotoxic myxobacterial metabolites. Structure elucidation and biosynthesis proposal. *Angew. Chem. Int. Ed. Engl.* 2017 Aug 1; **56**(32):9614-9618
DOI: 10.1002/anie.201704790

Michael Hoffmann,[1] David Auerbach,[1] **Fabian Panter**, Thomas Hoffmann, Pieter C. Dorrestein and Rolf Müller*: Homospermidine Lipids: A Compound Class Specifically Formed during Fruiting Body Formation of *Myxococcus xanthus* DK1622. *ACS Chem. Biol.* 2018 Jan 19; **13**(1):273-280
DOI: 10.1021/acschembio.7b00816

---

[1] These Authors contributed equally to the manuscript

Marvin Meusel, Franziska Hufsky, **Fabian Panter**, Daniel Krug, Rolf Müller and Sebastian Böcker*: Predicting the Presence of Uncommon Elements in Unknown Biomolecules from Isotope Patterns. *Anal. Chem.* 2016 Aug 2; **88**(15):7556-66

DOI: 10.1021/acs.analchem.6b01015

Publikationen die nicht Teil dieser Arbeit sind

Eilien Schulz, Adriely Goes, Ronald Garcia, **Fabian Panter**, Marcus Koch, Rolf Müller, Kathrin Fuhrmann and Gregor Fuhrmann*: Biocompatible bacteria-derived vesicles show inherent antimicrobial activity. *J. Control. Release* 2018 Nov 28; **290**:46-55

DOI: 10.1016/j.jconrel.2018.09.030

Joachim J. Hug, **Fabian Panter**, Daniel Krug and Rolf Müller*: Genome mining reveals Uncommon Alkylpyrones as Type III PKS Products from Myxobacteria. *J. Ind. Microbiol. Biotechnol.* Accepted manuscript, online preview

DOI: 10.1007/s10295-018-2105-6

Domen Pogorevc, **Fabian Panter**, Carolina Schillinger, Silke C. Wenzel and Rolf Müller*: Production optimization and biosynthesis revision of Corallopyronin A, a potent anti-filarial antibiotic. Manuscript in preparation

Tagungsbeiträge

Fabian Panter, Daniel Krug und Rolf Müller **(2016)** The Fulvuthiacenes – Secondary metabolites from *Myxococcus fulvus* with an unprecedented structural feature. (Poster) **Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM) Workshop, Freiburg, Germany**

Fabian Panter, Daniel Krug und Rolf Müller **(2017)** The Fulvuthiacenes – Secondary metabolites from *Myxococcus fulvus* with an unprecedented terminal ketene acetal moiety. (Oral Communication) **Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM) Workshop, Tübingen, Germany**

Fabian Panter, Daniel Krug und Rolf Müller **(2017)** The Fulvuthiacenes – Secondary metabolites from *Myxococcus fulvus* with an unprecedented structural feature. (Poster) **VI. Symposium of the Interdisciplinary Graduate School of Natural Product Research, Saarbrücken, Germany**

Fabian Panter, Daniel Krug, Sascha Baumann und Rolf Müller **(2018)** Self-resistance guided genome mining uncovers the pyxidicycline class of myxobacterial natural products. (Poster) **30. Irseer Naturstofftage, Gesellschaft für Chemische Technik und Biotechnologie e.V. (DECHEMA), Irsee, Germany**

Fabian Panter, Daniel Krug und Rolf Müller **(2018)** Mass spectrometrical observation of the fulvuthiacene biosynthesis (Poster) **European Mass Spectrometry Conference (EMSC), Saarbrücken, Germany**

Fabian Panter, Daniel Krug, Sascha Baumann und Rolf Müller **(2018)** Self-resistance guided genome mining uncovers new topoisomerase inhibitors from myxobacteria (Oral Communication) **8th Helmholtz Institut für Pharmazeutische Forschung Saarland (HIPS) Symposium, Saarbrücken, Germany**

Fabian Panter, Daniel Krug, Sascha Baumann und Rolf Müller **(2018)** Self-resistance guided genome mining uncovers the pyxidicycline class of myxobacterial natural products (Poster) **VII. Symposium of the Interdisciplinary Graduate School of Natural Product Research, Saarbrücken, Germany**

Fabian Panter, Daniel Krug, Sascha Baumann und Rolf Müller **(2018)** Self-resistance guided Genome Mining uncovers the Pyxidicycline Class of Myxobacterial Natural Products (Oral Communication) **Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM) Workshop, Frankfurt am Main, Germany**

Fabian Panter, Daniel Krug, Sascha Baumann und Rolf Müller **(2018)** Self-resistance guided Genome Mining uncovers the Pyxidicycline Class of Myxobacterial Natural Products (Poster) **3rd European Conference on Natural Products, Gesellschaft für Chemische Technik und Biotechnologie e.V. (DECHEMA), Frankfurt am Main, Germany**

Zusammenfassung

Bakterielle Sekundärmetabolite haben oft starke biologische Aktivität und hohe „Target"-spezifität, weshalb sie wertvolle Wirkstoffleitstrukturen darstellen. Es besteht jedoch eine große Diskrepanz zwischen der Anzahl der aus Bakterien unter Laborbedingungen nachweisbaren Sekundärmetaboliten und der in ihrem Genom codierten Produktionskapazität für Naturstoffe. Sowohl die Zuordnung bakterieller Sekundärmetabolom-Bestandteile zu Biosynthesegenclustern als auch die Entdeckung neuer Kandidatensubstanzen verlangen daher nach verbesserten analytischen Methoden gekoppelt mit Genomanalysetechniken. Methoden, die auf statistischer Analyse großer Mengen hochauflösender Massenspektrometriedaten sowie automatischer Genclusterannotation beruhen, verändern derzeit die Naturstoffforschung grundlegend. Diese Arbeit beschreibt Beiträge zu Massenspektrometrie- und Genomikgeleiteter Identifizierung, Strukturaufklärung, Beschreibung von Biosyntheserouten sowie Evaluierung der Bioaktivität fünf verschiedener neuartiger Grundstrukturen von Sekundärmetaboliten aus Myxobakterien. Spezielles Augenmerk liegt dabei auf der Verbesserung der Arbeitsabläufe um Sekundärmetaboliten mittels Massenspektrometrie zu identifizieren. Die Studien zeigen, wie durch erfolgreiche Kombination mehrerer Ansätze zur Identifizierung neuer bakterieller Sekundärmetabolite und zur Beschreibung ihrer Diversität die Lücke zwischen Genomik- und Metabolomikbasierten Ansätzen geschlossen werden kann.

Abstract

Bacterial secondary metabolites often exhibit potent biological activities with precise target specificity, making them a valuable asset for the identification of drug leads. However, a striking discrepancy is regularly seen between the genome-inscribed capacity for the production of secondary metabolites and the number of known compound classes detected from many microorganisms under laboratory cultivation conditions. Methods using improved analytical approaches to connect the bacterial secondary metabolome to biosynthetic gene clusters, aiming to uncover new candidate compounds from high resolution mass spectrometric 'big data' by statistical data treatment, as well as automated secondary metabolite biosynthetic gene cluster annotation coupled to various genome mining techniques are therefore about to transform natural product discovery. This work describes contributions to mass spectrometry- and genomics-guided compound identification, structure elucidation, investigation of biosynthesis routes and biological activity evaluation covering five different myxobacterial secondary metabolite scaffolds. Special focus lies on improving workflows for secondary metabolite identification using mass spectrometry data. By successfully combining several approaches to find novel bacterial secondary metabolites, this study underlines current prospects for closing the gap between genomics- and metabolomics-centric approaches to access bacterial secondary metabolite diversity.

Table of contents

1 Introduction

Natural products, small molecules synthesized by living organisms, are an extremely diverse group of compounds that have been used by humans throughout history. The importance of natural products in human civilization is well exemplified by traditional Chinese medicine, a more than 2000 years old science that treats all kinds of illnesses with – among others – herbal remedies. Advances in technology provided the means to analyze the biologically active ingredients in these remedies, which led to a boost in pharmaceutical development.[1] While plant natural products have been used for a long time, the emergence of natural products of microbial origin was set off by the discovery of penicillin by Alexander Fleming in 1928, which also marked the starting point of antibiotics discovery. Ever since, bacterial and fungal natural products, termed secondary metabolites, have proven to be a valuable source of pharmaceutically relevant substances. These compounds are metabolites, which the host organism produces, but which are not imminently necessary for the host organism's survival.[2] From an evolutionary perspective, it is commonly thought that secondary metabolites confer a competitive advantage to the producing organism in terms of securing its habitat, e.g. by outcompeting concurring organisms as part of a permanent rivalry for food sources. Evolution has promoted the development of species that are especially fit to outcompete their contestants, which is often based on their ability to produce specific secondary metabolite for this purpose. It is thus unsurprising that many known secondary metabolites have evolved to be bactericidal, cytotoxic or to block virulence of competing species. As natural evolutionary processes have been driving the optimization of effectiveness and specificity of secondary metabolites over an extended timeframe, these compounds have also contributed many important lead structures to human medicine over several decades.[3] Especially in the field of antibiotics, secondary metabolites have been foundational for the discovery of new active scaffolds.[4] But not only antibiotics discovery has relied heavily upon natural products in order to obtain drug leads. To date, most of the clinically used drugs are either natural products, or semisynthetic compounds and synthetic compounds based on a natural product pharmacophores, revealing the large importance of natural products for drug discovery and pharmacological development.[6]
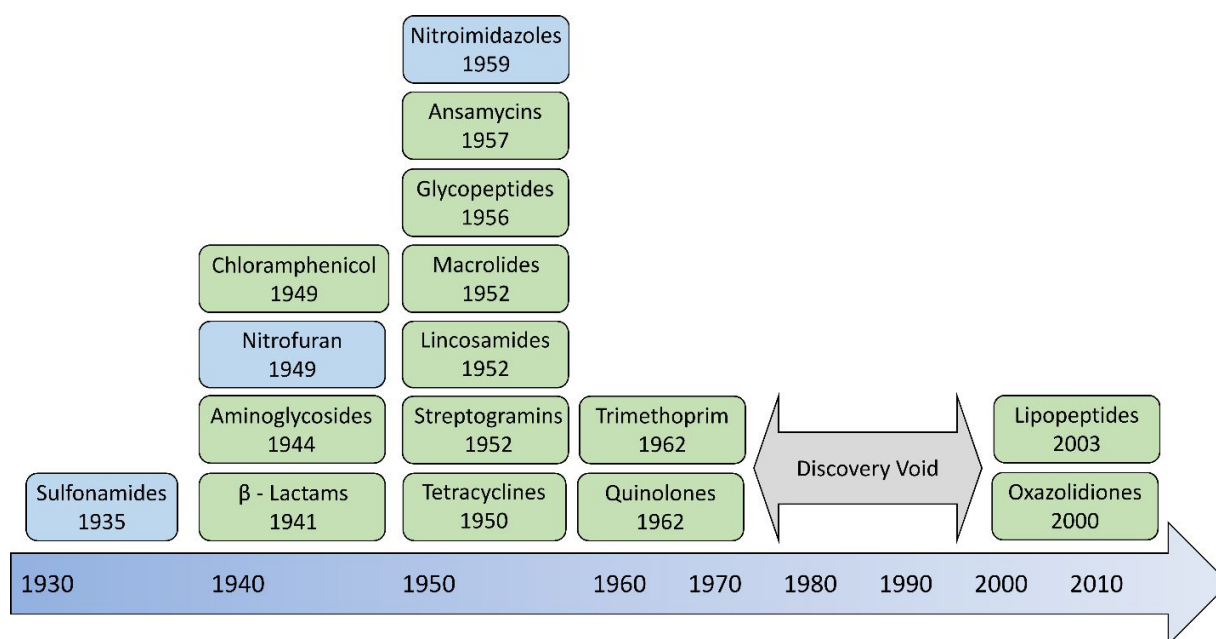
Figure 1. *Discovery of antibiotics classes over several decades: natural products derived drug scaffolds depicted in green, synthetic antibiotic scaffolds depicted in blue.*[5]

There is currently an urgent need for new antibiotics that comes with the idea to refocus antibiotics discovery towards new natural products, since, as depicted in Figure 1, very few fundamentally new antibiotics were approved by the FDA and EMA since the late 1960's, while at the same time antibiotics resistance skyrocketed.[7] This bleak picture comprises several threatening aspects. On the one hand there are Gram positive pathogen borne infections, often considered a more treatable problem compared to infections caused by their Gram negative counterparts, which have become increasingly deadly.[8] More and more of these pathogens such as the ESKAPE panel pathogens – namely *Enterococcus faecium Staphylococcus aureus Klebsiella pneumoniae Acinetobacter baumannii Pseudomonas aeruginosa* and *Enterobacter* – are resistant to many common antibiotics forcing hospitals to resort to reserve antibiotics that are often accompanied with detrimental side effects. Gram negative pathogens – often gram negative representatives of the ESKAPE panel known for nosocomial infections – are now among the main driving forces for raising fears that infective diseases might become a major cause of death by 2050.[9] Combined with efforts to improve sanitation in all parts of human life, especially in Hospital environments, new antibiotics classes are desperately needed to prevent the advent of a 'post antibiotics era' in the near future.[10] As mentioned above, most antibiotics scaffolds used today were first discovered from bacteria and fungi. This historical perspective translates into prospects that bacterial secondary metabolomes might still contain a large and diverse reservoir of new bioactive agents that could become part of the solution to the massive antimicrobial resistance problem.

1.1 Secondary metabolite production in light of species differentiation

The three main sources of biologically active secondary metabolites found in Nature are plants, fungi and eubacteria. While other sources of secondary metabolites such as archaea have been hypothesized, their production potential for novel biologically active scaffolds remains to be shown.[11] These three main secondary metabolite sources synthesize compounds that are very different in view of their structures and biogenesis. Plant metabolites are dominated by alkaloids, small molecules rich in nitrogen, that are biosynthesized using a combination of amino acids and terpene precursors.[12] Fungi on the other hand mainly rely on non-ribosomal peptides, alkaloids, terpenes and polyketides that represent their secondary metabolite arsenal.[13] Last but not least, bacteria are the biggest contributors to the known secondary metabolite diversity on earth, providing a wide variety of non-ribosomal peptides, ribosomally synthesized and post translationally modified peptides, alkaloids, terpenes and polyketides. According to the current phylogeny as well as metagenomics analyses, the vast majority of DNA existing on earth is of bacterial origin and most of the bacteria have not been cultured.[14]
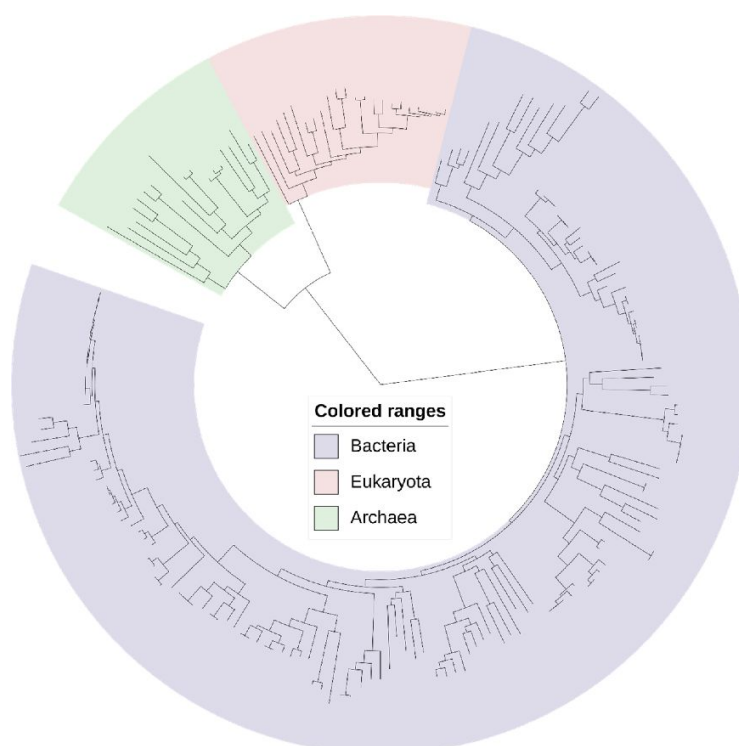
Figure 2. *Tree of life depicting the genetic diversity of Archaea, Eukaryota and Bacteria created by iTOL[15]*

Figure 2 shows that only a fraction of all different species are 'higher organisms' or part of the archaea clade. During evolution, as bacteria were competing for limited space and resources, they evolved a number of means to outcompete other bacteria such as efficient replication and fast growth as well as production of toxins and secondary metabolites. It is therefore not surprising that most of the secondary metabolite encoding biosynthetic gene clusters (BGCs) known today are of bacterial origin. Among the bacteria kingdom, some phyla have proven to be highly prolific secondary metabolite producers, such as the actinobacteria.[16] However, as the main effort to uncover novel bacterial secondary metabolites relied on this phylum, these bacteria have been extensively screened leading to decreased chances for novel findings. Nowadays, renewed efforts to detect microbial secondary metabolites therefore involve screening microorganisms found in uncommon habitats e.g. marine bacteria, plant or sponge symbionts, and bacteria from the human microbiome. These approaches are boosted by the development of genetic tools to access the 'hidden' biosynthetic potential and making use of the deeper insights offered by state of the art analytical methods.[17] One of the bacterial orders that rather recently received more attention as secondary metabolite producer are the myxobacteria.

1.2   Myxobacteria as prolific secondary metabolite producers

Myxobacteria are ubiquitous soil dwelling gram negative eubacteria belonging to the δ-proteobacteria clade. They feature "social behavior" that is unique among bacteria, such as fruiting body formation which goes along with a sophisticated cell differentiation process.[18]
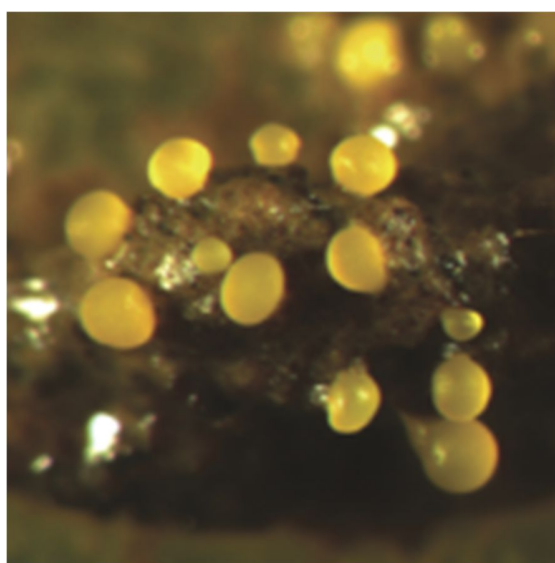
Figure 3. *Fruiting bodies of the myxobacterium* Myxococcus xanthus.[19]

Furthermore, myxobacteria of the cystobacterinae subclade are predatory organisms that thrive through feeding on other soil-dwelling microbes. Interestingly, this predation process is a cooperative effort by myxobacteria, where a deadly cocktail of antibiotic compounds and secreted enzymes allows them to collectively feed on prey organisms.[18,20]

Figure 4. *Cooperative preying behavior displayed by* Myxococcus xanthus *(left) upon an* E. coli *colony (right)*

Figure 4 shows the coordinated preying behavior of *M. xanthus* while moving over and preying on *E. coli*, macroscopically visible as rippling patterns.[21] As it has been shown for *M. xanthus* DK1622, this predatory behavior of myxobacteria is often reliant on "chemical warfare".[22] Thus it stands to reason that many of these compounds used to support prey activity exhibit strong biological activities.

1.3   **Myxobacterial secondary metabolites and their bioactivities**

Secondary metabolites that have previously been isolated from myxobacteria cover a large chemical space. Myxobacterial secondary metabolite structures belong to several classes of secondary metabolites including polyketides, non-ribosomal peptides, polyketide-non-ribosomal peptide hybrids, ribosomally produced and post-translationally modified peptides and terpenes.[23] The vast majority of known compounds from myxobacterial strains belong to the polyketide and non-ribosomal peptide secondary metabolite classes or hybrid structures thereof. Polyketide biosynthesis alone has afforded a plethora of different myxobacterial structures exhibiting various bioactivities, including cytotoxic and anti-bacterial activities as prominent examples.[24]

Myxopyronin A
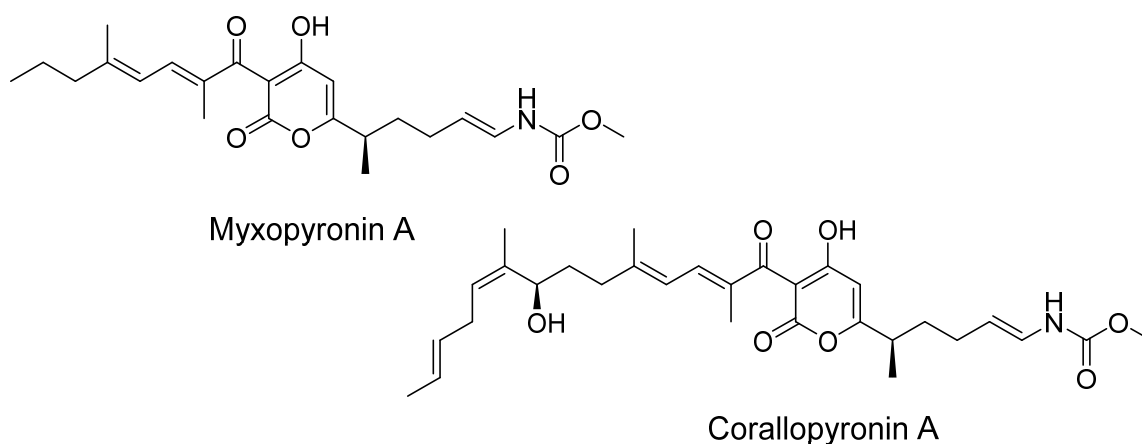
Corallopyronin A

Scheme 1. *Myxopyronin and Corallopyronin, strong α pyrone antibiotics with a selectivity for Gram positive pathogens* [25,26]

A prominent example for promising antibiotic agents from myxobacteria are the α-pyrone antibiotics of the myxopyronin and corallopyronin family. These compounds that were isolated from *Myxococcus fulvus* Mx f50 (myxopyronin) and *Corallococcus coralloides* Cc c127 (corallopyronin) are strong inhibitors of the bacterial RNA polymerase (RNAP) that act by binding to the 'switch' region of RNAP.[25,26,27] These compounds do not share the binding pocket with the most widely used RNAP inhibitor Rifampicin and are therefore still able to kill rifampicin resistant bacteria. Corallopyronin A is investigated as a drug to target sleeping sickness because the compound is able to kill the *Wolbachia* bacteria that live in symbiosis within the sleeping sickness parasite.[28]
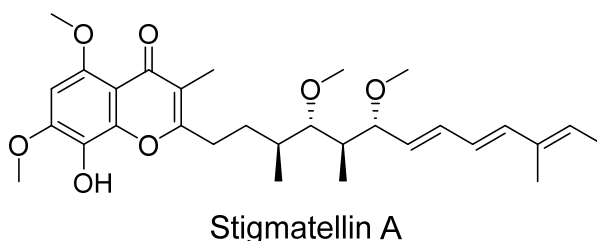
Stigmatellin A

Scheme 2. *The cytotoxin stigmatellin produced by* Stigmatella aurantiaca

Other polyketides from myxobacteria such as the stigmatellins are strong cytotoxins.[29] Stigmatellin is a polyketide that acts by blocking the cytochrome bc-1 complex.[30] Therefore, respiratory processes are inhibited which leads to quick cell death. Stigmatellins have been used as tool compounds for understanding the action of bc-1 inhibitors on the heme centers of this important protein as well as deciphering the reduction of the protein's iron-sulfur clusters once the protein is inhibited.[31]

The other rather large group of myxobacterial secondary metabolites are the non-ribosomal peptides. These types of peptidic secondary metabolites often display of unusual activities and building blocks reaching far beyond the 21 canonical amino acids. One particular example from the ever expanding panel of non-ribosomal peptides that was recently discovered from different *Cystobacter* species is mainly composed of head-to-tail linked p-aminobenzoeic acid building blocks and was thus named cystobactamid. Members of this compound class bind very strongly to the gyrase (topoisomerase II) of a wide range of gram positive and gram negative bacteria, notably without presenting any apparent cytotoxicity.[32]
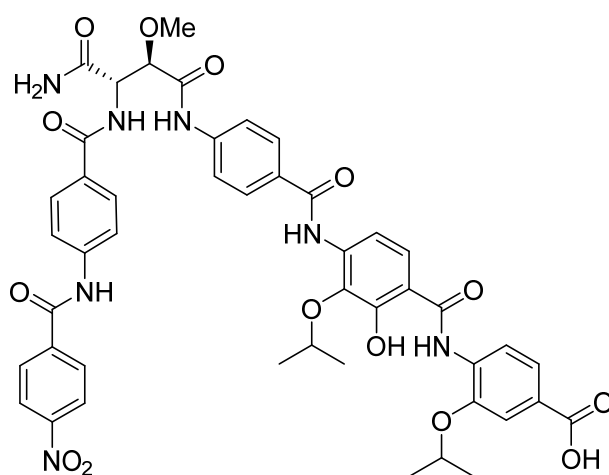
Cystobactamid 919-2

Scheme 3. Structure of the topoisomerase II inhibitor cystobactamid.

On the contrary, *Pyxidicoccus fallax* An d48 produces non-ribosomal peptides exhibiting strong cytotoxicity: The so-called tubulysins are non-ribosomal peptides belonging to a class of strong cytotoxins that act on mammalian cells – as the name suggests – by dissolution of the tubulin skeleton.[33]
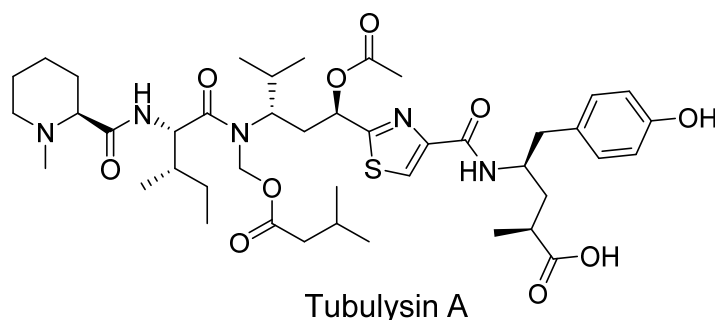
Tubulysin A

Scheme 4. Structure of the tubulin depolymerizing agent tubulysin A.

The peptide chain of the tubulysins is completely different from ribosomally produced peptides as the building blocks used for tubulysin biosynthesis are in part non-canonical amino acids such as pipecolic acid, and in part heavily modified by acylation, N-methylation or cyclisation to form the thiazole ring structure.[33,34]

Many myxobacterial secondary metabolites are in fact polyketide/non-ribosomal peptide hybrids. Among those, the most prominent candidate is epothilone, a tubulin stabilizing agent that is clinically used in chemptherapy. The corresponding drug, the aza-analogue of epothilone B, is called ixabepilone, which is marketed as Ixempra® by Bristol-Meyers Squibb.[35]

Epothilone B   Ixabepilone
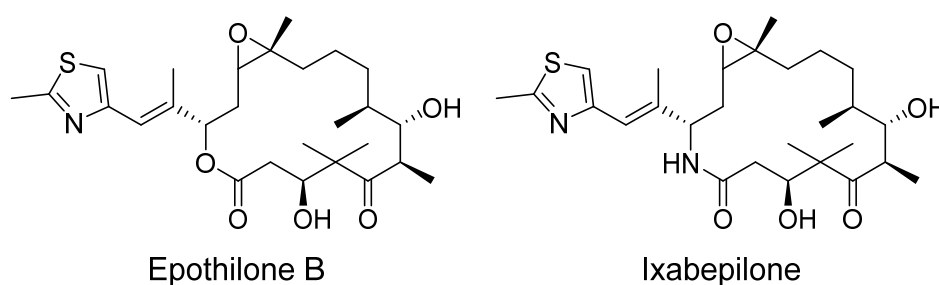
Scheme 5. *Structures s of Epothilone B and the clinically used Ixabepilone*

As this compound contains a thiazole moiety that originated from cysteine, this compound thus qualifies as a hybrid structure made from polyketide synthase and non-ribosomal peptide synthase.[36] Apart from epothilone, there is another myxobacterial compound showing strong cytotoxicity by interfering with the cytoskeleton called rhizopodin. This compound produced by *Myxococcus stipitatus* leads to cell death in eukaryotes as stabilization of the cell's actin skeleton impairs cell division.[37] While this compound was first thought to be monomeric, the compound has later been shown to consist of a C2-symmetric dimer.[38]
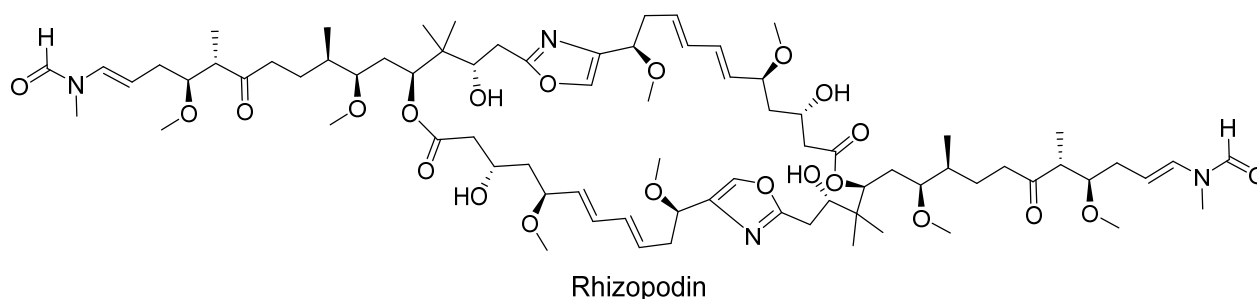
Rhizopodin

Scheme 6. *Structure of the actin stabilizing agent Rhizopodin*

While there are many examples of myxobacterial secondary metabolites exhibiting potent biological activities as listed in this paragraph, a number of myxobacterial secondary metabolites have been uncovered, where the biological target remains to be determined. One prime example is Antalid, a non-ribosomal peptide – polyketide hybrid for which no apparent biological activity has been found yet.[39]
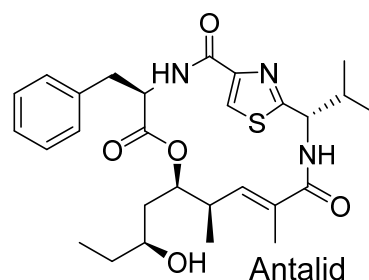
Scheme 7. *Structure of Antalid from* Polyangium spumosum.

Still, this scaffold was developed during myxobacterial evolution and it is likely that production of this secondary metabolite gave the bacterium a competitive advantage. It is therefore likely that this compound shows strong affinity for a target structure in an unknown competitor's proteome or metabolome and this secondary metabolite – target interaction – albeit as yet unrecognized - might in the future become useful for therapeutic applications.

In addition to the classical non-ribosomal peptides, polyketides and hybrid secondary metabolite structures, myxobacterial secondary metabolites also cover ribosomally produced and post-translationally modified peptides (RIPPs). The prime myxobacterial example in this case is an alkaloid structure called crocagin produced by *Chondromyces crocatus* Cm c5.[40]
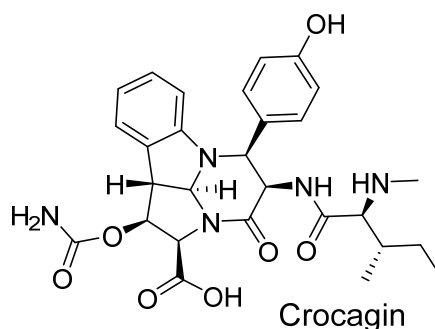
Scheme 8. *Structure of crocagin from* Chondromyces crocatus *Cm c5*

As concluded from the crocagin structure, this compound is constructed from the amino acids Ile – Tyr – Trp, which are subsequently processed through oxidative coupling reactions to form the crocagin

backbone. The structure is a strong inhibitor of the carbon storage regulator protein CsrA that is involved in pathogen virulence and is therefore under investigation as an anti-infective.[40]

Finally, the myxobacterial secondary metabolism also includes terpenes, although their discovery rate is significantly lower than the discovery rate for the compound classes described above. As terpene clusters are rather abundant in myxobacterial genomes this seems rather surprising. On the other hand, many terpenes, especially those having a low oxygen content, are hardly detectible in electrospray ionization (ESI) based mass spectrometry (MS) screening workflows while being very non-polar, which might be one of the reasons why these compounds are often missed in LC-MS screening. One of these terpenes was isolated from *Cystobacter ferrugineus* Cb fe23 and is called Cystodienoeic acid, a cytotoxic diterpene structure that exemplifies the potential of myxobacteria to synthesize uncommon terpene structures.[41]
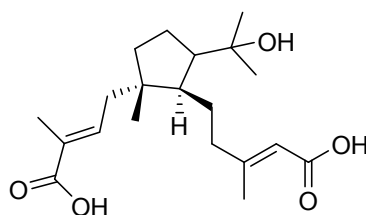
Scheme 9. *Structure of the myxobacterial terpene cystodienoeic acid*

The myxobacterial secondary metabolite structures are a chemically very diverse group of secondary metabolites that, as shown in this paragraph, span a wide range of different biological functions. All these compounds are synthesized using a small number of biological precursors like amino acids, acyl-CoA esters and isoprene pyrophosphate. To achieve synthesis of such complex molecules from these rather simple precursors, bacteria evolved a complex set of biosynthetic enzymes that catalyze the transformation of these building blocks into secondary metabolites. In bacteria, these enzymes are grouped together on the bacterial chromosome to form biosynthetic gene clusters (BGCs). The chemical diversity seen in myxobacterial natural products – as exemplified in this section – is thus also reflected by the amount and diversity of biosynthetic gene clusters in the myxobacterial genomes.

1.4   The biosynthetic proficiency of myxobacteria

Myxobacteria possess the largest bacterial genomes known to date, with chromosomes featuring up to 15 MBps. The genomes of species in the Sorangineae subclade are particularly large.[42] In secondary metabolite research, however, the number of gene clusters potentially encoding secondary metabolite

biosynthesis pathways is of higher importance than the absolute genome size, as this number reflects the theoretical secondary metabolite production capacity more accurately. The number of biosynthetic pathways per strain is mostly estimated through analysis by antiSMASH, a program that detects and annotates secondary metabolite biosynthetic enzymes using hidden markov models.[43] This tool is able to predict prototypical secondary metabolite biosynthesis pathways such as non-ribosomal peptide synthetases (NRPS) or polyketide synthases (PKS) but also ribosomally produced and post-translationally modified peptides (RiPPs) such as lantipeptides.[44] Furthermore, antiSMASH is able to predict terpene and indole biosynthetic pathways in the myxobacterial genomes.[45] These capabilities agree well with most of the compounds characterized from myxobacteria to date, which are of NRPS or PKS origin or are based on PKS-NRPS hybrid pathways. The myxobacterial pathways which are the subject of this thesis also fall into these categories, and the underlying biosynthetic machinery is thus described in more detail in the following.

1.5 Polyketide and non-ribosomal peptide biosynthesis

1.5.1 Multimodular polyketide biosynthesis

Multimodular polyketide synthases are large megaenzymes that function like an assembly line in secondary metabolite biosynthesis. These megaenzymes are organized in modules, grouping domains into functional units. A domain is the smallest functional unit of the megaenzyme that is responsible for a specific catalytic transformation. The core biosynthetic principle of polyketide biosynthesis is the construction of large molecules based on condensation of malonyl-CoA on an acyl carrier protein (ACP) bound carboxyl group which subsequently creates a 1,3-diketo moiety that is further processed. The biosynthetic step is catalyzed in a multistep process by ketosynthase (KS) and acyl transferase (AT) domains as depicted in Scheme 10. Alternative extender units such as methylmalonyl-CoA or methoxymalonyl-CoA can also be used as extender units in polyketide biosynthesis.[46] Substrate specificity for the different CoA esters as extender units is conferred by the AT domain. Polyketide synthase pathways can be divided into several groups depending on their architecture. Multimodular polyketide synthases - also termed type I polyketide biosynthesis pathways - are characterized by a modular organization of PKS modules meaning that every module catalyzes chain elongation by one malonyl-CoA unit, variably followed by different tailoring steps such as ketoreductase domains (KR) to produce an alcohol, KR and dehydratase domains (DH) to produce a double bond and finally KR, DH and an enoyl-reductase domain (ER) that produces a saturated C_2 unit.[47]
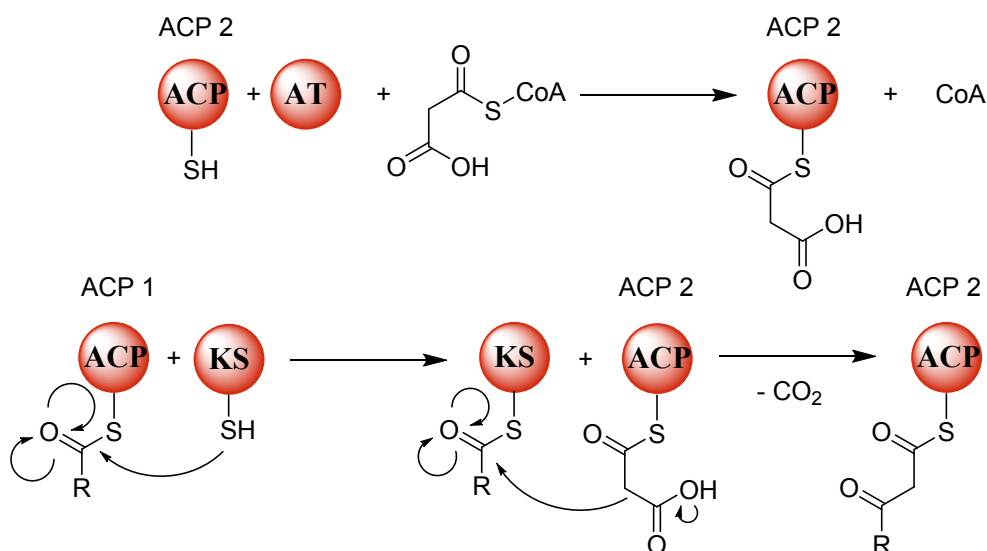
Scheme 10. *Schematic view of the decarboxylative claisen condensation that constitutes the core reaction of polyketide biosynthesis*
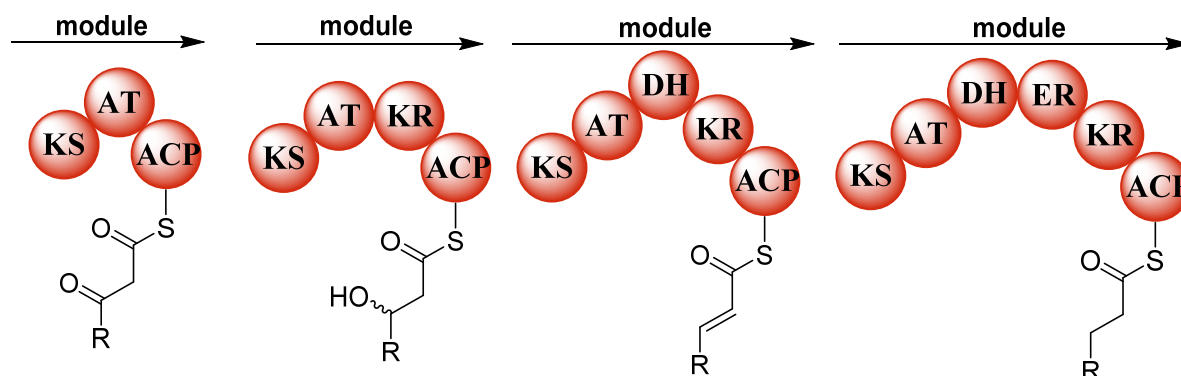
Scheme 11. *Schematic view of the different module organizations in modular type I polyketide synthases as well as the corresponding redox state of the C_2 unit that is used as an extender.*

However, not all multimodular polyketide synthases are organized strictly as depicted in Scheme 11. In comparison to these architectures that are called *cis* polyketide synthases, since the acyl transferase functionalities are encodes as domains *in-cis* on the megasynthase, the other type of modular type I polyketide synthase proteins are named as *trans* AT type I polyketide synthases.[48,49] In this case, the AT is encoded on a separate open reading frame (orf) near to the biosynthetic gene cluster. Instead of AT domains located *in cis*, these megasynthases possess so called docking domains that recognize the stand alone acyl transferase and dock the AT domain so that it is located at the right position within the multidomain enzyme complex to perform catalysis.[49]
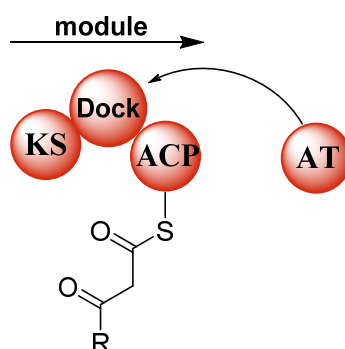
Scheme 12. *Schematic view of polyketide extension on a* trans *AT type I PKS assembly line*

1.5.2 Iterative polyketide biosynthesis

Contrary to PKS biosynthesis in modular type I PKS systems, PKS type II and type III systems work iteratively. Instead of having one module per malonyl-CoA incorporation, these types of polyketide synthases work by polykondensation of malonyl-CoA to a poly-1,3-diketone chain that is subsequently cyclized and tailored to form the final products.[50,51] PKS type II systems are characterized by the so-called minimal PKS system consisting of a catalytically active ketosynthase called KSα, a protein controlling the chain length named chain length factor (CLF) or KSβ, and the ACP that bears the growing polyketide chain. These three subunits form a tightly interacting protein complex that represents the catalytically active PKS machinery.
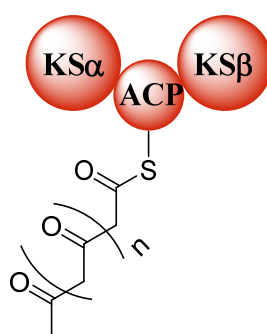
Scheme 13. *Schematic view of polyketide extension in PKS type-II systems*

After creation of this nascent polyketide chain growing on the ACP as seen in Scheme 13, whose length is determined by the KSβ, the core structure is typically aromatized as in subsequent reactions that create the polyaromatic backbone structure of PKS type-II products. Examples for this eaction sequence can be seen in the chelocardin or tetracycline biosynthesis.[52] It is worth mentioning that to date most bacterial secondary metabolites of PKS type-II origin have been reported from gram-positive bacteria while the corresponding pathways also exist in gram-negatives. Type-III PKS systems on the other hand are also

iterative like PKS type-II systems. Contrary to PKS type-II systems, these pathways rely on a single enzyme of the Chalcone/Stilbene synthase (CHS) enzyme family that is responsible for polyketide chain formation as an active homodimer.[53] Polyketide synthesis in PKS type-III systems is independent of Acyl carrier proteins and relies on a ping pong mechanism of the two CHS which pass the growing polyketide chain on in every catalysis cycle. The CHS enzyme that possesses a thiolase fold catalyzes a decarboxylative thioester claisen condensation where the growing polyketide chain remains bound to coenzyme A instead of an acyl carrier protein.[51]
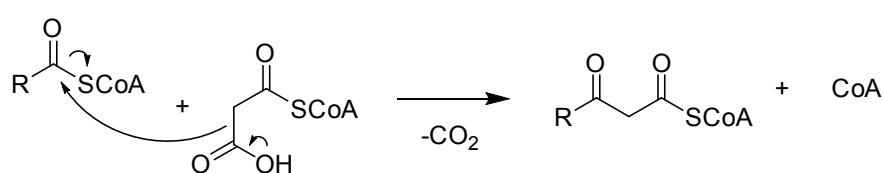
Scheme 14. *Schematic view of polyketide extension catalyzed by PKS type-III systems*

Like in PKS type-III systems, PKS type-II systems create a mostly polyaromatic core structure out of malonyl-CoA building blocks. This core structure is subsequently decorated by tailoring enzymes to synthesize the final secondary metabolite. These tailoring enzymes frequently comprise oxidase and reductase enzymes, aminotransferases, acyl transferases, methyl transferases and halogenases.[54]

1.6   **Non-ribosomal peptide biosynthesis**

The other large family of modular assembly lines other than type I PKS systems (section 1.5.1) are the non-ribosomal peptide biosynthetic enzymes. Contrary to most other peptides in living organisms, non-ribosomal peptides are not synthesized by the ribosome but instead their formation relies on catalysis mediated by megasynthetase enzymes. As described earlier, megasynthetase enzymes are grouped into functional units called modules consisting of a set of domains linked to one amino acid extension step. Generally, peptide bond formation is achieved by activation of the carboxylic acid function of an amino acid as an adenosylmonophosphatidyl ester catalyzed by the adenylation domain (A) and subsequent condensation onto the phosphopantheinyl arm of a peptidyl carrier protein (PCP) domain. Subsequently, a condensation domain (C) tethers the incoming nascent non-ribosomal peptide to the freshly bound amino acid and the amino group of said amino acid forms a peptide bond elongating the non-ribosomal peptide by one amino acid.
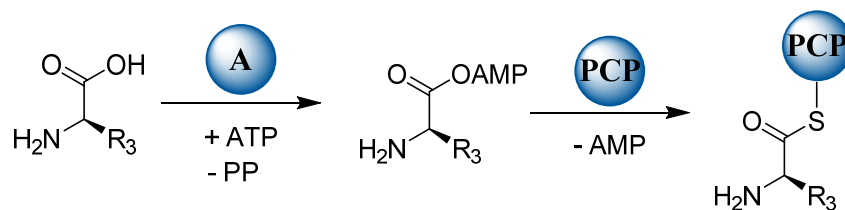
Scheme 15. *Loading of an amino acid on a peptidyl carrier protein catalyzed by an A domain*

A combination of multiple of the loading steps visible in Scheme 15 finally leads to stepwise elongation of the non-ribosomal peptide as described in Scheme 16.
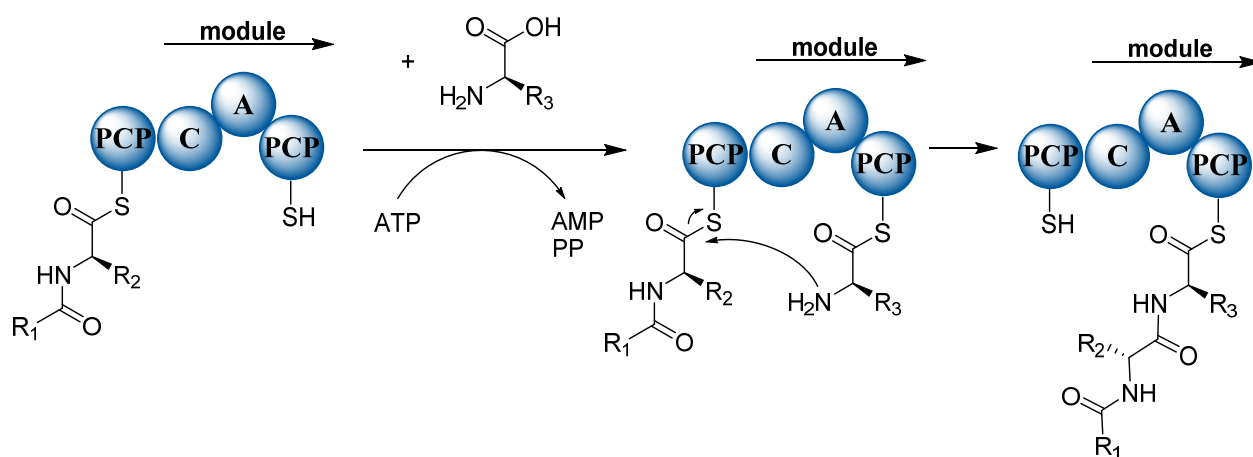
Scheme 16. *Schematic view of non-ribosomal peptide biosynthesis on the assembly line*

While non-ribosomal peptide biosynthesis is significantly slower than ribosomal peptide biosynthesis, non-ribosomal peptide synthesis opens up access to a chemical space that is a lot more diverse as it is not limited to the usage of the 21 canonical amino acids. Adenylation and condensation domains have been shown to incorporate diverse building blocks such as p-aminobenzoeic acid, β-amino acids and pipecolic acid for example; but in theory any small building block featuring an amino group as well as a carboxylic acid could be used.[32,34,55] Even more uncommon amino acids such as tubuvalin, that is incorporated in tubulysin biosynthesis or 6-chloromethyl-5-methoxy-pipecolic acid can be incorporated.[34,56] Even more variable is the incorporation of starter units that are loaded onto the first PCP domain. Biosynthesis may start with incorporation of a fatty acyl group as in cystomanamide or vioprolide biosynthesis, but also with 2,3-dihydroxybenzoeic acid as in bacillibactin biosynthesis for example.[57]

1.7   Non-ribosomal peptide - polyketide hybrid biosynthesis

Non-ribosomal peptides and polyketides both rely on biosynthetic schemes that involve condensation of an extender unit to a carboxylic acid moiety, thereby creating another terminal carboxylic acid. While polyketide biosynthesis forms an 1,3-diketone, non-ribosomal peptide biosynthesis relies on formation of a peptide bond.
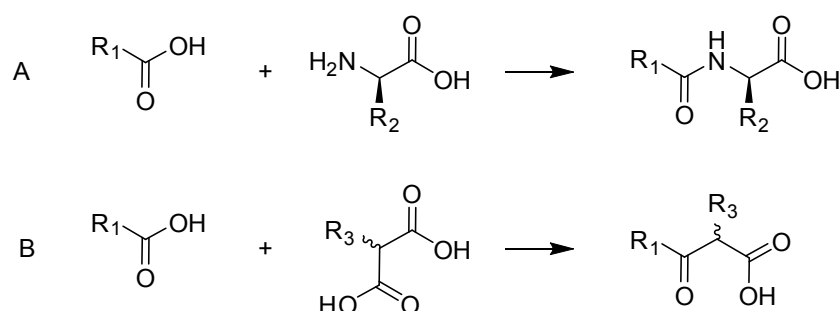
Scheme 17. *Comparative analysis of the overall reaction in NRPS and PKS biosynthesis*

Evolution has consequently come up with a number of PKS-NRPS hybrid biosynthesis pathway that combine both reaction principles. Examples for these types of biosynthesis – among many others – are antalid and epothilone that have already been discussed in section 1.4.[36,39]

The presence of these assembly lines in bacterial genomes therefore significantly further their potential to biosynthesize structurally diverse secondary metabolite molecules that are well adapted to a certain target, which gives the producer bacterium a competitive edge over its competitors. In many instances, these molecules have proven extremely valuable for human use not only to combat infections but to provide treatment options for a number of different human diseases as they bind to pharmaceutically relevant targets in the human body.

1.8   The missing link in compound to gene cluster correlation

With the significant drop in genome sequencing cost, which is largely attributed to the development of next generation sequencing techniques, the number of genomes available to the scientific community exploded rapidly.[58] Although secondary metabolite biosynthetic gene clusters are abundant in bacteria not evenly distributed across different bacteria and many bacterial families have only limited biosynthetic potential regarding secondary metabolite production, secondary metabolite production potential seems to correlate with genome size.[59] As myxobacteria have the largest genomes identified so far, this finding

underlines the importance of investigating the biosynthetic potential thereof. Still, on the one hand the vast majority of all biosynthetic gene clusters predicted so far have not been assigned to its corresponding secondary metabolite yet while on the other hand the number of gene clusters predicted by tools such as antiSMASH exceeds the number of known bacterial secondary metabolites already.[61] Although many bacterial secondary metabolite gene clusters are transcriptionally tightly controlled making it close to impossible to observe the corresponding secondary metabolite, improved analytical techniques are a key factor to close the gap between the number of biosynthetic gene clusters and the number of observed secondary metabolites.

1.9   Challenges accessing the secondary metabolome

Bacterial secondary metabolism is extremely complex. Consequently, bacterial extracts often containing more than 1000 molecular features detectable in a single LC-MS run.[60] As in many cases, only one technique is applied in order to separate and analyze the bacterial secondary metabolome, it is even likely that this number of molecular features per extract is still a grave underestimation of the fundamental analytical problem. As the so called 'low hanging fruits' of the bacterial secondary metabolome, namely secondary metabolites that are abundantly produced and easy to detect due to prominent UV and/or MS signals, are increasingly harvested already, analysis turns towards looking at compounds that are less easy to detect and isolate. Additionally, in the past many compounds have been isolated on the basis of biological activity detected in an enzyme or whole cell assay. Still, many bacterial secondary metabolites are clinically used on targets that are different from the activity that was initially found and reported. Furthermore, as many natural products have stringent target specificities, potent biological activities might be missed in 'standard' screening approaches. It seems therefore likely that many secondary metabolite scaffolds for whom no primary activity has been found initially can in fact be used on different cellular targets later. In order to discover novel secondary metabolite scaffolds potentially targeting novel structures and exhibiting novel biological activities, great efforts to utilize state of the art analytics with downstream big data processing to obtain the previously 'hidden' metabolome appear justified. Recent developments in analytical methods regarding limit of detection (LOD) and separation power led to an immense increase of detectable molecular entities. In order to transform this "well of complexity" into a feasible starting point for natural products discovery, a combination of highly standardized LC conditions combined with fast scanning high resolution MS data allows for a statistical data treatment approach that allows unprecedented insights into the myxobacterial secondary metabolome.

1.9.1   Molecular features as a valuable tool in LC-HRMS

If LC conditions are tightly controlled, a whole LC-MS chromatogram can be dissected into so called molecular features, solely characterized by exact mass and standardized retention time. The molecular feature is noise-free and combines all isotopic peaks of one molecular entity seen in the mass spectrometer and assigns it the retention time of the intensity maximum as well as the observed mass/charge ratio of the monoisotopic peak. If the molecular features present in a LC-MS run of a bacterial extract are plotted we obtain a more informative picture of the molecular entities constituting a bacterial extract compared to a view of all raw signals.
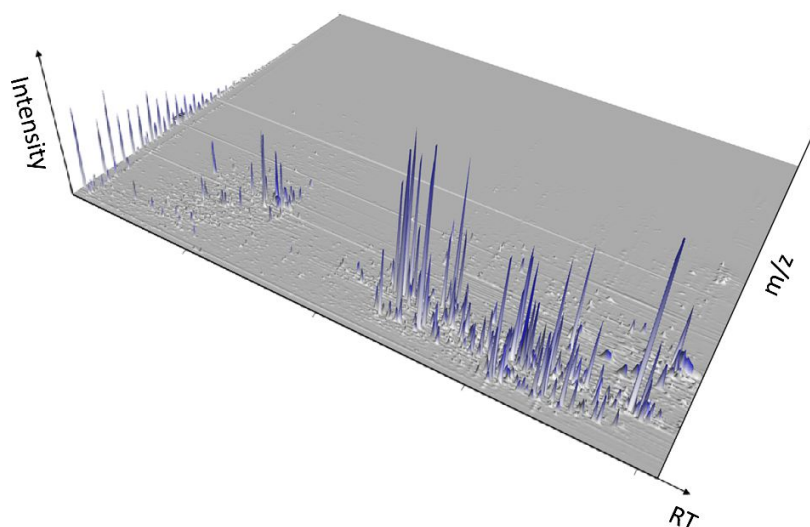
Figure 5. 3D plot of all molecular features present in a methanol extract of P. fallax *MCy10649 in a retention time-intensity-mass/charge ratio diagram*

While in former times secondary metabolite detection mainly focused on the molecular features that are visible in the base peak chromatogram (BPC), this diagram shows that there are many features 'hidden' by features that ionize well and are of higher abundance. As a word of caution, we must consider that the secondary metabolome consists of a variety of effective and specialized compounds with a high energy demand during their production, consequently we are bound to find many of them among the molecular features of low intensity. In order to mine the low abundance metabolome and to obtain an unbiased view on the myxobacterial secondary metabolome, we must introduce methods of statistical data treatment based on the molecular features so that finding new molecules becomes less a matter of serendipitous discovery.[60]

1.9.2 Bucketing to classify features in an LC-MS run

One way to make sets of molecular features accessible for statistical data treatment is the binning of these features in matrices to prepare them for comparative statistics. A technique to achieve this data binning for LC-MS data is 'bucketing'. Bucketing consists of dividing the whole RT and m/z value plane depicted in Figure 5 into rectangular subspaces of a defined m/z value width called Δ m/z and a defined RT value width called Δ RT.[60] These subspaces are called buckets. With this method, LC-MS chromatograms can be binned into matrices of all different possible buckets that get assigned the Intensity value of the molecular feature inside this bucket or zero if no molecular feature is present in this bucket.
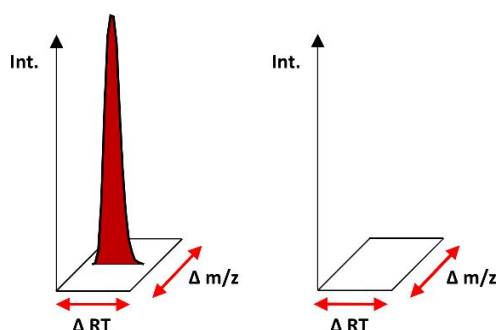
Figure 6. Schematic view of a filled LC-MS bucket (left side) and an empty LC-MS bucket (right side)

Provided the parameters of an LC-MS system are well controlled so that neither measurement of retention time nor measurement of the high resolution mass of a feature varies significantly during measurement of one sample set, the hereby created matrix, called 'bucket table' can be used to apply comparative statistics.[62]

1.9.3 Principal component analysis as a means to treat LC-MS 'big data'

As bacterial extracts often contain more than thousand molecular features detected in one single LC-MS run, it would be close to impossible to conduct meaningful analysis on these bucket tables by manual inspection of the data.[60] An extremely valuable tool to apply comparative statistics to the rather large data sets represented by a bucket table is principal component analysis (PCA).[63] PCA is a data analysis tool for tables representing observations linked by several interconnected variables.[64] This method consists of reducing the complexity of the bucket table which represents the complete LC-MS dataset into ´eigenvector´ type data subsets that are called principle components. One can now identify the principal component (PC) that accounts for most of the difference between the samples which is called PC1 and the principal component which accounts for the second biggest difference PC2 and so forth while the total number of principal components equals the number of data sets e.g. LC-MS chromatograms one started

the analysis with. In a contribution analysis one can now determine which buckets, meaning which molecular features, meaning which LC-MS observations largely contribute to PC1, indicating these molecular features can be used to distinguish analyses from each other.[63,64] Metabolomics experiments can now efficiently make use of this technique as these experiments most often rely on comparison of two or more different groups of experiments. Examples from myxobacterial secondary metabolomics where this technique was efficiently employed include complete separation of the bacterial metabolome from its medium derived matrix peaks or differential profiling of bacterial mutants and their wild types to detect metabolites associated with said genetic change.[60,62,65] One of the drawbacks of solely applying PCA is that the amount of information available about the identified metabolite linked MS signals remains limited. Provided mass spectrometry data were acquired using a high resolution mass spectrometer one can use a combination of the feature's mass accuracy and its relative isotope ratio to predict the compounds sum formula.[66] Still, with regard to chemical complexity of molecules originating from bacterial secondary metabolism, further structural information is needed in order to make educated guesses on which molecules should be followed up in a prospective isolation endaveour.

1.10   Database assisted secondary metabolite dereplication

One of the major problems of natural products research are rediscoveries of already known substances. Thus, in addition to methods to discriminate between significant molecular features and the rather noisy background, there is high demand for methods to reproducibly annotate already known molecular features. The most important determinant in this workflow are databases that store a maximum amount of information available about known secondary metabolites relevant to the study at hand. If analytical conditions are well controlled, parameters other than the molecular mass such as metabolite retention times can be used for extended secondary metabolite dereplication. If one has a comprehensive database that stores high resolution masses as well as retention time on a specific system – such as the myxobase in our case – one can use secondary metabolite annotation tools such as Bruker Compass target analysis (TA) or Bruker Metaboscape to annotate all the 'knowns' in a bacterial extract thus limiting the possibilities of rediscoveries. There are a couple of on-line tools available to facilitate this process such as DEREP-NP for example.[67]

1.11 Extracting structural similarity information from MS data

In comparison to other analytical techniques such as NMR, the big advantage of mass spectrometry lies in its sensitivity. LC-MS is able to generate limited structural information about compounds from small amounts and/or from crude mixtures. The only way to use mass spectrometry to extract structural information from mass spectrometry experiments is the application of multidimensional mass spectrometry. Multidimensional mass spectrometry in general describes the isolation of a molecular ion called precursor ion in the mass spectrometer followed by fragmentation of said ion into so called product ions. Isolation of a molecular ion followed by fragmentation and product ion analysis is called MS^2 or tandem MS while isolation of a product ion from this first fragmentation step followed by another fragmentation step is called MS^3. Fragmentation of molecules in a mass spectrometer is mainly done by colliding ionized molecules with collision gas (mostly Ar or N_2) in a collision cell or by bombarding the molecules with fast electrons. Fragmenting molecules by colliding them with a collision gas is called collision induced dissociation (CID). One of the basic principles used in multidimensional mass spectrometry is that molecules preferentially break at specific bonds in dissociation reactions creating smaller, more stable fragments. All these fragments combined are called the fragmentation pattern. It is therefore unsurprising that similar molecules present – in most cases – similar fragmentation patterns.

1.11.1 Computational metrics to assign structural information to fragment spectra

Fragmentation patterns are generally complex, as they usually comprise large numbers of mass peaks, and they are often counter-intuitive, as ion gas phase chemistry significantly differs from chemical reactions occurring in solution.[68] Thus, mass spectrometry has relied on computer tools that allow comparison of the fragmentation pattern at hand in term of their similarity. This would allow to either identify the substance behind this fragmentation pattern or assigning it to a group of similar fragmentation patterns. Met Frag for example is an web based tool that can be used to identify an unknown analyte solely using the exact mass of the parent ion and its fragmentation pattern as a peak list.[69] While the Met Frag framework has tremendous identificative power as it relies on a large number of reference spectra, the method does not assign unknown compounds to related mass spectra whose structures are already known. For the identification of previously unknown secondary metabolites, this method is thus of minor interest. Methods to increase the accuracy of the molecular formula prediction by taking sub formula of the fragments into account can help to determine the molecular formula by limiting the amount of possible sum formulas for one high resolution mass.[70] Especially for larger secondary metabolites this is of vital importance since even high resolution MS is not able to unambiguously pinpoint the molecular formula

for values above 600-700 Da. Finally, the fragmentation pattern interpretation tool of highest value for bacterial secondary metabolomics is GNPS.[71] This program relies on a metric called spectral networking. Spectral networking is a technique where spectra are compared to each other and spectra having similar fragmentation patterns are linked to each other. Thus all input fragment spectra are positioned in a network in a way that similar molecular features cluster together.
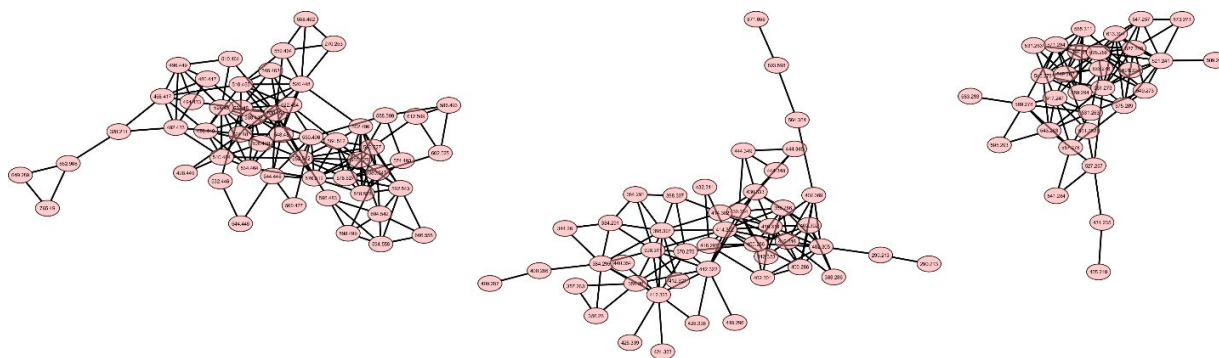
Figure 7. Example for parts of a spectral network from MS² spectra acquired from a Myxococcus xanthus *DK1622 crude extract (Every node represents the MS² spectrum of a molecular feature, every dash represents a similarity relationship based on a cosine score value of more than 0.75)*

In this network, families of structurally novel secondary metabolites would cluster together as all their derivatives fragment very similarly while overall the fragmentation pattern would be significantly different from all other fragmentation patterns. Thus, looking for outlier clusters in this spectral network is likely to uncover novel bacterial secondary metabolites with uncommon chemical structures. Therefore, a statistics based spectral networking approach to chemical screening in secondary metabolomics has significantly increased chances to find novel bacterial secondary metabolites with uncommon chemical structures and interesting biological activities than the traditional 'grind and find' approach.

1.12 Mining the 'hidden' secondary metabolome by pathway activation

Under laboratory conditions, bacteria have little incentive to produce secondary metabolites to outcompete competitors as sterile laboratory conditions aim to represent 'unnatural' quasi optimal growth conditions. Under these conditions, bacteria have a large incentive to replicate as fast as possible while a lot of external stimuli, which would trigger interaction and defense mechanism are missing in bacterial monocultures. One way to circumvent this problem is to 'force' bacteria to produce certain secondary metabolites through the pathway specific activation of secondary metabolite gene clusters.

There are several different techniques to achieve secondary metabolite BGC activation in a bacterial strain which depend on necessary genetic tools being developed to alter the strain's genome. The most conceivable approach to activate or overexpress a biosynthetic gene cluster is promotor insertion upstream of the biosynthetic gene cluster, which has been successfully applied to Myxobacteria.[62,65] Other techniques applied to activate 'cryptic' biosynthetic gene clusters in bacterial wild type strains include deletion or overexpression of pathway specific regulators or overexpression of the exporter machinery.[72]

1.13 Heterologous expression of biosynthetic gene clusters

A common problem in bacterial secondary metabolite discovery – and a point that has oftentimes been made regarding myxobacteria - is that the host strains that harbor biosynthetic potential are often hard to cultivate, intolerant to stress and inaccessible to genetic tools for targeted genetic manipulation. Moreover, many secondary metabolites, especially those found by described highly sensitive analytical techniques, are initially produced in minute quantities only. These kinds of strains are thus unlikely to be used as host strains for industrial scale fermentation, as problems with productivity, genetic stability of the strain, reaching sufficient cell density and strain growth speed are foreseeable. Thus, after identification of a promising secondary metabolite and characterization of its pathway, said secondary metabolite pathways are often transferred into so-called chassis microbes. These microbes are well adapted to transcription and translation of secondary metabolite gene clusters, rather easy to handle – as protocols for genetic manipulation, fermentation condition optimization and upscaling have already been developed – and have a streamlined metabolism that limits side product biosynthesis.[73] On an industrial scale, heterologous expression of secondary metabolites is in many cases more likely to result in a process that is able to compete with the costs of large scale chemical synthesis. This is especially true for biosynthetic gene clusters stemming from bacteria like myxobacteria that have very slow growth kinetics and thus limited space time yield. Another application of heterologous expression in secondary metabolite discovery is the activation of 'silent' secondary metabolite pathways. The activation of the 'hidden' secondary metabolite potential offers additional potential to obtain secondary metabolites that are not produced in the wild type.[65] Even for rather well-characterized bacteria such as the myxobacterial model strain *Myxococcus xanthus* DK1622, of the 18 predicted NRPS-PKS pathways encoded in the genome only 5 have been to date connected to their corresponding metabolites.[74] As cloning of biosynthetic genes from one organism and transferring these assembly lines to one of these chassis hosts extracts these BGCs from their natural regulatory network, it is likely that these assembly lines which were tightly controlled

in their host organisms are producing their corresponding secondary metabolites in the heterologous host.[73] This does not only extend to 'silent' biosynthetic gene clusters, one can also apply this technique to biosynthetic gene clusters of so far 'uncultured' bacteria such as symbionts or bacteria only known from metagenomics analyses. Modern heterologous expression techniques allow mining this previously hidden secondary metabolite potential. *E. coli* based Rec/ET direct cloning and *S. cerevisiae* based transformation associated recombination (TAR) of secondary metabolite gene clusters are two techniques that allow efficient cloning of larger genomic regions as for example a biosynthetic gene cluster.

1.13.1 Direct cloning of biosynthetic gene clusters

Direct cloning of biosynthetic gene clusters describes the technique to cut an entire BGC out of its locus in the genomic DNA followed by a recombineering step to recombine the ends of said BGC with homologous ends on a designed capture vector which leads to a synthetic vector harboring the entire BGC in the strain corresponding to the capture vector. The other commonly used recombineering technique is yeast based TAR, which allows direct cloning of larger fragments as yeast cells are more adapted to handling of larger DNA fragments.[75]
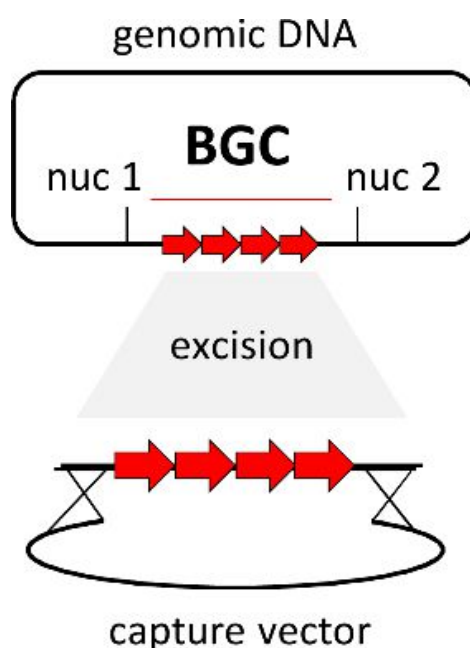
Figure 8. Schematic view of a direct cloning experiment for a BGC flanked by the restriction endonucleases nuc 1 and nuc 2 into a capture vector using a recombineering technique

First, restriction endonucleases that cut in the flanking regions to the desired biosynthetic gene cluster are employed to free the desired biosynthetic gene cluster from the bulk DNA. Additionally, so-

called anchor sequences that are inversely homologous to the genomic region one wants to clone are put on a so-called capture vector. Both the digested genomic DNA as well as the capture vector are then transformed into an *E. coli* strain harboring the *redα, redβ* and *recE, recT, redγ, recA* under separate promotors.[76] Both of these gene sets can be used to perform Red/ET recombineering, but while for linear to circular homologous recombination e.g. for plasmid modification the red proteins have shown to have good recombination activity, linear to linear direct cloning of entire gene clusters seems to work more efficiently using the rec genes.[77] This technique was successfully employed for direct cloning of a wide range of biosynthetic gene clusters as for example the alkylpyrones, syringolins or salinomycins.[78]

1.13.2 PCR based TAR cloning for secondary metabolite cluster assembly

A new recombineering technique that emerged, which is especially useful for assembly and heterologous expression of smaller biosynthetic gene clusters, is recombineering based assembly of PCR products spanning a whole biosynthetic gene cluster in yeast.
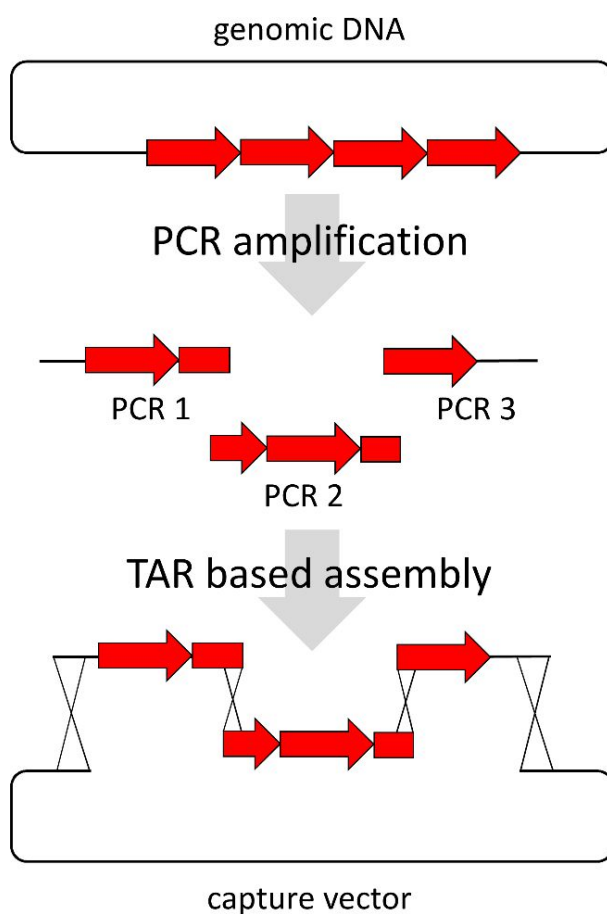
Figure 9. Schematic view of the protocol for PCR based transformation associated recombination to assemble biosynthetic gene clusters

Contrary to *E. coli* strains used in Rec or Red ET experiments, yeast cells have high recombinase activity. Therefore it is possible to co-transform several PCR products with overhangs into said yeast which will subsequently assemble the heterologous expression vector in a series of recombination events as can be seen in Figure 9.[65,79] So in total, one can assemble a variety of different PCR products with overlapping ends in one transformation step to the final expression vector. The big advantage of this approach is the possibility to reorganize a BGC during the assembly step including the possibility of introducing foreign DNA into the gene cluster during the assembly. Connectivity of the fragments during the recombination steps is guided by homology arms introduced via the primers as can be seen in Figure 9. Furthermore one has the ability to assemble DNA from multiple origins by solely adapting the overhangs necessary for recombination. That way, direct cloning of a biosynthetic gene cluster combined with gene cluster reorganization and gene cluster manipulation e.g. promotor exchanges is possible using only one transformation step.[65] Still, this technique has significant drawbacks, as on the one hand PCR amplification of the complete cluster can lead to SNPs or frame shifts in the assembled biosynthetic gene cluster, while this recombineering technique is inherently bad for the assembly of BGCs with repetitive homologous sequences. These repeats of homologous genetic regions, which occur for example in PKS or NRPS modules that extend a secondary metabolite with the same building block, might be recognized by the recombinases and parts of these megaenzymes might be crossed out. To assemble such biosynthesis pathways containing these repetitive sequences, dummy sequences can be used during the assembly process that are later replaced by the repetitive elements using conventional cloning.

1.14   Outline of this work

In this work, myxobacterial secondary metabolism is approached from two sides. Firstly, bacterial secondary metabolism is investigated from a genetic point of view in order to evaluate the genome-encoded potential for the production of secondary metabolites, aimed at connecting selected BGCs to yet unknown compounds. Secondly, a stringent chemical-analytical perspective is applied in order to gain access to the fraction of compounds from the secondary metabolome that are actually produced and seizable. Although myxobacterial genomes and secondary metabolomes are fundamentally interconnected, their connection is neither as straightforward nor as well studied as the interconnection of genes and proteins. It is therefore an additional objective of this work to improve the genome- and metabolome-mining toolbox to further strengthen our capability for both the "gene-to-compound" as well as the "compound first" discovery approach (Figure 10).
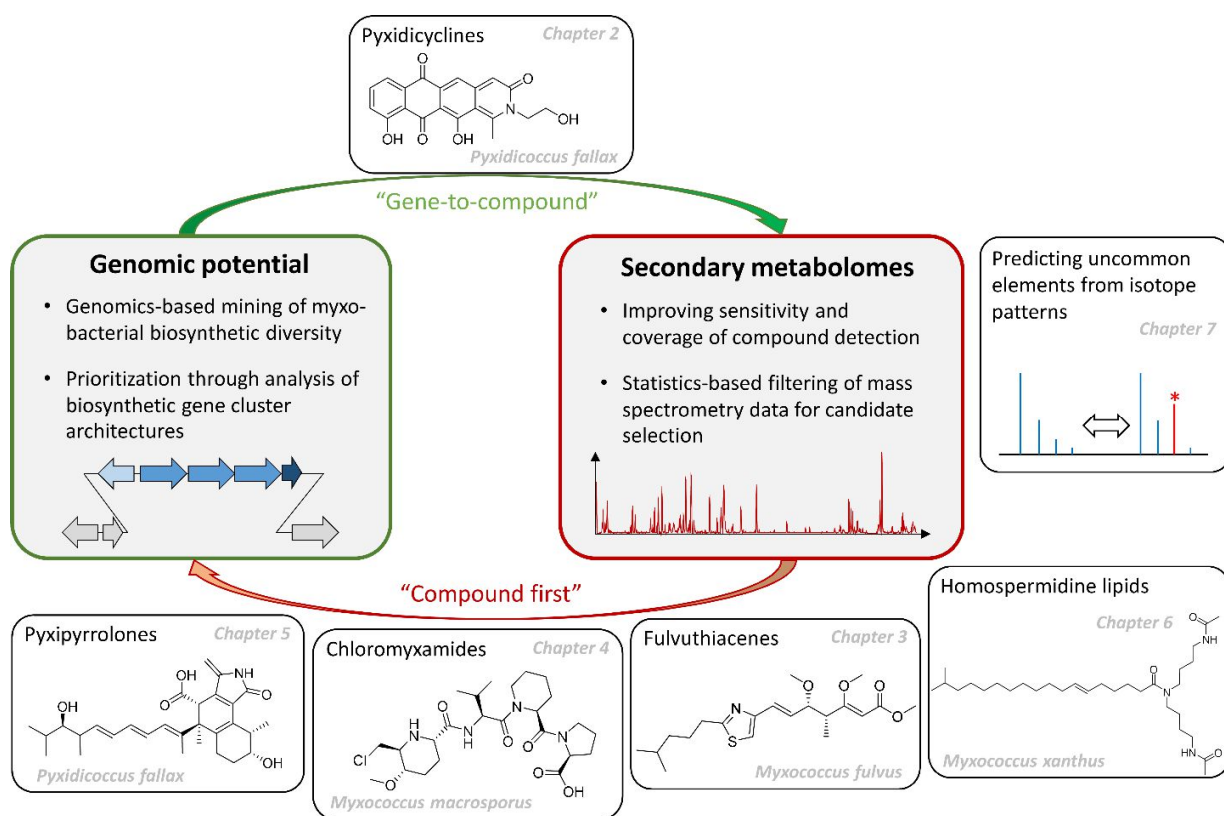
***Figure 10**. Representation of the different parts of this thesis as part of the secondary metabolome mining workflow.*

By combining *in silico* analysis, genetic manipulation techniques, state of the art mass spectrometry amended by statistical data treatment and also the full complement of methods for natural product isolation and structural characterization, this work highlights the tremendous chances for the discovery of novel natural products offered by myxobacteria throughout the following six chapters. The study on pyxidicyclines (chapter 2) exemplifies the genomics-driven discovery of a new myxobacterial secondary metabolite class ultimately revealed as potent topoisomerase inhibitors: *in silico* based gene cluster prioritization is carried out based on self-resistance determinants encoded within the BGC and statistics-based mass spectrometry is employed to distinguish myxobacterial mutant- from wild type-metabolomes to facilitate compound identification, using both BGC activation as well as targeted gene inactivation. In the study leading to discovery und full structure elucidation of fulvuthiacenes (chapter 3), the focus is on improving methods to identify secondary metabolite families from bacterial crude extracts including the assignment of minor derivatives, based on high resolution mass spectrometry and also exploiting tandem-MS data for tentative compound family identification. In line with the high value of high-resolution MS data for mining the secondary metabolomes, a method is described (chapter 7) to automatically predict the presence of uncommon elements solely from isotope patterns, which enhances the amount of information that can be drawn from any high resolution MS spectrum. This technique can become a

discovery tool in itself, as it enhances prospects for finding novel uncommon myxobacterial secondary metabolites - exemplified by the discovery of new chlorine-containing tetrapeptides called the chloromyxamides from *Myxococcus* (chapter 4). Furthermore, this work contains contributions to the discovery of two additional new secondary metabolite families from myxobacteria: the cytotoxic pyxipyrrolones (chapter 5) and the homospermidine lipids, a compound class specifically formed during fruiting body formation of *Myxococcus xanthus* (chapter 6).

The work on pyxidicyclines, fulvuthiacenes, chloromyxamides, pyxipyrrolones and the homospermidine lipids is accompanied by elucidation of the corresponding biosynthesis, leading in each case to a biosynthesis model that explains structural peculiarities found in these scaffolds. After dissecting these structures using retrobiosynthetic reasoning, *in silico* methods for generalized analysis of biosynthetic gene clusters are first used to approach the biosynthetic architecture of a secondary metabolite biosynthesis pathway and analysis is refined with predictions reliant on sequence or 3D alignment to provide hints for the enzymatic functions of tailoring enzymes involved in compound maturation. Moreover, the results from targeted mutagenesis in the producer strains, results from feeding experiments using stable isotope labelled precursors as well as heterologous expression are presented to underpin the molecular basis for compound formation.

1.15   References in the Introduction

[1]   D. Normile, *Science (New York, N.Y.)* **2003**, *299*, 188.

[2]   R. D. Firn, C. G. Jones, *Mol. Microbiol.* **2000**, *37*, 989.

[3]   G. M. Cragg, D. J. Newman, *Biochimica et biophysica acta* **2013**, *1830*, 3670.

[4]   D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2012**, *75*, 311.

[5]   J. Bérdy, *J. Antibiot. (Tokyo)* **2012**, *65*, 441.

[6]   D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2016**, *79*, 629.

[7]   a) L. Freire-Moran, B. Aronsson, C. Manz, I. C. Gyssens, A. D. So, D. L. Monnet, O. Cars, *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy* **2011**, *14*, 118; b) B. Spellberg, M. Blaser, R. J. Guidos, H. W. Boucher, J. S. Bradley, B. I. Eisenstein, D. Gerding, R. Lynfield, L. B. Reller, J. Rex et al., *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **2011**, *52 Suppl 5*, S397-428.

[8]   *Antimicrobial Resistance: Global report on surveillance*.

[9]   a) H. W. Boucher, G. H. Talbot, J. S. Bradley, J. E. Edwards, D. Gilbert, L. B. Rice, M. Scheld, B. Spellberg, J. Bartlett, *Clin Infect Dis* **2009**, *48*, 1; b) WHO, *Antimicrobial Resistance - Global Report on*

Surveillance 2014; c) *Reducing risks, promoting healthy life*. *The world health report 2002*, World Health Organization (WHO), Geneva, **2002**.

[10] B. Spellberg, *Critical care (London, England)* **2014**, *18*, 228.

[11] J. C. Charlesworth, B. P. Burns, *Archaea (Vancouver, B.C.)* **2015**, *2015*, 282035.

[12] P. J. Facchini, *Annu. Rev. Plant Physiol. Plant Molec. Biol.* **2001**, *52*, 29.

[13] N. P. Keller, G. Turner, J. W. Bennett, *Nat. Rev. Microbiol.* **2005**, *3*, 937.

[14] a) T. J. Sharpton, S. J. Riesenfeld, S. W. Kembel, J. Ladau, J. P. O'Dwyer, J. L. Green, J. A. Eisen, K. S. Pollard, *PLoS computational biology* **2011**, *7*; b) P. Yarza, P. Yilmaz, E. Pruesse, F. O. Glockner, W. Ludwig, K.-H. Schleifer, W. B. Whitman, J. Euzeby, R. Amann, R. Rossello-Mora, *Nat. Rev. Microbiol.* **2014**, *12*, 635.

[15] I. Letunic, P. Bork, *Nucleic acids research* **2016**, *44*, W242-5.

[16] L. Katz, R. H. Baltz, *J. Ind. Microbiol. Biotechnol.* **2016**, *43*, 155.

[17] a) P. Manivasagan, J. Venkatesan, K. Sivakumar, S.-K. Kim, *Microbiological research* **2014**, *169*, 262; b) J. J. Hug, C. D. Bader, M. Remškar, K. Cirnski, R. Müller, *Antibiotics* **2018**, *7*, 44.

[18] J. Munoz-Dorado, F. J. Marcos-Torres, E. Garcia-Bravo, A. Moraleda-Munoz, J. Perez, *Front. Microbiol.* **2016**, *7*, 781.

[19] *PLoS Biol* **2005**, *3*, e398.

[20] J. E. Berleman, J. R. Kirby, *J. Bacteriol.* **2007**, *189*, 5675.

[21] J. E. Berleman, T. Chumley, P. Cheung, J. R. Kirby, *J. Bacteriol.* **2006**, *188*, 5888.

[22] Y. Xiao, X. Wei, R. Ebright, D. Wall, *Journal of bacteriology* **2011**, *193*, 4626.

[23] J. Herrmann, A. A. Fayad, R. Müller, *Nat. Prod. Rep.* **2017**, *34*, 135.

[24] K. J. Weissman, R. Müller, *Nat. Prod. Rep.* **2010**, *27*, 1276.

[25] W. Kohl, H. Irschik, H. Reichenbach, G. Höfle, *Liebigs Ann. Chem.* **1983**, *1983*, 1656.

[26] R. Jansen, G. Höfle, H. Irschik, H. Reichenbach, *Liebigs Ann. Chem.* **1985**, *1985*, 822.

[27] a) H. Irschik, R. Jansen, G. Höfle, K. Gerth, H. Reichenbach, *J. Antibiot.* **1985**, *38*, 145; b) H. Irschik, K. Gerth, G. Höfle, W. Kohl, H. Reichenbach, *J. Antibiot.* **1983**, *36*, 1651; c) J. Mukhopadhyay, K. Das, S. Ismail, D. Koppstein, M. Jang, B. Hudson, S. Sarafianos, S. Tuske, J. Patel, R. Jansen et al., *Cell* **2008**, *135*, 295.

[28] T. F. Schäberle, A. Schiefer, A. Schmitz, G. M. König, A. Hoerauf, K. Pfarr, *International journal of medical microbiology : IJMM* **2014**, *304*, 72.

[29] a) G. Höfle, B. Kunze, C. Zorzin, H. Reichenbach, *Liebigs Ann. Chem.* **1984**, *8*, 1883; b) B. Kunze, T. Kemmer, G. Höfle, H. Reichenbach, *J. Antibiot.* **1984**, *37*, 454.

[30] J. P. Di Rago, J. Y. Coppee, A. M. Colson, *J. Biol. Chem.* **1989**, *264*, 14543.

[31] a) M. Ritter, H. Palsdottir, M. Abe, W. Mäntele, C. Hunte, H. Miyoshi, P. Hellwig, *Biochemistry* **2004**; b) B. Gurung, Y. Linda, C. A. Yu, *Biophys. J.* **2007**, 501.

[32] S. Baumann, J. Herrmann, R. Raju, H. Steinmetz, K. I. Mohr, S. Hüttel, K. Harmrolfs, M. Stadler, R. Müller, *Angew. Chem. Int. Ed.* **2014**, *53*, 14605.

[33] H. Steinmetz, N. Glaser, E. Herdtweck, F. Sasse, H. Reichenbach, G. Höfle, *Angew. Chem. Int. Ed. Engl.* **2004**, *43*, 4888.

[34] A. Sandmann, F. Sasse, R. Müller, *Chem. Biol.* **2004**, *11*, 1071.

[35] a) K.-H. Altmann, *The epothilones. An outstanding family of anti-tumour agents : from soil to the clinic / authors, K.H. Altmann … [et al.]*, Springer, Wien, **2009**; b) G. Höfle, N. Bedorf, H. Steinmetz, D. Schomburg, K. Gerth, H. Reichenbach, *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 1567; c) E. Rivera, J. Lee, A. Davies, *Oncologist* **2008**, *13*, 1207.

[36] J. Mulzer, K. H. Altmann, G. Hofle, R. Müller, K. Prantz, *Comptes Rendus Chimie* **2008**, *11*, 1336.

[37] F. Sasse, H. Steinmetz, G. Höfle, H. Reichenbach, *J. Antibiot.* **1993**, *46*, 741.

[38] G. Hagelueken, S. C. Albrecht, H. Steinmetz, R. Jansen, D. W. Heinz, M. Kalesse, W. D. Schubert, *Angew. Chem. Int. Ed. Engl.* **2009**, *48*, 595.

[39] T. Tautz, J. Hoffmann, T. Hoffmann, H. Steinmetz, P. Washausen, B. Kunze, V. Huch, A. Kitsche, H. Reichenbach, G. Höfle et al., *Org. Lett.* **2016**, *18*, 2560.

[40] K. Viehrig, F. Surup, C. Volz, J. Herrmann, A. Abou Fayad, S. Adam, J. Kohnke, D. Trauner, R. Müller, *Angew. Chem. Int. Ed.* **2017**.

[41] R. Raju, K. I. Mohr, S. Bernecker, J. Herrmann, R. Müller, *J Antibiot (Tokyo)* **2015**, *68*, 473.

[42] a) K. Han, Z. F. Li, R. Peng, L. P. Zhu, T. Zhou, L. G. Wang, S. G. Li, X. B. Zhang, W. Hu, Z. H. Wu et al., *Sci. Rep.* **2013**, *3*, 1; b) S. Pradella, A. Hans, C. Sproer, H. Reichenbach, K. Gerth, S. Beyer, *Arch. Microbiol.* **2002**, *178*, 484; c) S. Schneiker, O. Perlova, O. Kaiser, K. Gerth, A. Alici, M. O. Altmeyer, D. Bartels, T. Bekel, S. Beyer, E. Bode et al., *Nat. Biotechnol.* **2007**, *25*, 1281.

[43] a) T. Weber, K. Blin, S. Duddela, D. Krug, H. U. Kim, R. Bruccoleri, S. Y. Lee, M. A. Fischbach, R. Müller, W. Wohlleben et al., *Nucleic Acids Res.* **2015**, *43*, W237-W243; b) K. Blin, H. U. Kim, M. H. Medema, T. Weber, *Briefings in bioinformatics* **2017**.

[44] a) P. D. Cotter, C. Hill, R. P. Ross, *Current Protein & Peptide Science* **2005**, *6*, 61; b) D. E. Cane, C. T. Walsh, *Chem. Biol.* **1999**, *6*, R319-R325.

[45] K. Gerth, S. Pradella, O. Perlova, S. Beyer, R. Müller, *J. Biotechnol.* **2003**, *106*, 233.

[46] a) A. A. Hunaiti, P. E. Kolattukudy, *Antimicrobial agents and chemotherapy* **1984**, *25*, 173; b) S. C. Wenzel, R. M. Williamson, C. Grünanger, J. Xu, K. Gerth, R. A. Martinez, S. J. Moss, B. J. Carroll, S. Grond, C. J. Unkefer et al., *J. Am. Chem. Soc.* **2006**, *128*, 14325.

[47] S. Dutta, J. R. Whicher, D. A. Hansen, W. A. Hale, J. A. Chemler, G. R. Congdon, A. R. Narayan, K. Hakansson, D. H. Sherman, J. L. Smith et al., *Nature* **2014**, *510*, 512.

[48] J. Piel, *Nat. Prod. Rep.* **2010**, *27*, 996.

[49] E. J. N. Helfrich, J. Piel, *Natural product reports* **2016**, *33*, 231.

[50] a) C. Hertweck, A. Luzhetskyy, Y. Rebets, A. Bechthold, *Nat. Prod. Rep.* **2007**, *24*, 162; b) Y. Katsuyama, Y. Ohnishi, *Complex Enzymes in Microbial Natural Product Biosynthesis, Part B: Polyketides, Aminocoumarins and Carbohydrates* **2012**, *515*, 359.

[51] F. Gross, N. Luniak, O. Perlova, N. Gaitatzis, H. Jenke-Kodama, K. Gerth, D. Gottschalk, E. Dittmann, R. Müller, *Arch. Microbiol.* **2006**, *185*, 28.

[52] a) T. Lukežič, U. Lešnik, A. Podgoršek, J. Horvat, T. Polak, M. Šala, B. Jenko, P. Raspor, P. R. Herron, I. S. Hunter et al., *Microbiology* **2013**, *159*, 2524; b) L. B. Pickens, Y. Tang, *J. Biol. Chem.* **2010**, *285*, 27509.

[53] M. B. Austin, J. P. Noel, *Nat. Prod. Rep.* **2003**, *20*, 79.

[54] a) K. H. van Pee, E. P. Patallo, *Appl. Microbiol. Biotechnol.* **2006**, *70*, 631; b) F. Boissier, F. Bardou, V. R. Guillet, S. Uttenweiler-Joseph, M. Daffe, A. Quemard, L. Mourey, *J. Biol. Chem.* **2006**, *281*, 4434; c) A. Binter, G. Oberdorfer, S. Hofzumahaus, S. Nerstheimer, G. Altenbacher, K. Gruber, P. Macheroux, *FEBS J.* **2011**, *278*, 4122; d) F. Surup, K. Viehrig, K. I. Mohr, J. Herrmann, R. Jansen, R. Müller, *Angew. Chem. Int. Ed. Engl.* **2014**, *49*, 13588.

[55] S. Rachid, D. Krug, K. J. Weissman, R. Muller, *The Journal of biological chemistry* **2007**, *282*, 21810.

[56] J. Gorges, F. Panter, L. Kjaerulff, T. Hoffmann, U. Kazmaier, R. Müller, *Angew. Chem. Int. Ed. Engl.* **2018**.

[57] a) L. Etzbach, A. Plaza, R. Garcia, S. Baumann, R. Müller, *Org. Lett.* **2014**, *16*, 2414; b) F. Yan, D. Auerbach, Y. Chai, L. Keller, Q. Tu, S. Hüttel, A. Glemser, H. A. Grab, T. Bach, Y. Zhang et al., *Angew. Chem. Int. Ed. Engl.* **2018**; c) J. J. May, T. M. Wendrich, M. A. Marahiel, *The Journal of biological chemistry* **2001**, *276*, 7209.

[58] J. Shendure, H. Ji, *Nat. Biotechnol.* **2008**, *26*, 1135.

[59] P. Cimermancic, M. H. Medema, J. Claesen, K. Kurita, Wieland Brown, Laura C, K. Mavrommatis, A. Pati, P. A. Godfrey, M. Koehrsen, J. Clardy et al., *Cell* **2014**, *158*, 412.

[60] T. Hoffmann, D. Krug, S. Hüttel, R. Müller, *Anal. Chem.* **2014**, *86*, 10780.

[61] K. Blin, V. Pascal Andreu, E. L. C. de Los Santos, F. Del Carratore, S. Y. Lee, M. H. Medema, T. Weber, *Nucleic acids research* **2018**.

[62] N. S. Cortina, D. Krug, A. Plaza, O. Revermann, R. Müller, *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 811.

[63] N. Kuhnert, R. Jaiswal, P. Eravuchira, R. M. El-Abassy, B. von der Kammer, A. Materny, *Anal. Methods* **2011**, *3*, 144.

[64] H. Abdi, L. J. Williams, *WIREs Computational Statistics* **2010**, *2*, 433.

[65] F. Panter, D. Krug, S. Baumann, R. Müller, *Chem. Sci.* **2018**, *9*, 4898.

[66] M. Meusel, F. Hufsky, F. Panter, D. Krug, R. Müller, S. Böcker, *Anal. Chem.* **2016**, *88*, 7556.

[67] C. L. Zani, A. R. Carroll, *Journal of natural products* **2017**, *80*, 1758.

[68] A. R. Dongré, J. L. Jones, Á. Somogyi, V. H. Wysocki, *J Am.Chem.Soc.* **1996**, *118*, 8365.

[69] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, S. Neumann, *Journal of cheminformatics* **2016**, *8*, 3.

[70] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12580.

[71] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan et al., *Nat. Biotechnol.* **2016**, *34*, 828.

[72] L. Huo, S. Rachid, M. Stadler, S. C. Wenzel, R. Müller, *Chem. Biol.* **2012**, *19*, 1278.

[73] T. Beites, M. V. Mendes, *Frontiers in microbiology* **2015**, *6*, 906.

[74] S. C. Wenzel, R. Müller, *Mol. Biosyst.* **2009**, *5*, 567.

[75] N. Kouprina, V. Larionov, *Nat. Protoc.* **2008**, *3*, 371.

[76] H. Wang, Z. Li, R. Jia, Y. Hou, J. Yin, X. Bian, A. Li, R. Muller, A. F. Stewart, J. Fu et al., *Nat. Protoc.* **2016**, *11*, 1175.

[77] J. Fu, X. Bian, S. Hu, H. Wang, F. Huang, P. M. Seibert, A. Plaza, L. Xia, R. Müller, A. F. Stewart et al., *Nat. Biotechnol.* **2012**, *30*, 440.

[78] a) X. Bian, F. Huang, F. A. Stewart, L. Xia, Y. Zhang, R. Müller, *ChemBioChem* **2012**, *13*, 1946; b) J. Yin, M. Hoffmann, X. Bian, Q. Tu, F. Yan, L. Xia, X. Ding, A. F. Stewart, R. Müller, J. Fu et al., *Sci. Rep.* **2015**, *5*, 15081; c) J. J. Hug, F. Panter, D. Krug, R. Müller, *J Ind Microbiol Biotechnol* **2018**, *accepted manuscript*.

[79] O. Bilyk, O. N. Sekurova, S. B. Zotchev, A. Luzhetskyy, *PLoS ONE* **2016**, *11*, e0158682.

Chapter 2

Self-resistance guided genome mining uncovers new topoisomerase inhibitors from myxobacteria

Fabian Panter, Daniel Krug, Sascha Baumann and Rolf Müller*

Affiliation

Department Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmaceutical Biotechnology, Saarland University, Campus E8.1, 66123 Saarbrücken, Germany

Contributions and Acknowledgements

Author's effort:

The author significantly contributed to the conception of this study, designed and performed experiments, evaluated and interpreted resulting data. The laboratory and *in silico* work regarding gene cluster analysis, isolation and structure elucidation of the Pyxidicyclines, except acquisition and interpretation of the X-ray crystallography data, as well as the cloning and heterologous expression efforts were performed by the author. *In vitro* evaluation of the Pyxidicyclines' topoisomerase inhibition potential was also performed by the author. Furthermore, the author contributed significantly to conceiving and writing this manuscript.

Contributions by others:

Daniel Krug contributed to conception and supervision of this study, contributed to conceiving, writing and editing of the manuscript. Sascha Baumann contributed through identification of the target biosynthetic gene cluster in the genome of *P. fallax* And48 by performing preliminary *in silico* analyses. Rolf Müller contributed by supervision of the project and conceiving, editing and proofreading of the manuscript.

2.1   Abstract

There is astounding discrepancy between the genome-inscribed production capacity and the set of known secondary metabolite classes from many microorganisms as detected under laboratory cultivation conditions. Genome-mining techniques are meant to fill this gap, but in order to favor discovery of structurally novel as well as bioactive compounds it is crucial to amend genomics-based strategies with selective filtering principles. In this study, we followed a self-resistance guided approach aiming at the discovery of inhibitors of topoisomerase, known as valid target in both cancer and antibiotic therapy. A common host self-defense mechanism against such inhibitors in bacteria is mediated by so-called pentapeptide repeat proteins (PRP). Genes encoding the biosynthetic machinery for production of an alleged topoisomerase inhibitor were found on the basis of their collocation adjacent to a predicted PRP in the genome of the myxobacterium Pyxidicoccus fallax An d48, but to date no matching compound has been reported from this bacterium. Activation of this peculiar polyketide synthase type-II gene cluster in the native host as well as its heterologous expression led to the structure elucidation of new natural products that were named pyxidicyclines and provided an insight into their biosynthesis. Subsequent topoisomerase inhibition assays showed strong affinity to - and inhibition of - unwinding topoisomerases such as E. coli topoisomerase IV and human topoisomerase I by pyxidicyclines as well as precise selectivity, since E. coli topoisomerase II (gyrase) was not inhibited at concentrations up to 50 μg/ml.

2.2   Introduction

Natural products of bacterial origin have continuously provided drug leads to combat infectious diseases and cancer.[1] However, as certain groups of microorganisms have been screened extensively for natural products, the discovery of novel secondary metabolite scaffolds becomes increasingly challenging. Recent approaches to find valuable novel and biologically active natural products include screening of underexploited bacterial species, genera and families in different culture conditions as well as improving the bioinformatics, genetic and analytical toolset to access the "hidden" secondary metabolome of respective microorganisms.[2–4] As the quest for new bioactive compounds is thriving, combinations of these methods are increasingly applied to uncover previously unidentified classes of natural products from a range of microbes, including – amongst others - marine streptomycetes, plant endosymbionts, cyanobacteria and the myxobacteria.[3,4,5] Myxobacteria are ubiquitous Gram-negative, soil dwelling bacteria that exert a unique life-style involving the formation of multicellular fruiting bodies.[6] Among myxobacteria the suborder *Cystobacterineae* - to which the genus *Pyxidicoccus* investigated in this work

belongs - is typically represented by predatory strains displaying a high level of sophisticated multi-cellular coordination and cooperation.[6] The complex life-style is reflected by large genome sizes exceeding 10 Mbps as observed from many myxobacteria.[7] Moreover, the sheer number of biosynthetic gene clusters present in myxobacterial genomes according to bioinformatics analysis suggests myxobacteria constitute a highly prolific source of natural products. Nonetheless, the number of characterised compound families is still strikingly lower than expected from the genome-encoded capacity.[8] Even for myxobacterial model strains such as *Myxococcus xanthus* DK1622 or *Stigmatella aurantiaca* DW4/3-1 only a fraction of the more than 15 secondary metabolite biosynthesis pathways found in their genomes have been assigned a natural product to date, indicating that the corresponding natural products are absent or below detection limits under standard laboratory conditions.[8,9] Whereas several reasons might account for this observable discrepancy, it has emerged as a common picture for secondary metabolite-producing microorganisms. A range of genome-mining approaches has consequently been devised in order to "harvest" the untapped genome-encoded promise for novel natural products.[10,11,12] One straight-forward concept to bring to light the natural products supposedly originating from biosynthetic pathways that are repressed under standard laboratory conditions employs gene cluster activation via insertion of a heterologous promotor upstream of biosynthesis genes.
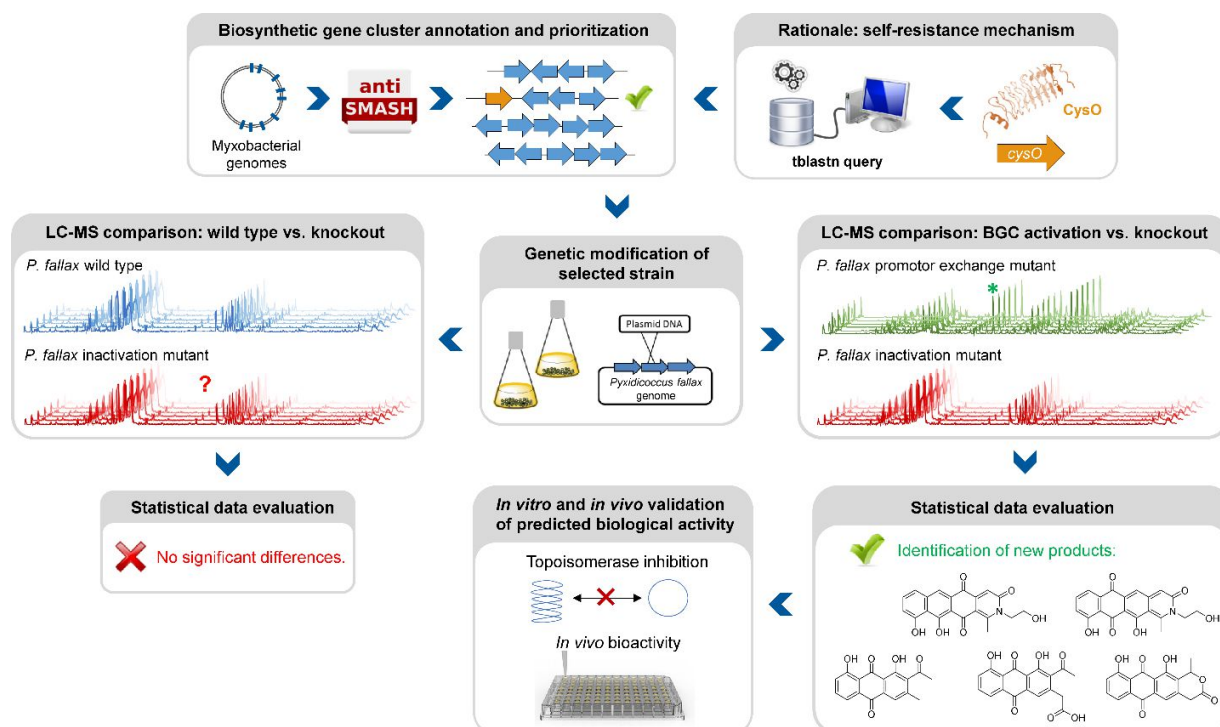
Figure 1. *Workflow scheme illustrating the genomics-based gene cluster prioritization and activation approach towards novel bioactive natural products from 'silent' gene clusters.*

This approach however is typically laborious and time-consuming. Success is critically depending on pre-established cultivation conditions, an understanding of the regulation of the respective pathway and genetic manipulation protocols specific to the target organism. A strong rationale is therefore required to prioritise the plethora of eligible biosynthetic gene clusters for well-directed gene cluster activation. A conceivable prioritization approach is based on the finding that potential genetic self-resistance determinants are often co-located to candidate biosynthetic gene clusters.[13] Considering topoisomerase inhibitors, the occurence of a gene encoding a pentapeptide repeat protein (PRP) might serve as an indicator for the nearby BGC to potentially produce a congener of this compound class. Adopting this strategy in the current study was inspired by a specific example recently reported from myxobacteria, i.e. the cystobactamid gene cluster where a pentapeptide repeat protein is responsible for self-resistance against the gyrase inhibitor cystobactamid, even allowing for simplified mode of action determination of this compound class.[14]

Along these lines, we report here the identification and full structural characterization of a new bioactive secondary metabolite family connected to a biosynthetic gene cluster not previously attributed to any natural product from the myxobacterium *Pyxidicoccus fallax* An d48. In agreement with the rationale used for their discovery, the new polyketide-type compounds named pyxidicyclines A (**1**) and B (**2**) display potent inhibition of bacterial and human topoisomerases. In addition, the heterologous expression of the underlying *pcy* gene cluster permits insights into the intriguing biosynthesis of a new myxobacterial secondary metabolite class featuring an unprecedented tetracene quinone scaffold.

2.3   Results and Discussion

2.3.1   Genome mining for topoisomerase inhibitors

The starting point of this study is defined by our aim to find novel topoisomerase inhibitors connected to matching BGCs hidden in the genomes of myxobacteria and guided by signature genes likely to encode the required self-resistance mechanism. We reasoned that CysO from the cystobactamide pathway in *Cystobacter ferrugineus* and AlbG from the albicidin pathway in *Xanthomonas albilineans* should be suitable templates to identify putative pentapeptide repeat proteins conferring self-resistance to topoisomerase poisons.[14,15] Searching for genes encoding topoisomerase-targeting pentapeptide repeat protein (TTPRP) sequences similar to AlbG and CysO yielded 31 candidate sequences among myxobacterial genomes in our in-house database.[15] Eight of those hit sequences were located close to uncharacterised biosynthetic gene clusters as predicted by antiSMASH (Figure 3).[10,11] Based on primary sequence

similarity, these genes were regarded as candidates which could plausibly be responsible for topoisomerase inhibitor self-resistance against the products afforded by said BGCs. In view of desired structural novelty of products to be discovered, our attention was drawn to a candidate encoded near a type II polyketide synthase (PKS) gene cluster present in the myxobacterial strains *P. fallax* An d30 and An d48. However, no compound resembling a type II PKS product has been found and no topoisomerase inhibitor was reported in these strains. While type II PKS compounds are well known from Gram-positive actinobacteria, there are only few examples of type II PKS products from Gram-negative bacteria as exemplified by anthraquinones from *Phodorhabdus luminescens* or aurachin from *Stigmatella aurantiaca*.[16] Therefore, the envisaged BGC was considered likely to be responsible for the biosynthesis of hitherto unknown secondary metabolites with uncommon structures. Besides the biosynthetic genes encoding polyketide synthase machinery, the *pcy* cluster features an obvious operon comprising ABC transporter genes and the 639 bp open reading frame (ORF) encoding the proposed PRP (*pcyQ-T*, Figure 3). Thus, this gene ensemble hints at a combined export- and topoisomerase protection mechanism allowing immunity of the bacterium to the hypothetical product. The proposed PRP in this cluster named PcyT was compared via MUSCLE alignment to CysO, AlbG as well as *E. coli* McbG, *M. tuberculosis* MfpA and Qnr proteins known to mediate resistance against the fluoroquinolone gyrase and topoisomerase IV inhibitor drug class (Figure 2).[17,18]
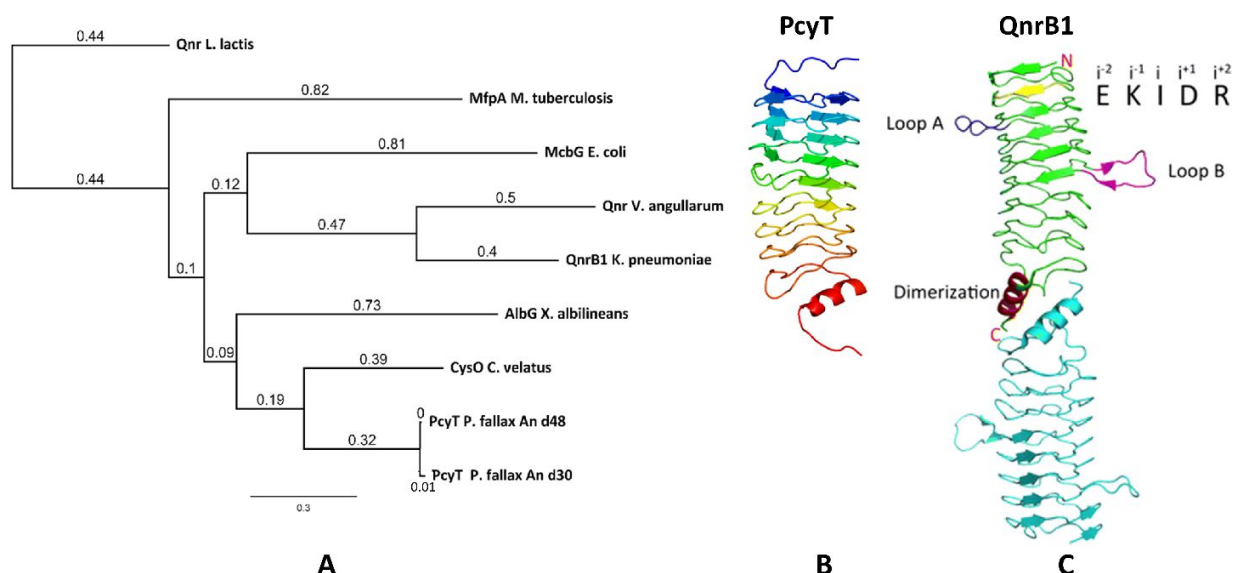
Figure 2. (A) Neighbour joining tree constructed from a MUSCLE alignment including TTPRP primary sequences known to confer topoisomerase inhibitor resistance, (B) comparison of the Phyre-2 3D structure homology model of PcyT from Pyxidicoccus fallax to (C) the protein crystal structure of the topoisomerase targeting pentapeptide repeat protein (TTPRP) QnrB1 from K. pneumoniae.

The alignment tree clearly supports functional similarities among these proteins, although amino acid sequence similarity between these short (200-300 AA) PRP is generally not higher than 50 %.[19] Their 3D structures are characterised by a consensus motif of five [S,T,A,V][D,N][L,F][S,T,R][G] repeats forming a right handed quadrilateral helix, indicating this 3D fold to be more important than the primary sequence.[19] Therefore, the 3D structure of PcyT was further analysed using the Phyre-2 homology modelling tool (section S 2.1-2.2).[20] The modelled structure of PcyT suggests the presence of the right handed quadrilateral helix structure thought to account for DNA mimicry plus a dimerization domain consisting of the alpha helical part and the terminal loop (depicted in red in Figure 2), both of which are required for mediating topoisomerase poison resistance.[15,19] The striking difference between the quinolone resistance protein QnrB1 (PDBe: 2xtx) and PcyT is the predicted absence of loop structures Loop A and Loop B suggesting PcyT falls into the category of "loopless" TTPRPs like CysO and AlbG.[19]

Thus, since *in silico* analysis suggested the possibility for production of a topoisomerase inhibitor by *P. fallax* we set out to identify the small molecules associated with said type II PKS BGC to ultimately assess their bioactivity. *P. fallax* An d48 has been the subject of natural product screening in the past and turned out to be a viable bioactive secondary metabolite producer.[21] However, no known small molecule from this strain resembles a structure likely to stem from type II polyketide synthase machinery. Furthermore, we could not detect any inhibition of *E. coli* DNA gyrase and *E. coli* DNA topoisomerase IV in assays using wild type crude extract (data not shown), pointing towards absence of a matching bioactive compound under standard laboratory conditions.

We thus took a closer look at the prioritised gene cluster (designated *pcy*) which comprises 20 ORFs encoded within three apparent operons (Figure 3). Operon 1 contains the minimal PKS genes *pcyA-I,* encoding a ketosynthase unit composed of the ketosynthase α (KSα, *pcyE*), the chain length factor (CLF or KSβ, *pcyF*) and the acyl carrier protein (ACP, *pcyG*) which are fundamental for the function of any type II PKS assembly line (Figure 3, Table 1, and KS assignments in sections S 2.3, S 2.4).[22] Operon 2 encodes proposed tailoring enzymes (*pcyJ-P)* while Operon 3 encodes ABC transporter-related genes as well as the putative resistance protein (*pcyQ-T)*.

Table 1. *ORFs within the pyxidicycline BGC, their nearest neighbours according to blastp against the nr protein data base at NCBI and putatively assigned functions (see also Table S 6).*

gene	putative function	closest annotated homologue – organism of origin
pcyA	precursor release from the acyl carrier protein	acyl-CoA thioester hydrolase – multispecies
pcyB	polyketide cyclisation (B-ring cyclisation)	polyketide cyclase – *Streptomyces nogalater*
pcyC	polyketide cyclisation (A-ring cyclisation)	polyketide cyclase – *Streptomyces atratus*
pcyD	C9' type ketoreductase	ketoacyl reductase – *Sorangium cellulosum*
pcyE	ketoacyl synthase KSα subunit	3-oxoacyl-ACP synthase II – *Dendrosporobacter quericolus*
pcyF	ketoacyl synthase KSβ subunit (chain length factor)	3-oxoacyl-ACP synthase II – *Pelosinus sp.*
pcyG	acyl carrier protein	acyl carrier protein – *Terriglobus saanensis*
pcyH	4'-phosphopanthenyl transferase	holo-ACP synthase – *Sorangium cellulosum*
pcyJ	serine incorporation	4-coumarate-CoA ligase family protein – *Modestobater species*
pcyK	heterocycle formation	muconolactone delta-isomerase – *Acidovorax valerianellae*
pcyL	heterocycle formation	polyketide cyclase – *Streptomyces species*
pcyQ	export	ABC transporter, ATP binding protein – *Myxococcus xanthus*
pcyR	export	ABC transporter permease – *Myxococcus xanthus*
pcyS	export	ABC transporter substrate binding protein – *Myxococcus xanthus*
pcyT	putative self-resistance mechanism	pentapeptide repeat containing protein – *Myxococcus fulvus*

   In order to connect a chemotype to the candidate gene cluster we initially followed a strategy combining single crossover inactivation with subsequent LC-MS profiling of 'inactivation' mutants vs. the wild type strain. Strikingly, disruption of either *pcyE* or *pcyC* located within Operon 1 did not lead to any detectable change in the *P. fallax* An d48 metabolome, as corroborated by in-depth statistical evaluation of the LC-MS chromatograms of replicate mutant extracts compared to wild type extracts (Figure 4 and sections S 3.1 to 3.4).

2.3.2   Reviving a 'silent' type II PKS gene cluster

Expecting the small molecule product of the *pcy* cluster to be absent or below the detection limit, gene cluster activation by insertion of a heterologous promotor was performed.[23] Since operon 1 contains the minimal PKS genes, promotor exchange via single crossover homologous recombination upstream of this operon should boost transcription of these genes and thus facilitate subsequent detection of the matching PKS-type metabolites. Similar pathway-refactoring strategies have been successfully employed for the activation of 'silent' gene clusters in *Streptomyces* and *Burkholderia*.[24][25] Furthermore, promotor insertions have been reported to improve secondary metabolite production by myxobacteria, although the approach did not afford a previously unknown myxobacterial natural product class to date.[2,26]
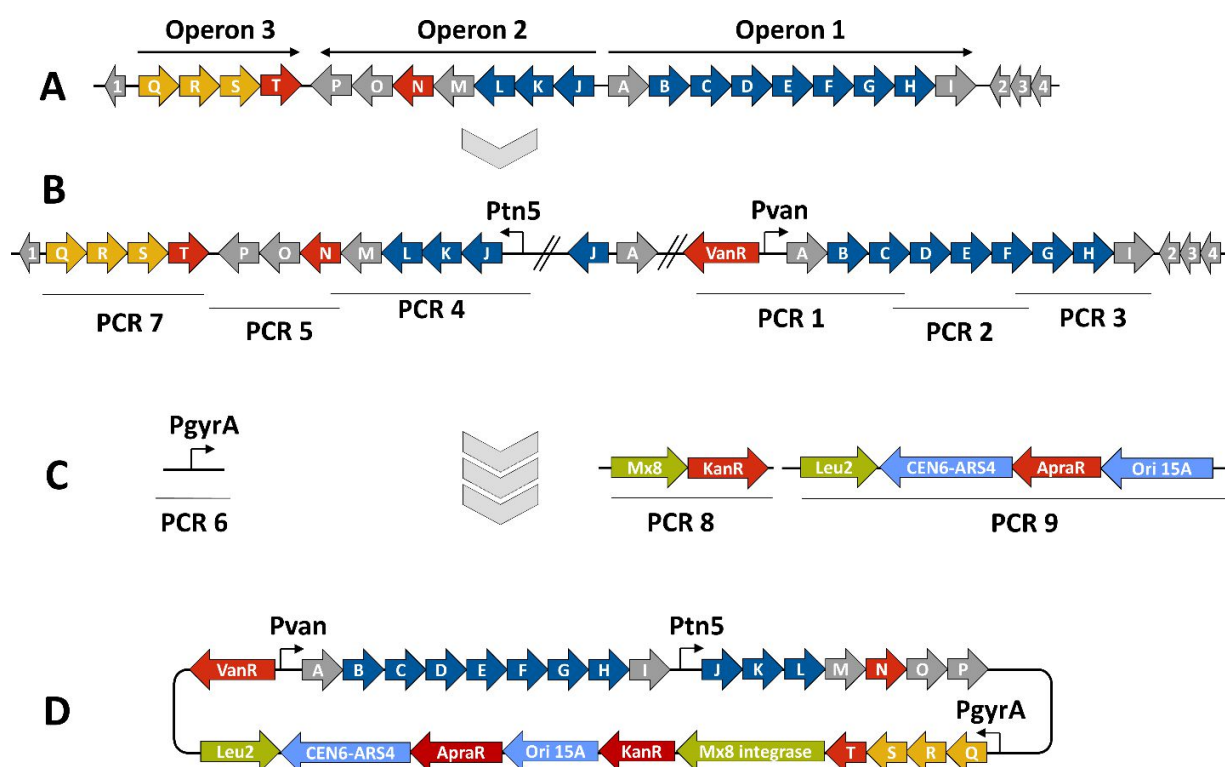
Figure 3. BGC activation and heterologous expression strategy for the pcy biosynthetic gene cluster. (A) the native gene cluster in An d48 and its operon structure, (B) the gene cluster activated by two promotor exchanges, (C) the TAR cloning strategy including PCR pieces necessary to complete the pCIY vector and (D) the assembled heterologous expression plasmid. PKS type-II biosynthesis genes depicted in blue, regulation and resistance genes in red, exporter genes in orange, auxiliary genes for the cloning strategy in light green and ORFs of unknown function in grey colour.

P. fallax An d48 mutants controlling the expression of the minimal PKS operon by a heterologous tn5 promotor and alternatively via the vanillate promotor/repressor system were created through single crossover homologous recombination using the plasmids pSBtn5 *pcyA* and pFPvan *pcyA*, respectively (section S 1.10). The vanillate promotor repressor system was chosen as it has been shown to work in myxobacteria.[27]
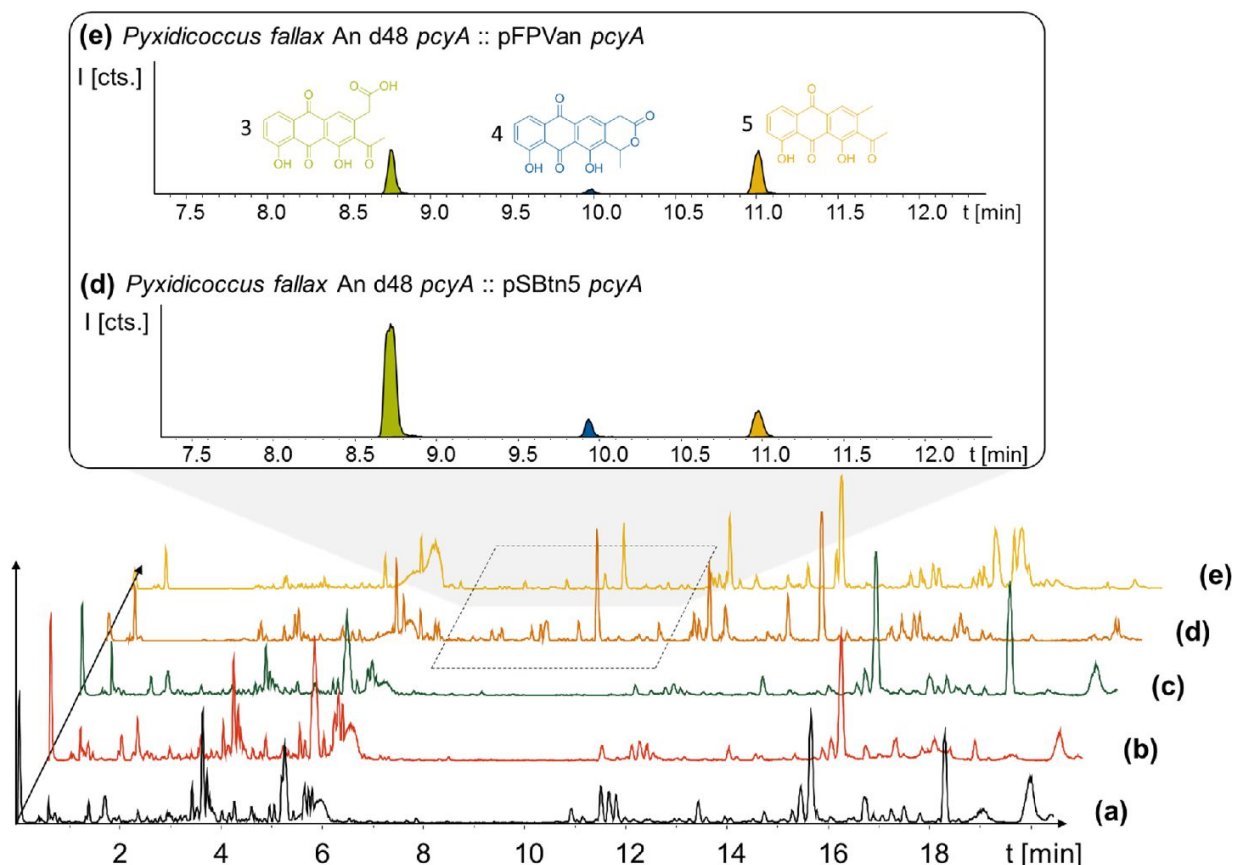
Figure 4. *BPC comparison of (**a**) Pyxidicoccus fallax An d48 wild type extract, the extracts of the (**b**) pcyC and (**c**) pcyE gene inactivation mutants, and anthraquinones produced in P. fallax An d48 (**d**) vanillate-induced and (**e**) tn5 activation mutants. The three anthraquinone metabolites 340 (**3**), 296 (**5**) and 325 (**4**) are displayed as LC-MS EICs at 341.06 [M+H]⁺ (light green), 325.07 [M+H]⁺ (dark blue) and 297.075 [M+H]⁺ (orange). The corresponding structure formula is shown in the same colour. Differences visible in traces (**a**) to (**e**) at retention times later than 12 minutes are unrelated to the investigated biosynthetic route.*

LC-MS analysis of the corresponding crude extracts revealed three newly emerging substances exhibiting strong UV absorption, indicating the presence of a chromophore as it would be expected from a type II PKS derived compound (Figure 4, Figure S 13). Their intensity is reduced to basal level shown by both LC-UV and LC-MS measurements if the An d48 *pcyA*::pFPVan *pcyA* mutant is grown without induction of the vanillate promotor, further indicating a direct link of found compounds to the target type II PKS machinery. Comparison of LC-UV and LC-MS peak areas suggests the tn5 promotor leads to higher expression levels than the induced vanillate promotor/repressor system (Table S 21). Purification of the new compounds appearing as a result of promotor integration into An d48 using a combination of chromatographic methods and NMR-based structure elucidation consequently revealed a series of anthraquinones (**3-5**) (Figure 4; section S 3.5.1 and 3.8), resembling literature-known secondary metabolites described from actinomycetes.[28,29] UV/Vis spectra and NMR signals agree well with the

previously published data (section S 3.5.1, 3.8 and 3.9). These anthraquinones show moderate in vivo activity against Gram-positive pathogens and efflux deficient *E. coli* (Table 2).

From a biosynthetic perspective **3** was described as a polyketide synthesis derailment product occurring when the frenolicin minimal PKS was complemented with the actinorhodin ketoreductase (KR) and the cyclases CYC1 and CYC2 from the griseorhodin and oxytetracycline pathways, respectively, and homologues of these genes also exist within the activated minimal PKS operon of the *pcy* gene cluster.[28] We thus reasoned that the products observed upon promotor insertion upstream of the *pcyA-I* operon could stem from incomplete processing of polyketide biosynthesis intermediates as a consequence of non-expressed tailoring enzymes located within the second operon *pcyJ-P*. Therefore, a second promotor was integrated via single crossover homologous recombination upstream of *pcyJ*, whereby this operon was activated in addition to the *pcyA-I* operon (Figure 3). The resulting 'dual promotor' mutant strain *P. fallax* An d48 *pcyA* :: pFPVan *pcyA*, *pcyJ* :: pFPtettn5 *pcyJ* (named An d48 *pcy* in short) was cultivated in small scale following successful confirmation of correct promotor integration via PCR. Notably, we found that this strain lysed upon vanillate induction when XAD-16 adsorber resin was absent from the medium, while cultivation and induction was possible with added XAD-16. Thus, the second promotor exchange apparently led to the production of biologically active entities other than the previously detected anthraquinones, and these yet unidentified products seemed to exhibit significant autotoxicity. Concomitantly, it was not possible to introduce simultaneously two constitutive tn5 promotors upstream of both operons *pcyA-I* and *pcyJ-P* despite several attempts (data not shown), further indicating autotoxicity effects associated with the observed metabolites.

2.3.3 Structure elucidation of the pyxidicycline class of natural products

Induction of the newly engineered double-promotor mutant strain with vanillate and subsequent LC-MS analysis of XAD-16 extracts led to the detection of two novel peaks with an UV absorption at 500 nm (Figure S 16). Isolation and structure elucidation of both candidate compounds (section S 3.5.2, 3.6, 3.7 and 3.9) ultimately revealed them as two novel type II PKS derived nitrogen containing tetracene quinones named pyxidicycline A (**1**) and B (**2**) (Figure 5). The two anthraquinones **3** and **5** referred to as polyketide derailment products are still present even though their yields are significantly reduced (Figure 5, Table S 21), whereas **4** could not be detected in the extract of An d48 *pcy*.
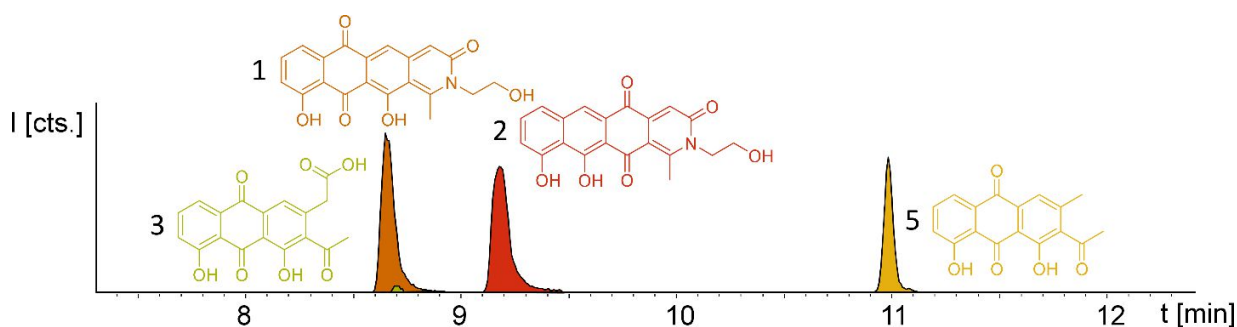
Figure 5. *Overlay of three EICs of the two novel mature type two PKS compounds (**1** and **2**) in Pyxidicoccus fallax An d48 pcy extract accompanied by two anthraquinones (**3** and **5**). **3** and **5** are displayed as LC-MS EICs at 341.06 [M+H]+ (light green) and 297.075 [M+H]+ (orange); the two pyxidicyclines are displayed as LC-MS EICs at 366.10 [M+H]+ (**1** in orange; **2** in red).*
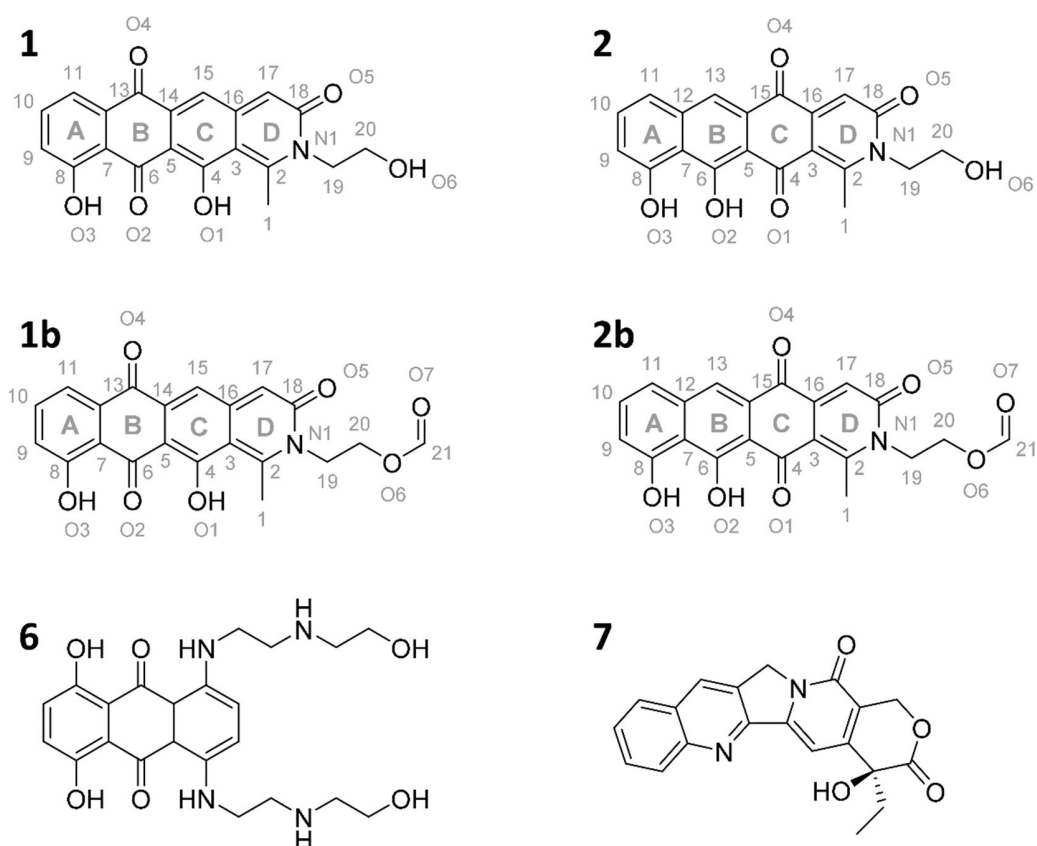
Figure 6. *Structures of the two pyxidicycline derivatives A (**1**) and B (**2**) as well as the corresponding formate esters (**1b** and **2b**) as determined by structure elucidation using NMR and X-ray crystallography. **6** and **7**, structures of mitoxanthrone and camptothecin.*[30,31]

The pyxidicyclines are difficult to dissolve in organic solvents except for DMSO and DMF, which unfortunately both degrade the compound over time. However, they are well soluble and stable in pure formic acid. Dissolution in formic acid leads to reversible formation of formate esters as shown by an increase of the compound's masses by 28 Daltons as seen in the LC-MS chromatogram. Fortunately, both isomers formed dark red crystals from formic acid solution and the corresponding crystal structures

obtained by single-crystal X-ray diffraction experiments afforded the structures depicted in Figure 6 (section S 3.6 and 3.7). The structures of both pyxidicycline formate esters were subsequently confirmed by NMR analysis in formic acid-d_2 (section S 3.11). These compounds represent a novel type II PKS scaffold, as nitrogen-containing tetracene quinone-like four ring structures have – to best of our knowledge - not been reported to date. Among the rare examples of type II PKS products containing nitrogen in their backbone structure (albeit not as part of a tetracene system) are the jadomycins and nitrogen-containing congeners of fredericamycins.[32,33]

The pyxidicyclines were subjected to antimicrobial and cytotoxic activity testing against a panel of bacterial pathogens and cancer cells. The anthraquinone precursors were tested against a reduced set since low isolation yields did not allow full bioactivity profiling. **1** and **2** show moderate antibacterial activity but strong cytotoxicity in the nano molar range (Table 2). The complete lack of activity against wild type Gram-negative pathogens is likely due to compound efflux as the efflux deficient *E. coli* ΔTolC strain shows the lowest MIC values among bacteria. While the yeast strains *C. albicans* and *P. anomala* show very high MIC values, the pyxidicyclines inhibit the growth of filamentous fungi displayed by *M. hiemalis* at intermediate concentrations (Table 2).

2.3.4   Investigating the pyxidicyclines' molecular target

Our motivation for activating the *pcy* BGC was based on the finding of pentapeptide repeat protein PcyT, which suggested the compounds afforded by this gene cluster could act as topoisomerase inhibitors. Structural similarity of pyxidicyclines to clinically used mitoxanthrone (Figure 6) - a topoisomerase II and topoisomerase IV inhibitor used in breast cancer treatment with an IC50 of 12 µM and 2 µM against *E. coli* topoisomerase II (gyrase) and topoisomerase IV, respectively [30] – increased our suspicion that pyxidicyclines might indeed target topoisomerases. Hence their topoisomerase inhibition activity towards two bacterial topoisomerases was tested, namely *E. coli* topoisomerase II (gyrase) and *E. coli* topoisomerase IV (section S 4). Contrary to the assumption based on the predicted structure of PcyT, the pyxidicyclines showed no inhibition of *E. coli* gyrase up to 60 µg/ml. We thus conclude that gyrase is not the molecular target as the *in vitro* MIC values of these substances are significantly lower (Table 2). On the other hand, topoisomerase IV was inhibited with IC50 values of 6.25-3.2 µg/ml for **1** and 3.2-1.6 µg/ml for **2** (section S 4). These IC50 values match the observed bactericidal activity against *E. coli* ΔTolC, the prokaryote in our testing panel that is most sensitive to pyxidicyclines. Therefore, topoisomerase IV inhibition is likely their main mode of action with respect to bactericidal effects. We reason that Gram-negative pathogens are likely pyxidicycline-resistant due to compound efflux, as topoisomerase IV inhibition should otherwise also kill wild type *E. coli* strains. As the pyxidicyclines exhibited pronounced

cytotoxicity in our bioassays, inhibition of human topoisomerase I was evaluated. This enzyme is the human counterpart for unwinding topoisomerases such as topoisomerase IV. **1** and **2** exert potent activity on this protein, which is an established target in anticancer research [34].

Table 2. *Minimum inhibitory concentrations (MIC) values of the pyxidicyclines A and B (**1** and **2**) and the precursor anthraquinones (**3-5**) against common pathogens and cancer cell lines.*

Test Organism	1	2	3	5	4
S. aureus Newman	16 µg/ml	32 µg/ml	2 µg/ml	16 µg/ml	n.d.
C. albicans DSM 1665	64 µg/ml	64 µg/ml	n.d.	n.d.	n.d.
M. luteus DSM 1790	16 µg/ml	32 µg/ml	n.d.	n.d.	n.d.
B. subtilis DSM 10	16 µg/ml	32 µg/ml	64 µg/ml	64 µg/ml	16 µg/ml
M. smegmatis mc2-155	32 µg/ml	64 µg/ml	n.d.	n.d.	n.d.
E. coli Δ TolC	4 µg/ml	2 µg/ml	8 µg/ml	64 µg/ml	16 µg/ml
P. aeruginosa PA 14	> 64 µg/ml	> 64 µg/ml	n.d.	n.d.	n.d.
C. violaceum DSM 30191	> 64 µg/ml	> 64 µg/ml	n.d.	n.d.	n.d.
E. coli DSM 1116	> 64 µg/ml	> 64 µg/ml	n.d.	n.d.	n.d.
M. hiemalis DSM 2656	16 µg/ml	16 µg/ml	n.d.	n.d.	n.d.
P. anomala DSM 6766	> 64 µg/ml	> 64 µg/ml	n.d.	n.d.	n.d.
HCT-116 (*Human colon carcinoma*)	0.06 µg/ml	0.03 µg/ml	n.d.	n.d.	n.d.

The IC50 values of **1** and **2** against human topoisomerase I were 1.6-0.4 µg/ml for **1** and 0.2-0.05 µg/ml for **2**, respectively. The MIC of pyxidicyclines against HCT-116 cells was determined to 0.06 µg/ml and 0.03 µg/ml for **1** and **2**, respectively, which is significantly lower than the on-target IC50 values determined in the cell free assay. This discrepancy has already been reported for the topoisomerase I inhibitor camptothecin (Figure 6),[31] which shows some structural similarity to the pyxidicyclines and this comparison also underlines topoisomerase I as the pyxidicycline's likely molecular target in mammalian cells.

Taken together, the pyxidicyclines are a novel class of type II PKS products with promising *in vitro* bioactivity based on the unprecedented nitrogen-containing tetracene quinone scaffold. Since the preferential inhibition of topoisomerase I by pxidicyclines might inspire further development of the compounds towards an anti-cancer drug, biotechnological production of the molecule is desirable. In order to pave the way for optimizing pyxidicycline production we thus decided to pursue the in-depth investigation of pyxidicycline biosynthesis. This was however met with challenges, as available genetic tools for the producer strain An d48 are limited to single-crossover recombination. We therefore sought to establish a heterologous expression system to facilitate investigation of this particular type II PKS biosynthetic pathway.

2.3.5 Heterologous expression of the *pcy* gene cluster

The myxobacterium *P. fallax* An d48 is related to model strains *M. xanthus* DK1622 and *S. aurantiaca* DW4/3-1 through membership of the same suborder *Cystobacterinae* and these two strains were thus chosen as host for heterologous expression. Since the promotor-activated *pcy* cluster in the *P. fallax* An d48 *pcy* mutant strain was already constructed, genomic DNA of this strain was used for PCR-based transformation assisted recombination (TAR) assembly of the gene cluster as described by Bylik et al. (Figure 3). [35] Through this assembly method, the whole cluster is reassembled during the cloning step and additional modifications can be engineered into the final construct by co-transformation of PCR products (Figure 3). The method also allows reorganising the native operon structure of the biosynthetic gene cluster. Thus, the minimal PKS operon (operon 1, *pcyA-I*) responsible for the creation of the polyketide chain was placed upstream of the tailoring enzyme operon (operon 2, *pcyJ-P*) likely responsible for the incorporation of an nitrogen-containing moiety into the molecule backbone (section S 1.12, Table 1) and subsequent dehydration to complete the fourth ring of the tetracyclic ring system. The putative resistance and export operon (operon 3, *pcyQ-T)* was relocated downstream of the tailoring enzyme operon. As already achieved in the An d48 double mutant, the minimal PKS system was kept under the control of the vanillate promotor/repressor system to control compound production through induction, thereby avoiding self-toxicity effects. The tailoring enzymes were subjected to control by the tn5 promotor as in the An d48 *pcy* mutant to ensure sufficient supply of tailoring enzymes for formation of mature pyxidicyclines. To boost the host's self-resistance towards the *pcy* cluster derived compounds, the putative resistance operon consisting of the ABC-transporter subunits and the pentapeptide repeat protein PcyT was placed under control of the native *P. fallax* An d48 gyrase subunit A promotor (PGyrA) to achieve strong and stable expression of the putative export and self-resistance system. That way, all *pcy* cluster related genes are transcribed in the same direction and the vanillate repressor's transcription and translation cannot interfere with the transcription of operon 2 or 3, and vice versa (Figure 3). The heterologous promotors for operon 1 and 2 were PCR amplified from the double mutant gDNA together with their gene starts, while the gyrase promotor was taken out of *P. fallax* An d48 by an additional PCR (section S 1.11). Finally, in addition to the vector described by Bilyk et al.[35] the pCIY vector backbone used in this study does not only carry two replication origins ori15A for *E. coli* and Cen-Ars6 for replication in *S. cerevisiae,* but also a yeast auxotrophy marker plus kanamycin and apramycin resistance genes. The vector additionally contains the Mx8 integrase system for later integration of the cloned construct into *S. aurantiaca* DW4/3-1 and *M. xanthus* DK1622 (Figure 3).[36] Therefore the vector assembled in yeast can be used for *M. xanthus* transformation following retransformation in *E. coli*. The assembled vector was subsequently inserted into *M. xanthus* DK1622 and *S. aurantiaca* DW4/3-1 via Mx8 integrase mediated

chromosome integration (section S 1.9.3 and 1.9.4). The *M. xanthus* DK1622 and *S. aurantiaca* DW4/3-1 attB :: attP pClY pyxidicycline cluster (*M. xanthus* DK1622 *pcy* and *S. aurantiaca* DW4/3-1 *pcy* in short) integration clones were cultivated, induced with vanillate, extracted and measured by LC-MS according to the standard protocol (section S 3.1 to 3.4). Compared to the non-induced strain we noted the appearance of two novel UV and MS signals that correspond to pyxidicycline A and B by retention time, high resolution mass and UV pattern indicating successful heterologous production of the pyxidicyclines.
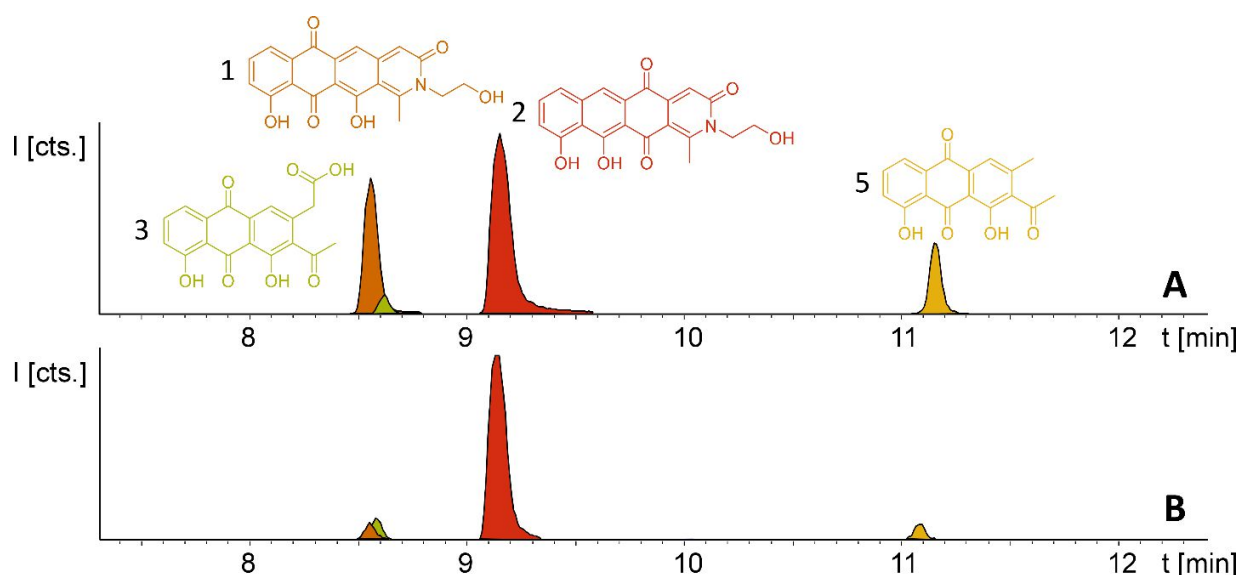
Figure 7. (**A**) Heterologous production of the pyxidicycline gene cluster in M. xanthus DK1622 pcy and (**B**) S. aurantiaca DW4/3-1 pcy. Compounds **3** and **5** are displayed as LC-MS EICs at 341.06 [M+H]⁺ (light green) and 297.075 [M+H]⁺ (orange); the two pyxidicyclines are displayed as LC-MS EICs at 366.10 [M+H]⁺ (**1** in orange; **2** in red).

Comparison of the relative LC-UV and LC-MS peak areas (Figure 7) reveals another peculiarity of this biosynthesis. The ratio of **1** to **2** shifts strongly depending on the host strain: **1** is by far the major product in the *P. fallax* An d48 *pcy* mutant while this ratio shifts to **2** in *M. xanthus* DK1622 *pcy.* Finally, *S. aurantiaca* DW4/3-1 *pcy* produces almost exclusively **2** (Figure 5, Figure 7, Table S 21) suggesting the corresponding oxidation to the quinone scaffold to be loosely controlled. This observation may be seen as a hint that this oxidation reaction is a spontaneous process as has been described for many anthraquinone type secondary metabolites such as steffimycin,[37] and that the regioselectivity of this reaction depends on the strain-specific cytoplasmic redox power.

2.3.6   Characterization of pyxidicycline biosynthesis

Following successful heterologous expression of the pyxidicyclines we set out to further investigate their biosynthesis. Formation of the nitrogen-containing D ring is probably mediated by the tailoring enzyme operon comprising *pcyJ*, *pcyK* and *pcyL*. Considering the structure of pyxidicyclines and in light of

biosynthetic logic, the nitrogen atom involved in fourth ring closure could stem from ethanolamine, serine or glycine incorporation. However, it was not obvious whether this moiety originally derives directly from incorporated ethanolamine, from serine that is decarboxylated after incorporation or from glycine that is subsequently reduced. LC-MS analysis of extracts following addition of stable isotope labelled precursors to a *M. xanthus* DK1622 *pcy* mutant culture showed that feeding of 3,3,2-serine-d_3 leads to a +2 Daltons mass shift explained by incorporation of serine into the molecule. In contrast, feeding of 2,2-glycine-d_2 or $^{13}C_2$ ethanolamine did not lead to any labelling of pyxidicyclines. The nitrogen atom present in the D ring of the tetracene quinone-like backbone as well as the attached side chain are therefore originating from serine, which is decarboxylated following incorporation. In order to localise the position of the two incorporated deuterium atoms the labelled molecule was subjected to LC-MS² analysis revealing the position of the incorporated deuterium atoms in the molecule (section S 3.10). This finding is additional proof that the ring nitrogen N1 and the adjacent 2-hydroxyethyl moiety consisting of C19, C20 and O6 derive from serine. To investigate the effect of the three genes *pcyJ*, *pcyK* and *pcyL* on pyxidicycline biosynthesis they were deleted by λ red prophage mediated recombination (RedE/T) (section 1.12). The resulting plasmids called pClY pyxidicycline cluster Δ *pcyJ,* Δ *pcyK* and Δ *pcyL* were subsequently transformed into *M. xanthus* DK1622 to study the consequences of deleting these tailoring enzymes for pyxidicycline biosynthesis. The corresponding *M. xanthus* DK1622 Mx8 integration mutants are called *M. xanthus* DK1622 *pcy* Δ *pcyJ,* Δ *pcyK* and Δ *pcyL.* As shown in Figure 8 presence of PcyJ is indispensable for formation of the final product. Therefore, the And48 mutants where only the minimal PKS is activated as well as the *M. xanthus* DK1622 *pcy* Δ *pcyJ* mutant show no pyxidicycline formation. Interestingly, the activation mutants of the minimal PKS in An d48 show production of **4** that cannot be observed in any of the heterologous expression systems. However, the connection of this metabolite to the biosynthesis of the **3** is explained by reduction of the ketone moiety by an An d48 enzyme, which does not have an isoenzyme in DK1622 or DW4/3-1, and subsequent intramolecular lactone formation (Figure 9). Furthermore, deletion of *pycK* and *pcyL* clearly impairs the production of the pyxidicyclines whereas not only the peaks of **1** and **2** become significantly smaller, but at the same time their biosynthetic precursors **3** and **5** accumulate significantly (Figure 8, Table S 21). Therefore, although the gene *pcyJ* is sufficient to obtain the final product from the reduced analogue of **3** (Figure 9), efficient recruitment of serine into the pyxidicyclines and subsequent decarboxylation is only possible if the biosynthetic gene cluster possesses functional copies of *pcyK* and *pcyL*. Since the role of *pcyO* in the biosynthesis was unclear, the gene was placed under the control of a strong constitutive T7A1 promotor to boost its expression.
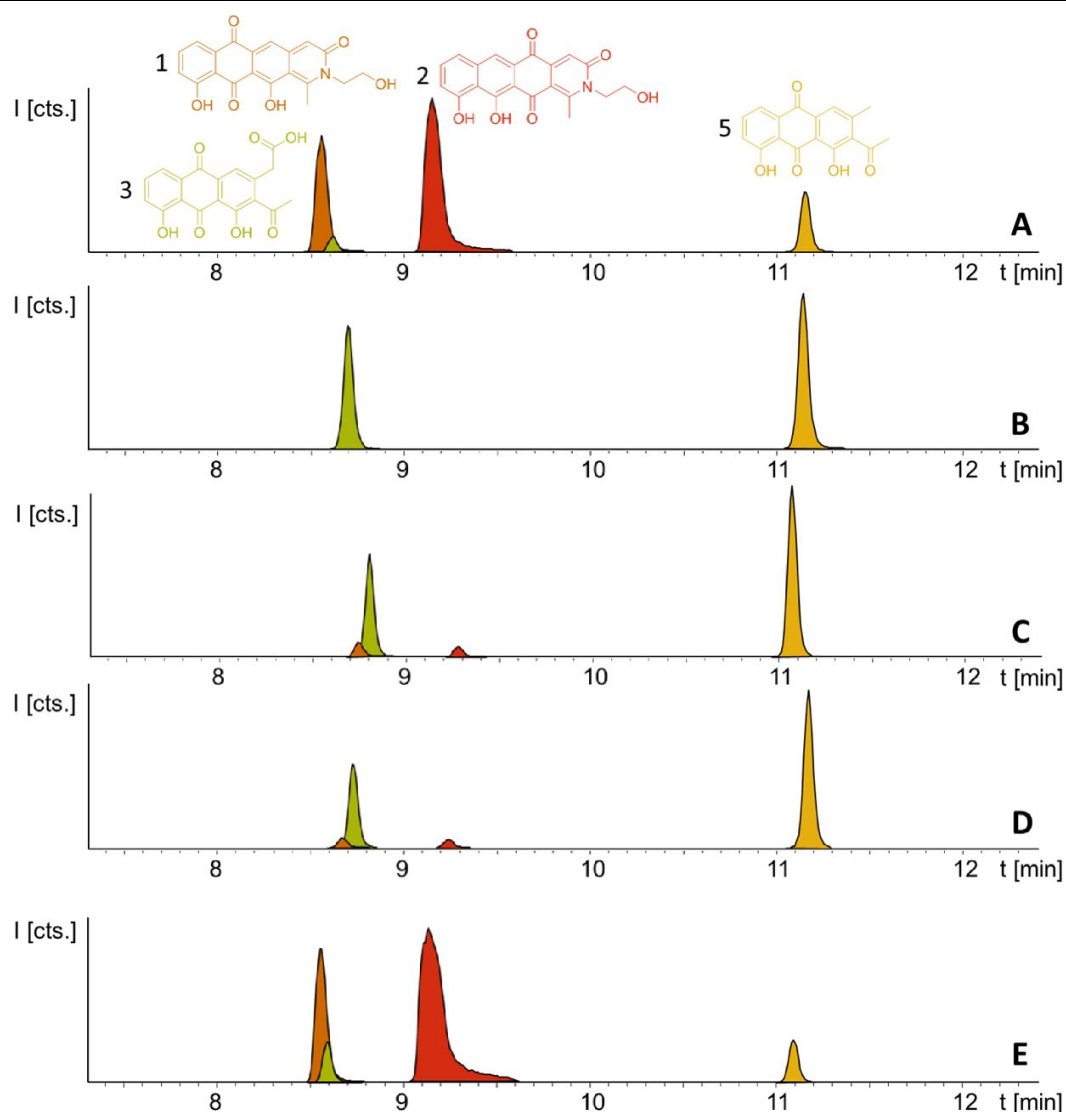
Figure 8. *(**A**) LC-MS chromatograms of M. xanthus DK1622 pcy; (**B**) pcy ΔpcyJ; (**C**) pcy ΔpcyK; (**D**) pcy ΔpcyL and (**E**) pcy T7A1 pcyO, producing the pyxidicyclines and pyxidicycline precursors. Compounds **3** and **5** are displayed as LC-MS EICs at 341.06 [M+H]⁺ (light green) and 297.075 [M+H]⁺ (orange); the two pyxidicyclines are displayed as LC-MS EICs at 366.10 [M+H]⁺ (**1** in orange; **2** in red).*

The *pcyO* gene shows similarities to transglycosylases that might be involved in glycosyl transfer to the pyxidicycline core. However, no novel products were obtained from this overexpression, but the mutant afforded slightly better yields of **1** and **2** when compared to the gene cluster on the initial pClY pyxidicycline cluster vector. A complete summary of the relative pyxidicycline and anthraquinone precursor productivities in the different strains can be found in Table S 21.

To complete the biosynthetic model for pyxidicyclines the assignment of the different cyclase proteins to their respective cyclisation reaction was done starting from *in silico* considerations. The anthraquinone backbone of the pyxidicyclines is probably assembled by two cyclases responsible for first

ring and second ring cyclisation, while third ring formation is thought to occur spontaneously. Zhang et al. could show this for structure **3**, called SEK 26 in their study.[28] To determine which cyclase enzyme is responsible for which cyclisation/aromatisation reaction of the pyxidicyclines we used a combination of gene deletion studies studies on the pClY pyxidicycline cluster plasmid and *in silico* analysis. From the deletion experiments, it is evident that aromatase/cyclase like enzymes PcyJ and PcyL cannot affect this biosynthetic step as they have been deleted in the heterologous expression host - showing only effects on the total pyxidicycline production but not on the strain's potential to produce the precursor anthraquinones (Figure 8). Furthermore, upon activation of the minimal PKS operon in the And48 mutant strains via tn5 or vanillate promotor the precursor anthraquinones were still produced but not the full-length pyxidicyclines. Therefore, we conclude that genes *pcyJ* to *pcyP* are not involved in anthraquinone precursor formation. The two remaining candidate cyclase/aromatase genes are *pcyB* and *pcyC*. The pyxidicyclines undergo C-7' to C-12' "S-type" first ring cyclisation like steffimycin and tetracycline.[38] The corresponding cyclase genes were compared to the first and second ring cyclases of the steffimycin and oxytetracycline pathway by MUSCLE alignment[17] (Figure S 12), since first and second ring cyclases typically differ by primary sequence.[38] As the cyclase PcyC resembles the oxytetracycline and steffimycin first ring cyclases it is safe to assume that PcyC is responsible for first (A) ring cyclisation (section S 2.6). On the contrary, PcyL is not part of the minimal PKS operon, which produces the anthraquinone core and has already been demonstrated to be involved in the formation and aromatisation of the nitrogen-containing D heterocycle by gene inactivation experiments. It is therefore not involved in A- and B-ring cyclisation steps. The second ring cyclases StfY and OxyN share high similarity with PcyB and all three proteins form a group with high phylogenetic distance to other proteins in the alignment tree (section S 2.6). PcyB is therefore likely to be the second (B) ring cyclase in pyxidicycline biosynthesis. The enzyme PcyJ is singled out in this alignment; this is unsurprising as our gene inactivation experiments determined the protein to be responsible for serine integration and heterocyclisation, an unprecedented step in type II polyketide biosynthesis. This enzyme was therefore expected to contain a rather uncommon cyclase/aromatase fold that does not fit very well to the enzymes known from streptomycetes (section S 2.5 and 2.6).

2.3.7   The biosynthetic route to pyxidicyclines

Based on the results of activation studies, stable isotope labelling, *in silico* analysis as well as the knockout experiments we are able to devise a comprehensive biosynthetic model for the formation of pyxidicyclines (Figure 9). The biosynthesis route shown with dotted arrows on the lower right hand side of the picture additionally illustrates formation of the anthraquinone shunt products mainly occurring if *pcyJ* is inactivated or repressed. Similar to known type II polyketide synthase pathways like the oxytetracycline

pathway,[39] the biosynthesis of pyxidicyclines starts with polycondensation of nine malonyl-CoA units by the KSα PcyE with its chain length factor PcyF, whereas intermediates are tethered to the ACP cyG. he next acting enzyme is the C-9' type ketoreductase (KR) PcyD reducing the nascent polyketide chain at C-9' counting from the thioester bond (C-10 in the final molecule). This step is necessary to pre-orientate the polyketide chain for the first ring cyclisation/aromatization catalysed by the C-7' to C-12' 'S type' cyclase PcyC that forms the pyxidicyclines' A ring.[38] The second ring cyclase/aromatase PcyB is subsequently responsible for ring closure and aromatization of the B ring by C-C bond formation between C5' and C14' on the nascent polyketide chain. As it has been shown for tetracycline biosynthesis, third ring (C) formation occurs spontaneously with these scaffolds.[39] This is also likely to be the case for the pyxidicyclines as the minimal PKS operon does not contain additional cyclases and it was shown experimentally that the minimal PKS operon is sufficient to produce anthraquinones 340, 296 and 324, all of which contain the C ring. Following C ring cyclisation the product is released and the AMP-dependent synthethase and ligase protein and aromatase/cyclase di-domain protein, PcyJ, attaches a serine moiety and closes the D ring with the help of PcyK and PcyL (section S 2.5). Serine recruitment into pyxidicyclines is notably different from amino acid incorporation during jadomycin biosynthesis, as it is not a spontaneous process but relies on catalysis by PcyJ.[32] Fredericamycin A formation on the other hand invloves catalysis by the asparagine synthethase homologue FdmV, leading to the nitrogen-containing six membered ring structure. FdmV does not transfer an amino acid but synthesizes a cyclic amide by transfer of an ammonium equivalent.[33] Our feeding experiments unambiguously showed that only 3,3,2 serine-d_3 feeding leads to labelling of the pyxidicyclines, indicating that serine is incorporated and decarboxylated on the scaffold, since $^{13}C_2$ monoethanolamine did not lead to any labelling (section S 3.10). Subsequent aromatization of the D ring can be spontaneous as indicated by the gene inactivation experiment of the muconolactone δ-isomerase type enzyme PcyK and the aromatase-like enzyme PcyL (Table 1). On the other hand, those experiments also revealed that the yield of pyxidicyclines compared to the corresponding anthraquinone shunt products drops dramatically upon inactivation of PcyK or PcyL (Table S 21). However, the exact timing for decarboxylation remains unresolved, as decarboxylation is too efficient to detect carboxylated derivatives by LC-MS. It is therefore unclear whether decarboxylation occurs before or after D ring aromatization. From the enzyme characteristics, we assume that PcyK - a muconolactone δ-isomerase homolog - is more likely to help with decarboxylation while the cyclase/aromatase homolog PcyL is more likely to assist D ring aromatization. Other enzymes cannot be involved in these transformation steps, as proven by successful production in the heterologous hosts *S. aurantiaca* DW 4/3-1 and *M. xanthus* DK1622, both considered unlikely to furnish suitable isoenzymes encoded in their genomes.
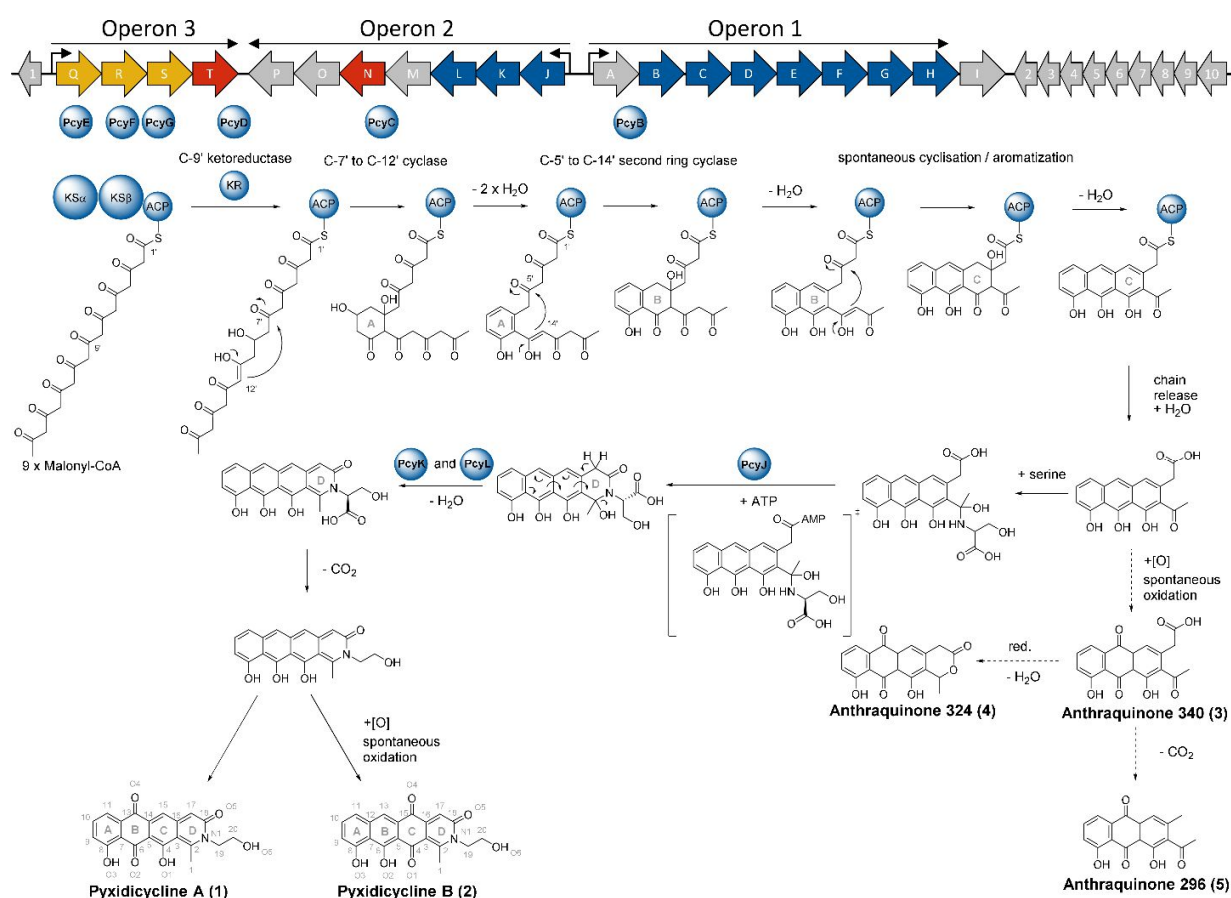
Figure 9. *Biosynthetic gene cluster organization in An d48 wild type and biosynthesis proposal for the pyxidicyclines (**1** and **2**) and the detected anthraquinone shunt products (**3-5**), according to experiments made in this study. PKS type II biosynthesis genes are depicted as blue arrows, regulatory and resistance elements are depicted as red arrows, exporter genes are depicted as yellow arrows and orfs of unknown function are depicted as grey arrows. Proteins participating in polyketide biosynthesis are depicted as blue spheres.*

2.4   Conclusion

This study presents an efficient approach for prioritizing particular BGCs out of over 93 myxobacterial genomes containing 10 to 30 BGCs each with focus on desired bioactivity and chemical novelty of the products to be discovered. Using the known host-defense mechanism of pentapeptide repeat proteins as the rationale behind our genome-mining strategy, a new small-molecule product was uncovered featuring an intriguing nitrogen-containing tetracene quinone scaffold constructed by a type II polyketide synthase pathway, a biosynthetic route rarely found in Gram-negative bacteria. Comprehensive *in vitro* bioactivity profiling ultimately allowed us to conclude the pyxidicyclines' molecular target. Consequently, the pyxidicyclines could become interesting leads mainly for anticancer research as they share the target and the order magnitude of activity with the camptothecins, an approved anti-cancer drug for an established anti-cancer drug target.[31,34]

As much as this work highlights the promise of contemporary genome-mining techniques applied to myxobacteria as a still underexploited group of microbes, it hints at the same time at the limitations currently faced by genomics-based methods for natural products discovery. While it was reasonably likely that the new compounds would exhibit biological activity in the direction of topoisomerases, it was impossible to forecast the precise range of bioactivities as well as the type of topoisomerase that would be targeted. Future developments in bioinformatics analysis are undoubtly needed for well-directed genome-mining, but this notion just as much underlines the importance to advance knowledge of biosynthetic pathway enzymatics and to develop the required genetic tools further, in order to increase our chances for translating genome-encoded potential for novel natural products into real candidate molecules, ideally for human therapy.

2.5 Acknowledgement

2.6 References

[1] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2016**, *79*, 629.

[2] N. S. Cortina, D. Krug, A. Plaza, O. Revermann, R. Müller, *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 811.

[3] D. Krug, R. Müller, *Nat. Prod. Rep.* **2014**, *31*, 768.

[4] S. Bertrand, O. Schumpp, N. Bohni, A. Bujard, A. Azzollini, M. Monod, K. Gindro, J.-L. Wolfender, *J. Chromatogr. A* **2013**, *1292*, 219.

[5] a) A. W. Schultz, D. C. Oh, J. R. Carney, R. T. Williamson, D. W. Udwary, P. R. Jensen, S. J. Gould, W. Fenical, B. S. Moore, *J. Am. Chem. Soc.* **2008**, *130*, 4507; b) F. Jüttner, A. K. Todorova, N. Walch, W. von Philipsborn, *Phytochemistry* **2001**, *57*, 613; c) J. Herrmann, A. A. Fayad, R. Müller, *Nat. Prod. Rep.* **2017**, *34*, 135.

[6] J. Munoz-Dorado, F. J. Marcos-Torres, E. Garcia-Bravo, A. Moraleda-Munoz, J. Perez, *Front. Microbiol.* **2016**, *7*, 781.

[7] a) N. Zaburannyi, B. Bunk, J. Maier, J. Overmann, R. Müller, *Appl. Environ. Microbiol.* **2016**, *82*, 1945; b) S. Schneiker, O. Perlova, O. Kaiser, K. Gerth, A. Alici, M. O. Altmeyer, D. Bartels, T. Bekel, S. Beyer, E. Bode et al., *Nat. Biotechnol.* **2007**, *25*, 1281.

[8] S. C. Wenzel, R. Müller, *Mol. Biosyst.* **2009**, *5*, 567.

[9] S. C. Wenzel, R. Müller, *Nat. Prod. Rep.* **2009**, *26*, 1385.

[10] K. Blin, T. Wolf, M. G. Chevrette, X. Lu, C. J. Schwalen, S. A. Kautsar, H. G. Suarez Duran, E. L. C. de Los Santos, H. U. Kim, M. Nave et al., *Nucleic Acids Res* **2017**, *45*, W36-W41.

[11] T. Weber, K. Blin, S. Duddela, D. Krug, H. U. Kim, R. Bruccoleri, S. Y. Lee, M. A. Fischbach, R. Müller, W. Wohlleben et al., *Nucleic Acids Res.* **2015**, *43*, W237-W243.

[12] a) M. Alanjary, B. Kronmiller, M. Adamek, K. Blin, T. Weber, D. Huson, B. Philmus, N. Ziemert, *Nucleic Acids Res* **2017**, *45*, W42-W48; b) M. A. Skinnider, N. J. Merwin, C. W. Johnston, N. A. Magarvey, *Nucleic Acids Res* **2017**, *45*, W49-W54.

[13] a) H.-H. Yeh, M. Ahuja, Y.-M. Chiang, C. E. Oakley, S. Moore, O. Yoon, H. Hajovsky, J.-W. Bok, N. P. Keller, C. C. C. Wang et al., *ACS chemical biology* **2016**, *11*, 2275; b) X. Tang, J. Li, N. Millán-Aguiñaga, J. J. Zhang, E. C. O'Neill, J. A. Ugalde, P. R. Jensen, s. M. Mantovani, B. S. Moore, *ACS chemical biology* **2015**, *10*, 2841.

[14] S. Baumann, J. Herrmann, R. Raju, H. Steinmetz, K. I. Mohr, S. Hüttel, K. Harmrolfs, M. Stadler, R. Müller, *Angew. Chem. Int. Ed.* **2014**, *53*, 14605.

[15] M. W. Vetting, S. S. Hegde, Y. Zhang, J. S. Blanchard, *Acta Crystallogr. Sect. F* **2011**, *67*, 296.

[16] a) D. Pistorius, Y. Li, A. Sandmann, R. Müller, *Mol. Biosyst.* **2011**, *7*, 3308; b) A. O. Brachmann, S. A. Joyce, H. Jenke-Kodama, G. Schwär, D. J. Clarke, H. B. Bode, *ChemBioChem* **2007**, *8*, 1721.

[17] R. C. Edgar, *Bmc Bioinform.* **2004**, *5*, 113.

[18] a) M. W. Vetting, S. S. Hegde, M. Wang, G. A. Jacoby, D. C. Hooper, J. S. Blanchard, *J. Biol. Chem.* **2011**, *286*, 25265; b) G. A. Jacoby, K. E. Walsh, D. M. Mills, V. J. Walker, H. Oh, A. Robicsek, D. C. Hooper, *Antimicrob. Agents Chemother.* **2006**, *50*, 1178.

[19] S. Shah, J. G. Heddle, *Appl. Biochem. Biotechnol.* **2014**, *98*, 9545.

[20] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, *Nat. Protoc.* **2015**, *10*, 845.

[21] a) Y. Chai, D. Pistorius, A. Ullrich, K. J. Weissman, U. Kazmaier, R. Müller, *Chem. Biol.* **2010**, *17*, 296; b) B. Kunze, W. Kohl, G. Höfle, H. Reichenbach, *J Antibiot (Tokyo)* **1985**, *38*, 1649; c) W. Trowitzsch, G. Reifenstahl, V. Wray, G. Höfle, *J. Antibiot.* **1980**, *33*, 1480.

[22] A. Das, C. Khosla, *Acc. Chem. Res.* **2009**, *42*, 631.

[23] P. J. Rutledge, G. L. Challis, *Nat. Rev. Microbiol.* **2015**, *13*, 509.

[24] T. S. Freestone, K.-S. Ju, B. Wang, H. Zhao, *ACS Synth. Biol.* **2017**, *6*, 217.

[25] a) J. B. Biggins, X. Liu, Z. Feng, S. F. Brady, *J. Am. Chem. Soc.* **2011**, *133*, 1638; b) J. Franke, K. Ishida, C. Hertweck, *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 11611.

[26] C. Olano, I. García, A. González, M. Rodriguez, D. Rozas, J. Rubio, M. Sánchez-Hidalgo, A. F. Braña, C. Méndez, J. A. Salas, *Microb Biotechnol* **2014**, *7*, 242.

[27] A. A. Iniesta, F. García-Heras, J. Abellón-Ruiz, A. Gallego-García, M. Elías-Arnanz, *J. Bacteriol.* **2012**, *194*, 5875.

[28] W. Zhang, Y. Li, Y. Tang, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 20683.

[29] a) J. R. D. McCormick, E. R. Jensen, *Adv. Ceram. Mater.* **1968**, *90*, 7126; b) V. Prikrylova, M. Podojil, P. Sedmera, J. Vokoun, Z. Vanek, C. H. Hassal, *J Antibiot (Tokyo)* **1978**, *31*, 855.

[30] J. A. Taylor, L. A. Mitchenall, M. Rejzek, R. A. Field, A. Maxwell, *PLoS ONE* **2013**, *8*, e58010.

[31] M. J. Luzzio, J. M. Besterman, D. L. Emerson, M. G. Evans, K. Lackey, P. L. Leitner, G. McIntyre, B. Morton, P. L. Myers, M. Peel et al., *J. Med. Chem.* **1995**, *38*, 395.

[32] N. Tibrewal, P. Pahari, G. Wang, M. K. Kharel, C. Morris, T. Downey, Y. Hou, T. S. Bugni, J. Rohr, *J. Am. Chem. Soc.* **2012**, *134*, 18181.

[33] E. Wendt-Pienkowski, Y. Huang, J. Zhang, B. Li, H. Jiang, H. Kwon, C. R. Hutchinson, B. Shen, *J. Am. Chem. Soc.* **2005**, *127*, 16442.

[34] Y. Pommier, *ACS Chem. Biol.* **2013**, *8*, 82.

[35] O. Bilyk, O. N. Sekurova, S. B. Zotchev, A. Luzhetskyy, *PLoS ONE* **2016**, *11*, e0158682.

[36] V. Magrini, C. Creighton, P. Youderian, *J. Bacteriol.* **1999**, *181*, 4050.

[37] S. Gullón, C. Olano, M. S. Abdelfattah, A. F. Braña, J. Rohr, C. Méndez, J. A. Salas, *Appl. Environ. Microbiol.* **2006**, *72*, 4172.

[38] H. Zhou, Y. Li, Y. Tang, *Nat. Prod. Rep.* **2010**, *27*, 839.

[39] L. B. Pickens, Y. Tang, *J. Biol. Chem.* **2010**, *285*, 27509.

Supporting Information

Self-resistance guided genome mining uncovers new topoisomerase inhibitors from myxobacteria

Affiliation

Department Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmaceutical Biotechnology, Saarland University, Campus E8.1, 66123 Saarbrücken, Germany

The printed version of the Supporting Information does not contain data that cannot be visualized satisfactorily on paper. To access this data please refer to the enclosed storage medium or the on-line version of this research paper.

S 2.1 Molecular biology protocols

S 2.1.1 Culture media used in this study

Table S 1. *Comprehensive list of culture medium recipes used for myxobacterial fermentation in this study*

CTT medium			
Amount	Ingredient	Concentration	Supplier
10 g/L	Bacto Casiton	-	BD
10 ml/L	Tris solution	1 M	Sigma Aldrich
1 ml/L	KH_2PO_4 solution	0.1 M	Sigma Aldrich
10 ml/L	$MgSO_4$ solution	0.8 M	Grüssing
Dissolved in milliq. water pH adjusted to 7.6 with 1N KOH			
TS-6 medium			
Amount	Ingredient	Concentration	Supplier
6g/L	Bacto Trypton	-	BD
10 ml/L	Tris solution	1 M	Sigma Aldrich
4 g/L	Starch (soluble)	-	Roth
10 ml/L	$MgSO_4$ solution	0.8 M	Grüssing
Dissolved in milliq. water pH adjusted to 7.6 with 1N KOH			
YM – Medium			
Amount	Ingredient	Concentration	Supplier
5 g/L	Bacto Pepton Phyton	-	BD
10 g/L	Glucose	-	Roth
3 g/L	Bacto Malt extract	-	BD
3 g/L	Bacto Yeast Extract	-	BD
1 g/L	$CaCl_2$	-	Sigma Aldrich
1 g/L	$MgSO_4$ x $7H_2O$	-	Grüssing
1 mL/L	$KHPO_4$ solution	0.1 M	Sigma
10 mM	TRIS	-	Sigma Aldrich
Dissolved in milliq. water pH adjusted to 7.5 with 1N KOH			
SC leucine dropout – Medium			
Amount	Ingredient	Concentration	Supplier
6.8 g/L	Yeast Nitrogen base	-	Sigma Aldrich
20 g/L	Glucose	-	Roth
5 g/L	Leucine dropout amino acid mix	-	Sigma Aldrich
Dissolved in milliq. water			
YPAD – Medium			
Amount	Ingredient	Concentration	Supplier
10 g/L	Yeast extract	-	BD
10 g/L	Phytopeptone	-	BD
20 g/L	Glucose	-	Roth

100 mg/L	Adenine hemisulfate	-	Sigma Aldrich
Dissolved in milliq. water			
YMS 20 – Medium			
Amount	Ingredient	Concentration	Supplier
5 g/L	Bacto Pepton Phyton	-	BD
10 g/L	Glucose	-	Roth
20 g/L	Soluble Starch	-	Roth
3 g/L	Bacto Malt extract	-	BD
3 g/L	Bacto Yeast Extract	-	BD
1 g/L	$CaCl_2$	-	Sigma Aldrich
1 g/L	$MgSO_4$ x $7H_2O$	-	Grüssing
1 mL/L	$KHPO_4$ solution	0.1 M	Sigma
10 mM	TRIS	-	Sigma Aldrich
Dissolved in milliq. water pH adjusted to 7.5 with 1N KOH			
LB medium			
Amount	Ingredient	Concentration	Supplier
10 g/L	Tryptone	-	Roth
5 g/L	Yeast extract	-	Roth
5 g/L	NaCl	-	Grüssing
Dissolved in milliq. water pH adjusted to 7.2 with 1N KOH			

For the preparation of Agar, 7.5 g/L agar is added to the medium before autoclaving, Soft agar is prepared by addition of 4 g/L agar to the medium broth.

S 2.1.2   Isolation of genomic DNA from myxobacteria

To isolate bigger amounts of higher quality genomic DNA e.g. to amplify the PCR products for TAR cloning, the alkaline lysis gDNA preparation protocol is used.

S 2.1.3   List of oligonucleotides used in this study

Oligonucleotides used in this study are ordered from Sigma Aldrich. Nucleotides with a length below 40 bp are ordered as desalted primers, while all longer oligonucleotides were ordered as RP-HPLC purified primers.

Table S 2. *List of all oligonucleotides used in this study*

Primer name	Primer sequence
FP_pSBtn5verif_in	CCTTTGGACAGCAAGCGAAC
FP_pSBtn5verif_out	CACTCATTAGGCACCCCAGG
And48_pcyA_PromEx_verif_fwd	ATTCATTGGAGTGAGGGC
And48_pcyA_PromEx_verif_rev	GATGTACTCGCAGCGGATGA
And48_pcyJ_PromEx_verif_fwd	TGTGATGGATGGACTTCGCG
And48_pcyJ_PromEx_verif_rev	CGAAGATGAGGACGTTGCCT
FP_pFPtettn5verif_in	GCCTCTTCGCTATTACGCCA
FP_pFPtettn5verif_out	CACTCATTAGGCACCCCAGG
FP pCR2.1 in	CCTCTAGATGCATGCTCGAGC
FP pCR2.1 out	GGATCCACTAGTAACGGCCG
FP t2PKS KS1 SCveriv fwd	CGGCAACTACTAACCCCACA
FP t2PKS KS1 SCveriv rev	CATGAAGTCGGTGGGGTTGA
And48 KS1 SC Primer rev	AAGACCTCCCTCGTAGTCCG
And48 KS1 SC Primer fwd	GGACCGGTACATCCAGTTCG
t2PKS_Synthlig_tn5fusion_fwd	AGACAGGATAAGGAGGTACAGCATATGTCCGAAACCGCATCCTTC
t2PKS_Synthlig_tn5fusion_RC	GAAGGATGCGGTTTCGGACATATGCTGTACCTCCTTATCCTGTCT
t2PKS_AMPdepSynth_rev	TCAGAATTCTCATGCTCCTGTGGGGATG
tn5_Promotor_primer_XbaI	AGTCTAGAGTAATACGACTCACTATAGGG
t2PKS_Synthlig_tn5fusion_fwd	AGACAGGATAAGGAGGTACAGCATATGTCCGAAACCGCATCCTTC
t2PKS_Synthlig_tn5fusion_RC	GAAGGATGCGGTTTCGGACATATGCTGTACCTCCTTATCCTGTCT
t2PKS_AMPdepSynth_rev	TCAGAATTCTCATGCTCCTGTGGGGATG
tn5_Promotor_primer_XbaI	AGTCTAGAGTAATACGACTCACTATAGGG
t2PKS_Synthlig_tn5fusion_fwd	AGACAGGATAAGGAGGTACAGCATATGTCCGAAACCGCATCCTTC
p15A_rev	GTGCACCTCTGGCTCACCGAC
p15A_fwd	TCTGTCCCTCCTGTTCAGCT
VanR_pCLY_fwd	GACTCCAACGTCAAAGGGCGATCAGTCGGCGCGAATGCTCCA
VanR_pCLY_RC	TGGAGCATTCGCGCCGACTGATCGCCCTTTGACGTTGGAGTC
pCLY_mx8_RC	GCGTACCCGAAGTTCACCGAAAATTCAGAAGAACTCGTCAAGAAGGCGATAGAAGGCG
pCLY_mx8_fwd	CGCCTTCTATCGCCTTCTTGACGAGTTCTTCTGAATTTTCGGTGAACTTCGGGTACGC
And48_t2PKS_PCR1_rev	CGGGCTGGAAGTAGGTGATG
And48_t2PKS_PCR2_fwd	CGCAACCAGGAGTTCGTGAT
And48_t2PKS_PCR2_rev	GCACGTATGAAATGGCGTCC
And48_t2PKS_PCR3_fwd	CACTCCCTTCACGTACCTGG
And48_t2PKS_PCR4_rev	AACCACTGGGAGATGTGCTG
And48_t2PKS_PCR5_fwd	GGAGGCCTACAAGGAATCCG
And48_PGyrA_fwd	TTCATGCGGAAGTAGCGAGGCGGAAGGGCCCTCATTTTTC
And48_PGyrA_rev	AAGATGGGAGGCCGGGGGACGCCATGAGCTCTGGAAGTGTTCGGA
And48_t2PKS_PCR3_rev	AAAACCCTGGCGTTACCCAACCTCCCCTACCCCGTAGTC
And48_t2PKS_PCR4_fwd	GACTACGGGGTAGGGGAGGTTGGGTAACGCCAGGGTTTT
And48_t2PKS_PCR5_rev	GAAAAATGAGGGCCCTTCCGCCTCGCTACTTCCGCATGAA
And48_t2PKS_PCR6_fwd	TCCGAACACTTCCAGAGCTCATGGCGTCCCCCGGCCTCCCATCTT
And48_t2PKS_PCR6_rev	CGCTCGACAAGCGCGGCCGTCCGGTTCAGCGGCTCTTCC
And48_t2PKS_resOperon_mx8	GGAAGAGCCGCTGAACCGGACGGCCGCGCTTGTCGAGCG
mx8_attB_up	CGAGGAGTACGGGACTTGAAC

mx8_attB_down	CGGATAGCTCAGCGGTAGAG
mx8_attP_up	CATGGTCGACGCCTACACGAC
mx8_attP_down	GGCTTGTGCCAGTCAACTGCG
CmR_Ampdep_fwd	GATGGCGCAGGGGATCAAGATCTGATCAAGAGACAGGATAAGGAGGTACAGCATCTGAACAGGAGGGACAGCTG
CmR_Ampdep_rev	CACCTGCGACGCGGACAGGTGGCGTGGGTCTTCCTGGGTGATGCAGGCAATCAACATTTACGCCCCGCCCTGCCACTCATCGCAGTA
CmR_Mucono_fwd	AGCAGAGCGCGGCGTCCCAGAAGGGCCCGTCCGCCATCCCCACAGGAGCATGAGCCTGAACAGGAGGGACAGCTG
CmR_Mucono_rev	TCGCCCGGAGCGGCCTCGGTCAGTACGGCGAACAGCATGGGTGGGTTCTCCGGTAGGGTTACGCCCCGCCCTGCCACTCATCGCAGTA
CmR_Cyclase_fwd	TCACCCACGCGATGGCCGAGAACGCCTGACCCCCTACCGGAGAACCCACCCCTGAACAGGAGGGACAGCTG
CmR_Cyclase_rev	TCAGGGGCCGAGAGCGGAACGCCGCGTCGTGTGTCCGGGAAGCTTCCGTCCATTTACGCCCCGCCCTGCCACTCATCGCAGTA
CmR_T7A1_fusion_fwd	TACTGCGATGAGTGGCAGGGCGGGGCGTAAATCCATTACGCCAAGCTTATCA
CmR_T7A1_fusion_fwd	TGATAAGCTTGGCGTAATGGATTTACGCCCCGCCCTGCCACTCATCGCAGTA
T7A1_glycohydr_rev	GCGGCGCTCTTCTGCGGAAGCGCGGTGGGCTTGGACAGGGCTTTGGTCATATGTACACCTCTCGATGG
TipA_CmR_fwd	CCGATTCACCCGTGACGCCATCATCGCGAACGCGGAGCGGCACGGCGAGTAGCTGAACAGGAGGGACAGCTG

S 2.1.4   Ligation protocol used in this study

Ligations performed in this study are done with Thermo scientific T4 ligase.

T4 DNA Ligase mix:

2µL vector DNA (1h dephosphorylated using FastAP (shrimp alkaline phosphatase, Thermo scientific))

2 – 6.5 µL insert DNA (triple the molar amount of the vector DNA)

1 µL T4 DNA ligase buffer

0.5 µL T4 DNA ligase

to 10µL ddH$_2$O

The ligase mix is incubated at 16°C for 15-24h and 1-5 µL ligation mix are subsequently electroporated into *E. coli* DH10ß using standard *E. coli* electroporation protocol.

S 2.1.5   List of polymerase chain reaction protocols used in this study

S 2.1.5.1   Thermo scientific Taq Polymerase

Taq Polymerase Protocol:

2µL NH$_4$SO$_4$ Buffer

(Alternatively 2µL KCl Buffer)

2µL MgCl$_2$ solution (25mM)

4µL dNTP's (2mM)

4μL 50% Glycerol

1μL Primer fwd and rev (50mM)

0.2 μL Taq DNA Polymerase

1μL gDNA Template (~50 ng/μL)

4.3μL ddH$_2$O

Taq Polymerase cycler program:

Table S 3. *Taq DNA Polymerase PCR program*

Step	Time [min:s]	Temperature [°C]
Initial denaturation	4:00	95
	0:30	95
Cycle, repeat 30x	0:15	63
	1:20	72
Final elongation	10	72
Store	forever	8

S 2.1.5.2   Thermo scientific Phusion Polymerase

Phusion Polymerase Protocol:

5μL GC Buffer

2.5μL dNTP's (2mM)

1.25μL DMSO

0.5μL Primer fwd and rev (100mM)

0.2 μL Phusion DNA Polymerase

0.5μL gDNA Template (~50 ng/μL)

14.55μL ddH$_2$O

Phusion Polymerase cycler program:

Table S 4. Phusion DNA Polymerase PCR program

Step	Time [min:s]	Temperature [°C]
Initial denaturation	4:00	95
	0:30	98
Cycle, repeat 30x	0:15	63
	1:20	72
Final elongation	10	72
Store	forever	8

Table S 5. Phusion DNA Polymerase PCR program for long PCR products and long oligonucleotides

Step	Time [min:s]	Temperature [°C]
Initial denaturation	4:00	95
	0:30	98
Cycle, repeat 2x	0:15	61.5
	6:00	68
	0:30	98
Cycle, repeat 30x	3:00	66
	6:00	68
Final elongation	10	72
Store	forever	8

S 2.1.6 *Pyxidicoccus fallax* An d48 antibiotics sensitivity tests

In order to determine available resistance markers for genetic manipulation of An d48 the strain's resistance towards different antibiotics was assayed by adding respective antibiotic concentrations to an YM medium agar plate. The strain shows resistance against 12 µg/ml tetracycline and partial resistance against 25 µg/ml zeocin. On the other hand the strain is sensitive to 50 µg/ml kanamycin, 25 µg/ml chloramphenicol and 12 µg/ml oxytetracycline. For practical reasons kanamycin and oxytetracycline is chosen as a second resistance marker, since the pCR2.1 Topo and pSWU22 vectors will be used. The pCR2.1 Topo is based on kanamycin. The pSWU22 vector harbors a tetracycline resistance cassette as well as a

multiple cloning site and has no overlap with the pCR2.1 Topo backbone as overlap could lead to wrong backbone-backbone recombinations.

S 2.1.7 Transformation protocols

Transformation protocol *E. coli DH10ß* TOPO TA kit cloning

1) Take competent *E. coli DH10ß* from -80°C on ice
2) Mix competent cells with pCR2.1 TOPO reaction mixture
   (0.5µL salt solution, 0.5µL pCR2.1 TOPO vector (linearized), 1-5 µL PCR product cleaned by gel electrophoresis incubated for 10 minutes)

3) Electroporate the cells at 1250 V, 25 µF, 200 Ohm and 1mm in a standard electroporation cuvette (100 µL).
4) Add 500 µL of fresh LB medium
5) Incubate on a shaker at 37°C for 1h
6) Plate on LB Kan50 Agar and incubate the plates overnight at 37°C
7) Pick around 6 clones and cultivate them for plasmid preparation via alkaline lysis

S 2.1.8 Plasmid verification

S 2.1.8.1 Plasmid preparation via alkaline lysis

1) Centrifuge 2-4 ml E. coli culture in LB medium in a 2ml Eppendorf tube (13000-15000 rpm table centrifuge 2 min)
2) Resuspend the Pellet in 250 µl Solution I (Store at +4°C after addition of RNAse A)
3) Incubate for 5 min approximately
4) Add 250 µL Solution II, mix well via inverting the Eppendorf tube
5) Add 250 µL Solution III (fast to avoid denaturing the supercoiled plasmid DNA), mix well via inverting the Eppendorf tube
6) Centrifuge for 30 minutes at 13000-15000 rpm with a table centrifuge
7) Transfer the supernatant into fresh 1.5 ml Eppendorf tubes
8) Add 850µL analytical grade Isopropanol
9) Centrifuge the sample for 30 min at 13000-15000 rpm, possibly at 4°C
10) Remove the supernatant and add 350 µL EtOH to wash the plasmid DNA
11) Centrifuge the mixture for 5 min and remove the ethanolic supernatant, this step can be repeated to clean up the DNA

12) Dry the pellet by inverting the tube on a clean piece of paper or at 65°C (make sure to avoid residual EtOH in your plasmid preparations in order to avoid re-precipitation of the vector over time)

13) Add 50 µL of nuclease free H_2O and resuspend the Pellet at 55°C for 1h.

S 2.1.8.2   Plasmid verification by restriction digestion

1) Find a restriction Enzyme that cuts one time on the vector backbone and one time on the insert.

2) Mix 8.5µL Plasmid from alkaline lysis with 1µL restriction enzyme buffer and 0.5 µL enzyme.

3) Incubate for 3-4 h

4) Verify the bands on agarose gel (0.8 %) by gel electrophoresis

S 2.1.9   Transformation of the myxobacterial strains used in this study

S 2.1.9.1   Devising a reliable mutagenesis protocol for the myxobacterium *p. fallax* And48

Electro competence of the myxobacterium *p. fallax* And48 was assayed with the plasmid pMycoMar-Kan that confers kanamycin resistance via the defective transposon *magellan-4* [1]. Following optimization of electroporation parameters, a reliable protocol was created that allowed site directed mutagenesis via single crossover homologous recombination (detailed information in the supplementary information).

S 2.1.10   Single crossover homologous recombination in *P. fallax* And48

Since Silica membrane prepared plasmid (GeneJET Plasmid Miniprep kit, Thermo scientific) transposon mutagenesis of And48 yields more than 400 clones per 500 µg plasmid while standard alkaline lysis led to clone numbers inferior to 10, all manipulations of *Pyxidicoccus fallax* And48 are done using silica membrane purified plasmid.

Transformation protocol

1) Centrifuge 2ml of an And48 culture in YM Medium at OD_{600} of approx. 0.3 at 8000 rpm for 2 minutes at a table centrifuge.

2) Wash the residual cell pellet 2 times with 500 µL autoclaved ddH_2O and discard the supernatant.

3) Resuspend cells in 50µL of ddH_2O, add 5 µL of plasmid solution (prepared from E. coli with Thermo Scientific Miniprep Kit) at a conc. of 0.3-0.4 ng/µL and transfer the solution into an electroporation cuvette.

4) Electroporation at 675 V, 400 Ohms, 25µF and 1mm cuvette settings for optimum electroporation efficiency.

5) Flush out the cells with 1 mL fresh YM medium, transfer the Cell suspension into a 2ml Eppendorf tube.

6) Incubate the Cells for 5h on a Shaker thermostated to 37°C at 300 rpm.

7) Plating the cells on Kan50 YM Agar after mixing the cell suspension with 3 ml of Kan50 YM Softagar.

8) Slightly red, spherical clones appear after 7-12 Days in the 30°C incubator.

S 2.1.11 Transformation protocol for *M. xanthus* DK1622

The transformation protocol for *M. fulvus* SBMx132 was elaborated from CTT medium cultures, since CTT medium leads to growth in perfect suspension. This is a prerequisite for efficient electro transformation of myxobacteria.

Transformation protocol

9) Centrifuge 2ml of an Mx132 culture in CTT Medium at OD_{600} of approx. 0.3 at 8000 rpm for 2 minutes at a table centrifuge.

10) Wash the residual cell pellet 2 times with 500 µL autoclaved ddH_2O and discard the supernatant.

11) Resuspend cells in 50µL of ddH_2O, add 5 µL of plasmid solution (prepared from E. coli with Thermo Scientific Miniprep Kit) at a conc. of 0.3-0.4 ng/µL and transfer the solution into an electroporation cuvette.

12) Electroporation at 675 V, 400 Ohms, 25µF and 1mm Cuvette Settings for optimum electroporation efficiency.

13) Flush the cells out with 1 mL fresh CTT medium, transfer the Cell suspension into a 2ml Eppendorf tube.

14) Incubate the Cells for 5h on a Shaker thermostated to 37°C at 300 rpm.

15) Plating the cells on Kan50 CTT Agar after mixing the cell suspension with 3 ml of Kan50 CTT Softagar.

16) Clones appear after 5-7 Days in the 30°C incubator.

S 2.1.12 Transformation protocol for *S. aurantiaca* DW 4/3-1

The transformation protocol for *S. aurantiaca DW4/3-1* was elaborated from TS-6 medium cultures, since TS-6 medium leads to growth in perfect suspension.

Transformation protocol

17) Centrifuge 2ml of a *S. aurantiaca* DW4/3-1 culture in TS-6 Medium at OD_{600} of approx. 0.3 at 8000 rpm for 2 minutes at a table centrifuge.

18) Wash the residual cell pellet 2 times with 500 μL autoclaved ddH$_2$O and discard the supernatant.

19) Resuspend cells in 50μL of ddH$_2$O, add 5 μL of plasmid solution (prepared from E. coli with Thermo Scientific Miniprep Kit) at a conc. of 0.3-0.4 ng/μL and transfer the solution into an electroporation cuvette.

20) Electroporation at 650 V, 400 Ohms, 25μF and 1mm cuvette Settings for optimum electroporation efficiency.

21) Flush the cells out with 1 mL fresh TS-6 medium, transfer the cell suspension into a 2ml Eppendorf tube.

22) Incubate the Cells for 5h on a Shaker thermostated to 37°C at 300 rpm.

23) Plating the cells on Kan50 TS-6 Agar after mixing the cell suspension with 3 ml of Kan50 TS-6 Softagar.

24) Clones appear after 10-14 Days in the 30°C incubator.

S 2.1.13   Elaboration of the plasmids used for single crossover integration

S 2.1.13.1   Single crossover inactivation plasmids for the pyxidicycline cluster

In order to disrupt the pyxidicycline cluster, ca. 1 kb single crossover homology arms targeting the KSα on the *pcyE* gene and the cyclase on *pcyC* were amplified from An d48 genomic according to the protocol for gDNA isolation from gram-negative bacteria (DNA Puregene Core Kit A, Quiagen) and ligated into pCR2.1-TOPO (TOPO-TA cloning, Thermo scientific TOPO-TA Kit). Single crossover integration was achieved by electro-transformation of An d48 using a protocol elaborated in this work (section S 1.9) and resulting mutants were checked for correct integration by PCR using the TAQ PCR protocol (section S 1.5). For single crossover inactivation of the pyxidicycline secondary metabolite cluster a pCR 2.1 vector from the TOPO-TA cloning kit (TOPO-TA cloning KIT, Thermo scientific) was used. Amplification of the 1 kb homology fragments was done with TAQ Polymerase (Thermo scientific) from And48 gDNA according to the TAQ polymerase amplification protocol (section S 1.5). The vector containing the KSα encoded on *pcyE* insert was verified by restriction digestion with the restriction Enzyme NcoI that cuts once on the pCR2.1 backbone and once on the KSα homology arm insert and analogously the vector containing homology to the cyclase *pcyC* was verified with NaeI. Two *E. coli* clones whose plasmids show the correct restriction pattern are selected for plasmid isolation with the plasmid isolation kit (GeneJET Plasmid Miniprep kit, Thermo scientific).
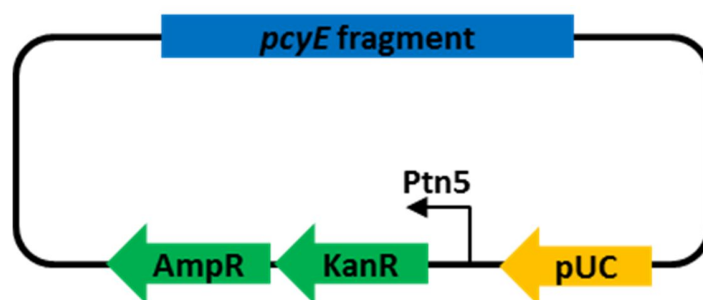
Figure S 1. *pCR2.1 vector from the pTOPO Kit carrying a 1kb homology arm for single crossover inactivation of the KSα on the pcyE gene.*

S 2.1.13.2   Plasmids used for promotor exchange of the *pcyA* gene

The first utilized construct was designed in order to exchange the alleged native promotor against a tn5 promotor. This promotor was chosen as it was known to successfully drive the kanamycin resistance gene expression in *M. xanthus*. The plasmids used for overexpression of the minimal PKS operon are called pSBtn5 and pFPVan, consisting of a pCR2.1 vector backbone from the TOPO-TA cloning kit (Thermo scientific, U.S.A) and a tn5 promotor or a vanillate promotor repressor system. The plasmids created in this section serve as tool to control the transcription and translation of operon 1 of the pyxidicycline cluster via a heterologous promotor.
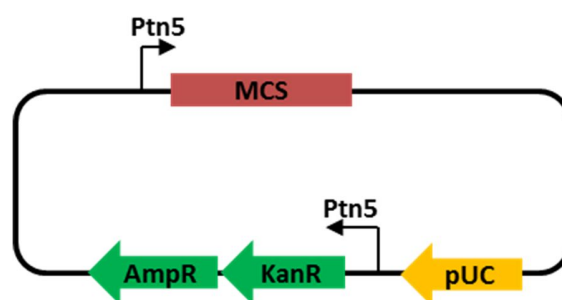
Figure S 2. *pSBtn5 vector used for placing the pyxidicycline gene cluster under the control of the contained tn5 promotor*

The first 1-1.5 kb of a biosynthetic assembly line are inserted into the pSBtn5 with the restriction enzymes EcoRI and NdeI that are introduced during PCR with restriction sites on the primers. NdeI is necessary as it contains the start codon and usage of EcoRI as a complementary enzyme facilitates cloning as both enzymes are working in Thermo scientific 1x Orange buffer. In order to get good cloning efficiency and avoid vector relegation clones it is necessary to dephosphorylate the vector backbone thoroughly

before ligation. This is done by utilizing the FastAP shrimp phosphatase (Thermo Scientific) which is also compatible with Thermo Scientific 1xO buffer. Electroporation of *Pyxidicoccus fallax* An d48 is done according to the devised protocol (see section S 2.1.9) with the assembled and purified pSBtn5 plasmid harboring the first gene of the type 2 PKS minimal PKS Operon. As this gene is called *pcyA*, this assembled plasmid will be called pSBtn5 *pcyA*.
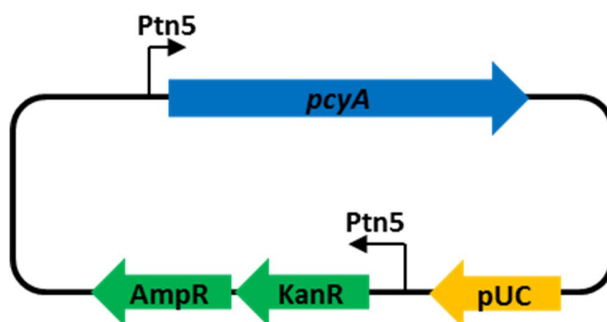
Figure S 3. pSBtn5 pcyA construct used for overexpression of p. fallax And48 pyxidicycline cluster

In order to avoid self-toxicity a second vector for single crossover insertion is created that harbors a promotor-repressor system inducible by vanillate. As the pCR2.1 Vector from the TOPO-TA Kit (Thermo scientific) works well for single crossover insertions of foreign DNA in *P. fallax* An d48, pCR2.1 is taken as a basis for this second vector system. The first gene of the t2PKS assembly line *pcyA* is amplified and the vanillate promotor-repressor system is added via overlap extension PCR using the Phusion PCR protocol that also introduced NdeI as well as EcoRI restriction sites on the Primers (Table S 3). This PCR fusion product is ligated into the cut vector via TOPO-TA cloning (TOPO TA Kit, Thermo scientific).
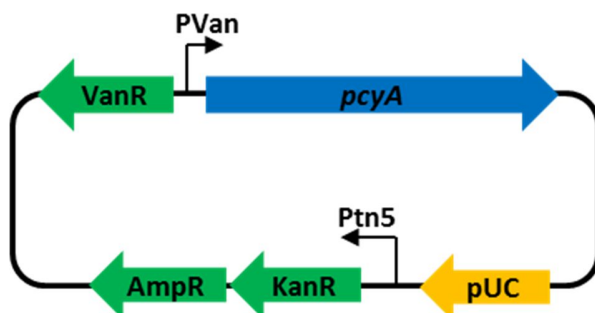
Figure S 4. pFPVan pcyA used for overexpression of the t2PKS cluster in Pyxidicoccus fallax *An d48*

Since both of these vectors are designed to work with the restriction enzymes EcoRI and NdeI, these gene cluster overexpression vectors can be utilized with the same PCR product into which NdeI and EcoRI restriction sites have been introduced.

S 2.1.13.3   Creation of an orthogonal tetR based vector for a second single crossover integration

In order to introduce a second promotor exchange into the bacterial chromosome of *Pyxidicoccus fallax* And48 *pcyA* :: pFPVan *pcyA,* a vector with an orthogonal resistance marker had to be created. As preliminary studies showed oxytetracycline to be a usable selection marker for *P. fallax* An d48 (section S 1.6), a vector harboring a p15A origin, the *tetR* gene as well as a fusion of a tn5 promotor to the *pcyJ* gene was created by standard restriction cloning based on the pSWU22 plasmid. This plasmid solely consists of the tetracycline exporter gene as resistance marker, a pUC origin and a multiple cloning site. The plasmid pSWU22 possesses a pUC origin which is not usable in this case as this might lead to single crossover integration of pUC :: pUC which is not desired. Therefore, the origin was exchanged against a p15A origin by restriction ligation using DraI and ApaLI. This also leads to elimination of the NdeI cutting site on the plasmid backbone that would impede efficient use of the created plasmid later. Overlap extension PCR with the Phusion PCR protocol was subsequently used to amplify a fusion PCR product consisting of a tn5 promoter and the *pcyJ* gene harboring XbaI cutting sites on the primer sequences. Standard restriction cloning is subsequently used to insert the tn5-*pcyJ* fusion sequence into the modified pSWU22 plasmid to create the pFPtettn5 *pcyJ* vector.
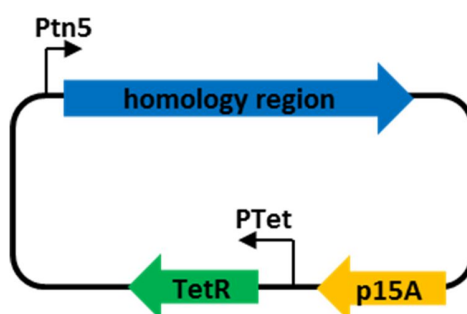
Figure S 5. *schematic view of the pFPtettn5 vector used for creation of the* Pyxidicoccus fallax *An d48* pcyA *:: pSBtn5* pcyA*;* pcyJ *:: pFPtettn5* pcyJ (*An d48* pcy) *mutant*

This allowed exchange of the native pcyJ promotor against a tn5 promotor via single crossover homologous recombination in *Pyxidicoccus fallax* An d48 *pcyA* :: pFPVan *pcyA* to obtain *Pyxidicoccus fallax* An d48 *pcyA* :: pFPVan *pcyA; pcyJ :: pFPtettn5 pcyJ* (An d48 *pcy). *Single crossover integration of this plasmid into the *Pyxidicoccus fallax* An d48 *pcyA* :: pFPVan *pcyA* strain was achieved by electroporation according to standard protocol (section S 1.9). As mentioned in the results, despite serious efforts we were unable to obtain integration clones of said plasmid into *Pyxidicoccus fallax* An d48 *pcyA* :: pSBtn5 *pcyA.* Plasmid integration verification was done by PCR using the TAQ PCR protocol (section 1.5).

S 2.1.13.4   Verification of plasmid integration into the myxobacterial strains

Verification of the myxobacterial mutant clones is done by PCR with TAQ DNA Polymerase (Thermo Scientific). Therefore, genomic DNA from the myxobacterium is isolated with the Puregene Core Kit A (Quiagen) according to the protocol for gDNA isolation from gram-negative bacteria. The Taq DNA Polymerase master mix is prepared according to Table S 3. Primers are in all cases 2 primers that bind on the genomic DNA next to the homologous region if homologous recombination is used or next to the integration site if phage integrase based integration is used as well as two complementary primers that bind on the integrated vector. Therefore the appearance of PCR products of the right size proves homologous recombination to have occurred in the correct position.

S 2.1.14   Protocols for transformation assisted recombination (TAR cloning)

S 2.1.14.1   Stock solutions needed for TAR cloning of biosynthetic gene clusters

PEG 3500 solution

50 % [w/v] PEG 3500 (Sigma Aldrich) in milliq Water

Single strand carrier DNA stock solution

2 mg/ml salmon sperm DNA (Sigma Aldrich) in autoclaved milliq Water

Boil for 5 min at 95 °C and freeze aliquots of 500 µL, these aliquots can be used several times

LiOAc stock

1M LiOAc in TE Buffer

TE Buffer

1 mM EDTA

10 mM Tris

in milliq water

pH 8.0

S 2.1.14.2   TAR cloning protocol adapted from Bylik et al. [2]

Previous day:

1) Inocculate the *S. cerevisiae* strain ATCC4004247 in 10 ml YPAD medium a 50 ml baffled flask over night at 30°C on a rotary shaker at 200 rpm. Simultaneous pre - warm one 300 ml baffled flask containing 50 ml of sterile YPAD medium to 30°C

TAR cloning day:

2) Determine approximate culture titer by measuring OD 600 of 10µL yeast culture in 1ml of milliq H_2O, OD of 0.1 is approximately 10^6 yeast cells.

3) Add ca. 2×10^8 yeast cells to the pre warmed YPAD medium and incubate for 4h at 30°C 200 rpm.

4) Denature one aliquot of the SS carrier DNA stock solution for 5 min at 95°C and chill the solution on ice.

5) Harvest the yeast cells at 3000g for 5 min in a 50 ml falcon, wash twice with 25 ml sterile milliq H_2O and re suspend the cells in 1000 µL of sterile milliq H_2O.

6) Pipet aliquots of 200 µL of this solution in 1.5 µL Eppendorf tubes and remove the water by centrifuging at 3000g for 5 min

7) Add 360µL of transformation mix to the pellets and mix by pipetting up and down

8) Place the Eppendorf tubes (closed) in a floating device and heat shock the *S. cerevisiae* cells at 42°C for 30 min

9) Centrifuge at top speed for 30s and remove residual transformation mix by pipetting

10) Re suspend the pellet in 1ml of sterile milliq H_2O

11) Streak the cells on a SC leucine dropout agar plate and incubate the plate for 2-3 days

12) Pick the clones on a replica plate and select for positive clones via colony PCR

13) *S. cerevisiae* clones that give a positive colony PCR signal are grown in 5 mL of sterile SC leucine dropout medium and the plasmid is prepared from 2ml of *S. cerevisiae* culture

S 2.1.14.3   Plasmid preparation from yeast strains

Plasmids from TAR cloning for heterologous expression were prepared from the Δ Leu2 leucine auxotrophy strain *S. cerevisiae* ATCC 4004247 [2].

Plasmid preparation follows the following scheme:

1) Centrifuge 1.5 ml *S. cerevisiae* culture in SC dropout medium in a 1.5 ml Eppendorf tube (13000-15000 rpm table centrifuge 2 min)

2) Re suspend the Pellet in 150 µl Solution I

3) Add 2ml of Zymolyase solution (10mg/ml) to each tube

4) Incubate for 30 min approximately

5) Add 150 µL Solution II, mix well via inverting the Eppendorf tube

6) Add 150 µL Solution III (fast to avoid denaturing the supercoiled plasmid DNA), mix well via inverting the Eppendorf tube

7) Centrifuge for 30 minutes at 13000-15000 rpm with a table centrifuge

8) Transfer the supernatant into fresh 1.5 ml Eppendorf tubes

9) Add 700µL analytical grade Isopropanol and mix well by inverting the tube several times

10) Centrifuge the sample for 30 min at 13000-15000 rpm, possibly at 4°C

11) Remove the supernatant and add 500 µL 70% EtOH to wash the plasmid DNA

12) Centrifuge the mixture for 5 min and remove the ethanolic supernatant, this step can be repeated to clean up the DNA

13) Dry the pellet by inverting the tube on a clean piece of paper or at 65°C (make sure to avoid residual EtOH in your plasmid preparation in order to avoid re-precipitation of the vector over time)

14) Add 30 µL of nuclease free H_2O and re suspend the pellet at 55°C for 1h.

S 2.1.15   TAR cloning based pyxidicycline gene cluster assembly

Pxyidicoccus fallax An d48 *pcyA* :: pFPVan *pcyA*, *pcyJ* :: pFPtetp15Atn5 *pcyJ* strain's genomic was prepared according to alkaline lysis genomic DNA preparation standard protocol for all further PCR reactions (section S 1.2). For naturally overlapping PCR fragments e.g. PCR2 and PCR3 the PCR product overlap necessary for TAR cloning was introduced by PCR, for non-overlapping fragments as PCR3 and PCR4, the overlap was introduced with the primer sequences (main text Figure 3 and Figure S 6). 25-30 Bp homology introduced through the primers are sufficient for recombination in yeast. Amplification of the cluster sequence was done by using the PCR program for long PCR products (section S 1.5.2). Co-transformation of the PCR products of the pyxidicycline cluster into the TAR yeast *S. cerevisiae* ATCC 4004247 is done at a total concentration of 25 pg x Number of nucleotides per PCR product to get an equimolar PCR product mix in the yeast transformation mix. The capture vector was designed to harbor all genes necessary for replication and auxotrophy selection in S. cerevisiae analogous to the vector described by Bylik et al.[2] The last fragment imperative for the heterologous expression of the pyxidicycline cluster in *Myxococcus xanthus* DK1622 is the Mx8 integrase prophage with the attP site that mediates stable and efficient integration of plasmids into the DK1622 chromosome's attB site [3],[1]. This fragment also contains a kanamycin resistance cassette necessary for the latter selection of integration clones in *M. xanthus* DK1622. Furthermore this resistance is helpful during the retransformation of the constructs in *E. coli* as since it is located on a PCR product different of the one harboring the p15A origin - it helps to prevent selecting plasmids consisting only of the pClY backbone as these would not confer kanamycin resistance. TAR cloning was done according to a protocol by Bylik et al.[2] that was slightly adapted (section S 1.11). Yeast clones appearing on leucine deficient medium were checked by PCR for a fragment of the *pcyE* gene by colony PCR. Clones that gave a positive result were prepared from yeast (Zymo Prep Kit, Zymo Research, U.S.A) and transformed into *E. coli* DH10ß by electroporation. Plasmid preparation and

plasmid digestion with the enzymes NdeI, BstEII and NotI were used to confirm the plasmid by restriction digestion. The vector pClY pyxidicycline cluster was sequenced by Illumina sequencing at GMAK (Brunswick, Germany). The sequence was proven correct except for a single nucleotide polymorphism that leads to a silent mutation in the ATP binding protein of the ABC transporter encoded on *pcyQ*. The plasmid is transformed into *M. xanthus* DK1622 and *S. aurantiaca* DW 4/3-1 according to standard protocol (section S 1.9.3 and 1.9.4). Correct integration via the Mx8 integrase was checked by PCR using the TAQ PCR protocol (section 1.5).[1]
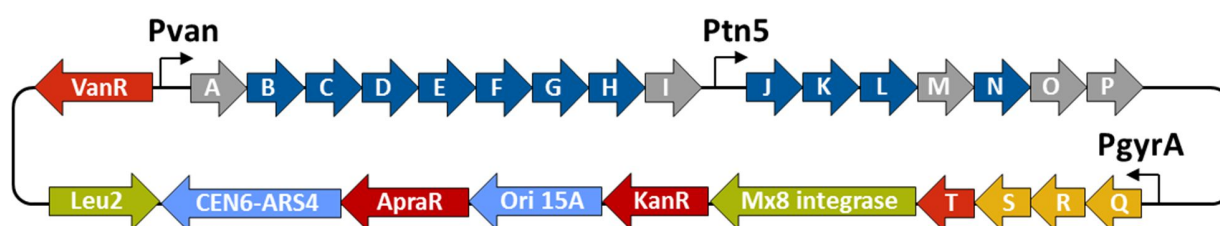
Figure S 6. *Schematic view of the pClY Pyxidicycline vector. PKS type-II biosynthesis genes depicted in blue, regulation and resistance genes in red, exporter genes in orange, auxiliary genes for the cloning strategy in light green and orfs of unknown function in gray.*

S 2.1.16   Protocols for λ-Red prophage recombination (Red E/T)

The strain used for all λ-Red based recombination experiments was *E. coli HS996* carrying the thermosensitive pSC101 Bad red gba plasmid. As the origin of this plasmid stops replicating at 37°C the strain has to be grown at 30°C for cryo cultures.

S 2.1.16.1   Incorporation of the target plasmid into the λ-Red recombination strain

1) Inoculate 25 ml LB containing 12 µg/ml oxytetracycline and grow the strain overnight at 30°C 200 rpm on a rotary shaker

2) Inoculate 20 µL of the well grown *E. coli* overnight culture into a 2 ml Eppendorf tube with a punctured lid containing 1.5 ml LB with 12 µg/ml oxytetracycline

3) Incubate the culture for 4h until OD reaches approx. 0.6

4) Mix target vector or Ligation mixture (ca.1-3µL) with competent *E. coli* and pipet the mixture into a Electroporation cuvette (100µL)

5) Electroporate the mixture at 1250V, 25µF, 200 Ohm, 1mm Cuvette

6) Add 1mL fresh sterile LB medium, incubate for 1h at 900 rpm 30°C in a 2ml Eppendorf tube with a punctured lid

7) Plate on LB Kan50 Otc12 Agar (provided the plasmid has kanamycin resistance, if not use appropriate antibiotic) and incubate the plates overnight at 30°C

8) Pick around 6 clones and cultivate them for plasmid preparation via alkaline lysis

9) Confirm the electroporated plasmid via restriction digestion

S 2.1.16.2   Modification of plasmids via Red E/T

1) Inoculate 25 ml LB containing 12 µg/ml oxytetracycline and 50 µg/ml Kanamycin and grow the strain overnight at 30°C 200 rpm on a rotary shaker

2) Inoculate 20 µL of the well grown *E. coli* overnight culture into a 2 ml Eppendorf tube with a punctured lid containing 1.5 ml LB with 12 µg/ml oxytetracycline and 50 µg/ml Kanamycin

3) Incubate the culture for 2h until OD reaches approx. 0.2 (sometimes longer incubation times are necessary)

4) Centrifuge the culture at 8000 x g for 5 minutes to pellet the cells

5) Wash the residual cell pellet 2 times with 1000 µL autoclaved milliqH_2O and discard the supernatant.

6) Re suspend the cell pellet in 50 µL autoclaved milliq H_2O and add the PCR product for modification

7) Electroporate the mixture at 1300V, 25µF, 600 Ohm, 1mm Cuvette

8) Add 1mL fresh sterile LB medium, incubate for 1h at 900 rpm 30°C in a 2ml Eppendorf tube with a punctured lid

9) Transfer the mixture into 50 ml LB Kan50 containing appropriate concentration of the antibiotic on the modification PCR product (in my case Chloramphenicol 25µg/ml) and incubate over night at 30°C

10) Prepare plasmid from this mixture via alkaline lysis

11) Electroporate the plasmid mix into *E. coli DH10ß* according to the *E. coli DH10ß* electroporation protocol and plate it on LB Kan50 Cm25

12) Pick around 4 clones and cultivate them for plasmid preparation via alkaline lysis

13) Confirm the electroporated plasmid via restriction digestion

The retransformation is necessary in order to avoid having clones harboring both plasmids (the modified and the unmodified one) in the same *E. coli* clone as well as to get rid of the pSC101 Bad Red gba plasmid

S 2.1.17   Elaboration of the Red E/T deletion plasmids utilized in this study

To elucidate functions of the genes in the *pcy* cluster's operon two, genes were subsequently replaced by a chloramphenicol cassette on the heterologous expression construct. Therefore, the plasmid

was transformed into the host strain *E. coli* HS996 harboring the pSC101BAD RedγβαA plasmid. Ca. 80 bp homology arms to the neighboring regions of the genes selected for deletion via Red/ET were put on primers suited to amplify the chloramphenicol resistance cassette (Table S 2). PCR products were obtained using the Phusion PCR protocol (section S 1.5). Subsequent Red/ET replacement of the target genes led to the creation of deletion plasmids for the genes *pcyJ*, *pcyK* and *pcyL.* For overexpression of the *pcyO* gene, a T7A1 promotor was fused to the chloramphenicol resistance cassette equipped with 80 bp homology arms to replace the *pcyO* promotor region. Red/ET recombination was done according to standard linear to circular Red/ET (section S 1.13). The resulting plasmids were transformed into *Myxococcus xanthus* DK1622 via Mx8 mediated integration and the resulting strains termed *M. xanthus* DK1622 *pcy Δ pcyJ*, *pcyK* and *pcyL,* depending on the deleted enzyme. The *M. xanthus* mutants were extracted according to standard protocol and analyzed by LC-MS. Pyxidicycline and anthraquinone production titers of M. xanthus DK1622 harboring the pClY pyxidicycline cluster plasmid were compared to the titer of the strains harboring the *pcyJ*, *pcyK* and *pcyL* defective plasmids by LC-MS EIC peak area integration (Table S 21).
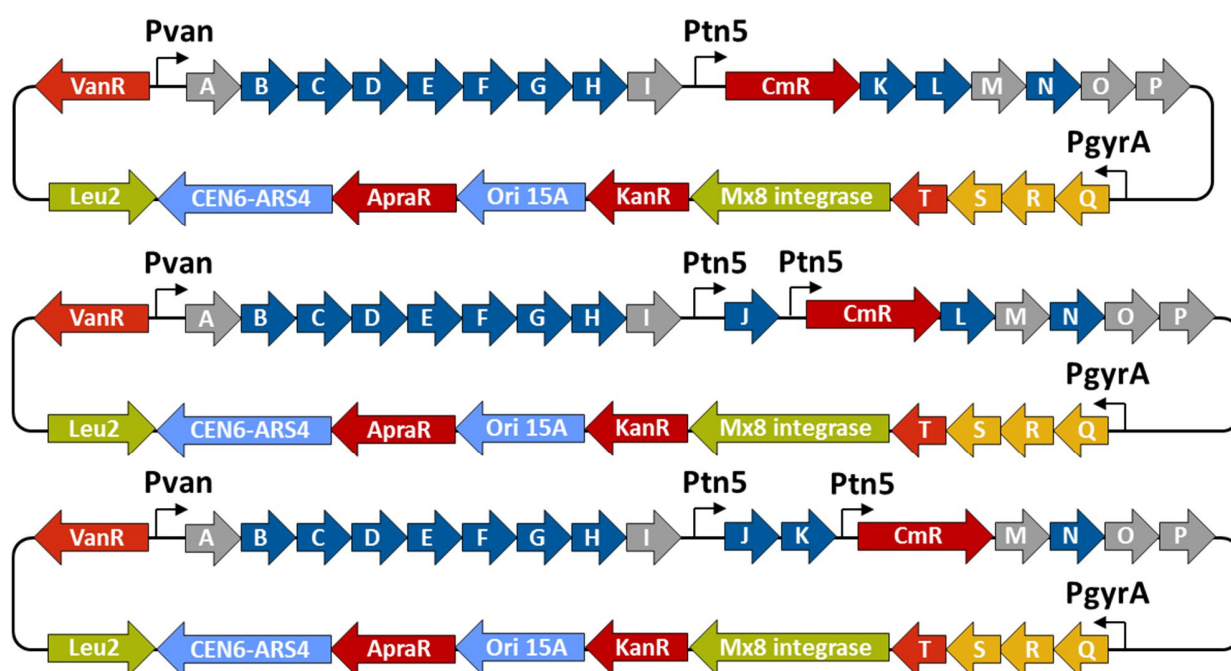
Figure S 7. *Schematic view of the pClY Pyxidicycline vector deletion constructs Δ* pcyJ*,* pcyK *and* pcyL*. PKS type-II biosynthesis genes depicted in blue, regulation and resistance genes in red, exporter genes in orange, auxiliary genes for the cloning strategy in light green and orfs of unknown function in gray.*

S 2.2 *In silico* analysis of the Pyxidicycline biosynthetic gene cluster

S 2.2.1 *In silico* analysis of the pentapeptide repeat protein

In order to analyze similarities of the found putative topoisomerase self-resistance proteins with self-resistance proteins against the gyrase poisons ciprofloxacin, albicidin and cystobactamid the pentapeptide repeat proteins were aligned via MUSCLE alignment. Alignment of the proteins in general shows them to have a rather low overall homology as the proteins are structural proteins with the relaxed consensus motiv [S,T,A,V][D,N][L,F][S,T,R][G] for every repeat. Still the similarity of PcyT to McbG in *E. coli* and CysO in *C. velatus* indicates that this protein might have a similar function.[4]

S 2.2.2 3D modelling of the pentapeptide repeat self-resistance protein via the Phyre² server

Since the 3D structure of pentapeptide repeat proteins is most likely more important than the primary sequence as these proteins are most likely structural proteins that mimic DNA to protect the hosts topoisomerases an *in silico* 3D structure prediction was done.

Figure S 8. *3D structure homology model of the pentapetide repeat protein putatively conferring topoisomerase self-resistance PcyT modelled by Phyre² [5]*

The best matching proteins found by psi-blast on the Phyre² server to obtain a 3D homology model were *Xanthomonas albilineans* AlbG (100.0 confidence scoring; 29 % identity) and the quinolone resistance protein QNR from *Aeromonas Hydrophilia* (99.9 confidence scoring; 17 % identity). Both of these proteins

confirm topoisomerase poison resistance and the devised homology model depicted in the picture above looks strikingly similar to described topoisomerase inhibitor resistance proteins.[4]

S 2.2.3 *In silico* analysis of the Pyxidicycline biosynthetic gene cluster

Table S 6. *Table of all open reading frames belonging to the pyxidicycline (pcy) cluster with proposed function and closest homologue according to a blastp search against the nr (non-redundant protein sequences) protein database at NCBI*

Name	Length [bp]	proposed function	closest annotated homologue – organism of origin	Identity [%] and length of alignment [AA]	Accession Nr.
pcyA	1128	precursor release from the acyl carrier protein	acyl-CoA thioester hydrolase – multispecies	27.7 / 361	WP_060519218
pcyB	852	polyketide cyclisation (B-ring cyclisation)	polyketide cyclase - *Streptomyces nogalater*	46.3 / 283	AAF01818
pcyC	963	polyketide cyclisation (A-ring cyclisation)	polyketide cyclase – *Streptomyces atratus*	32.9 / 325	WP_072488808
pcyD	798	C9' type ketoreductase	ketoacyl reductase – *Sorangium cellulosum*	69.1 / 259	WP_061624890
pcyE	1281	ketoacyl synthase KSα subunit	3-oxoacyl-ACP synthase II – *Dendrosporobacter quericolus*	56.0 / 425	SDL92203
pcyF	1242	ketoacyl synthase KSβ subunit (chain length factor)	3-oxoacyl-ACP synthase II – *Pelosinus sp.*	37.7 / 414	WP_038670561
pcyG	237	acyl carrier protein	acyl carrier protein – *Terriglobus saanensis*	33.3 / 78	WP_013566858
pcyH	372	4'-phosphopanthenyl transferase	holo-ACP synthase – *Sorangium cellulosum*	61.5 / 122	WP_061624886
pcyI	249	-	hypothetical protein – *Myxococcus fulvus*	37.9 / 66	WP_052771001
pcyJ	2088	serine incorporation	4-coumarate-CoA ligase family protein – *Modestobater species*	37.3 / 542	WP_082557217
pcyK	303	heterocyle formation	muconolactone delta-isomerase – *Acidovorax valerianellae*	32.5 / 83	SDC81513
pcyL	717	heterocyle formation	polyketide cyclase – *Streptomyces species*	29.3 / 232	WP_053683051
pcyM	378	-	-	-	-
pcyN	762	regulation	MerR family transcriptional regulator – *Myxococcus fulvus*	76.6 / 252	WP_074955963
pcyO	2862	-	glycosyl hydrolase family 31 – *Mucilaginibacter paludis*	60.4 / 934	WP_008504858
pcyP	720	-	hypothetical protein – *Mucilaginibacter paludis*	32.2 / 233	WP_008504859
pcyQ	816	export	ABC transporter, ATP binding protein – *Myxococcus xanthus*	93.0 / 272	ABF90137
pcyR	783	export	ABC transporter permease - *Myxococcus xanthus*	90.7 / 259	WP_011552893
pcyS	1035	export	ABC transporter substrate binding protein - *Myxococcus xanthus*	85.4 / 343	WP_011552892
pcyT	639	putative self-resistance mechanism	pentapeptide repeat containing protein – *Myxococcus fulvus*	70.3 / 209	WP_046715273
orf1	1431	-	serine/threonine protein kinase – *Myxococcus xanthus*	58.5 / 470	WP_011552901
orf2	993	-	hydrolase – *Myxococcus fulvus*	73.4 / 308	WP_013941483
orf3	417	-	elongation factor GreAB – *Myxococcus hansupus*	86.2 / 138	WP_002640736
orf4	714	-	hypothetical protein – *Stigmatella aurantiaca*	30.3 / 238	WP_075011090
orf5	1047	-	hypothetical protein – *Archangium violaceum*	30.8 / 156	WP_043405726
orf6	2163	-	serine/threonine protein kinase – *Archangium gephyra*	38.8 / 707	WP_047858246

orf7	375	-	-	-	-
orf8	1461	-	dihydropyrimidine dehydrogenase subunit A – *Myxococcus xanthus*	86.2 / 486	WP_011553928
orf9	4566	-	glutamate synthase large subunit – *Myxococcus virescens*	90.8 / 1520	SDF16482

The native Pyxidicycline gene cluster will be accessible at GenBank under Accession number MH048639 upon publication. All plasmid maps will be provided upon request.

S 2.2.4   Identification of KSα and KSβ *in silico*

Both ketosynthase proteins PcyE and PcyF in this assembly line are predicted to be 3-oxoacyl-ACP synthase 2 proteins. Therefore, sequence alignment is done in order to find out which of the sequences acts as the KSα protein and which protein acts as the KSβ. These functions can be identified via NCBI protein blast as, contrary to the KSβ, it possesses all catalytically active residues (see Figure S 9 and Figure S 10). In the PcyF amino acid sequence the catalytically active cysteine residue is mutated to arginine rendering it catalytically inactive.[6]
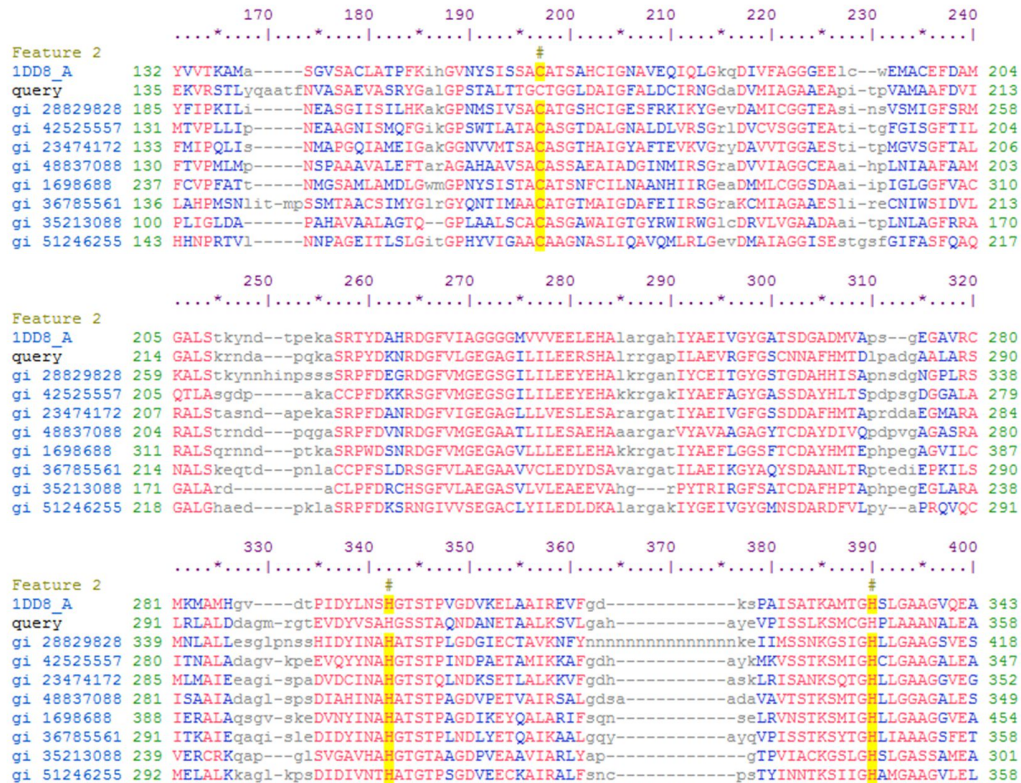
Figure S 9. Identification of the KSα protein on the gene pcyE *via blast alignment against the non-redundant protein sequences collection (nr) on the NCBI protein server*
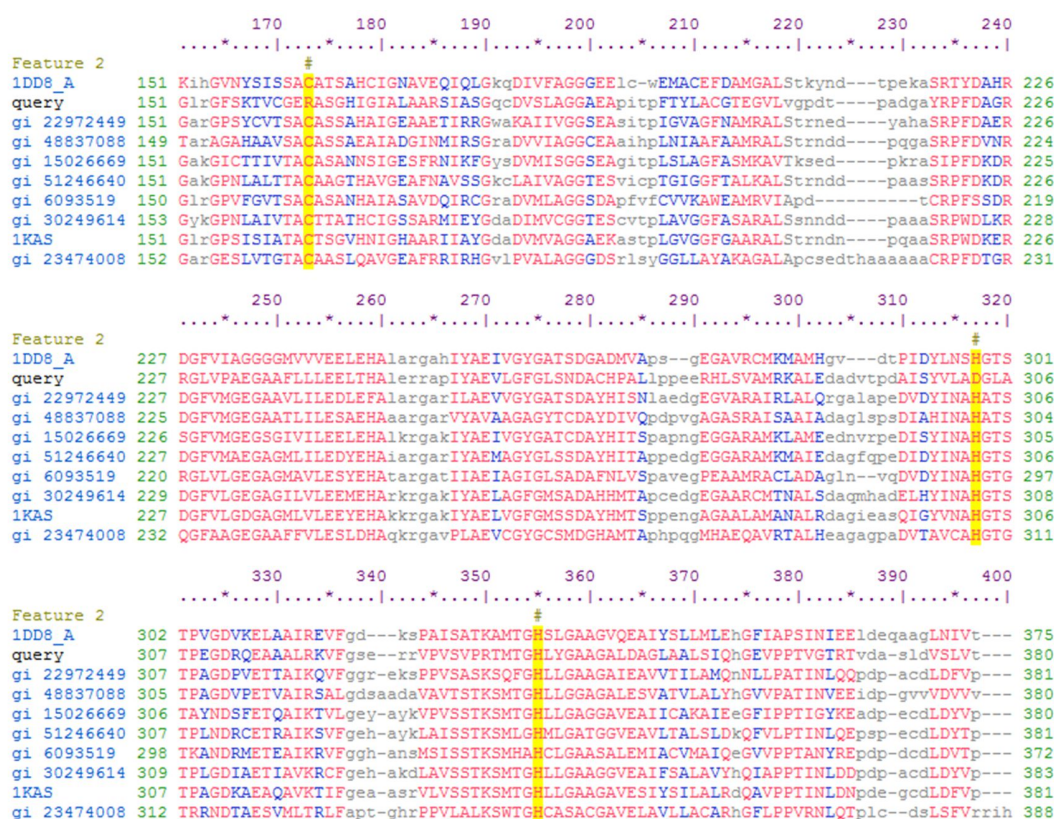
Figure S 10. *Identification of the KSβ protein on the gene* pcyF *via blast alignment against the non-redundant protein sequences collection (nr) on the NCBI protein server*

S 2.2.5   *In silico* analysis of the PcyJ Protein

The PcyJ protein is responsible for amide bond formation between the polyketide intermediate of the pyxidicyline biosynthesis and serine to form the D ring in pyxidicycline A and B. Therefore the protein has to have acyl transfer activity to form an amide bond between serine and the polyketide precursor as well as hetero cyclisation activity.

Conserved domain search hits in the PcyJ protein
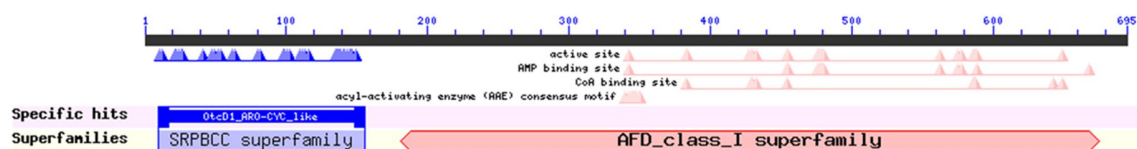
Figure S 11. *NCBI conserved domain search hits in the PcyJ protein stating the di-domain nature of the protein as a AMP dependent synthetase/ligase – aromatase/cyclase fusion protein*

Analysis of the PcyJ protein by NCBI conserved domain search [7] reveals an aromatase/cyclase domain (blue) putatively responsible for hetero cyclisation and an AMP-dependent synthetase and ligase function putatively responsible for amide bond formation.

S 2.2.6   *In silico* assignment of the pyxidicyclines' cyclase proteins

During maturation of the pyxidicyclines there are several cyclase like proteins involved to form the tetracene quinone like core structure. In order to assign the first and second ring cyclases which was not possible by knockout experiments we relied on protein alignment via the MUSCLE [8] algorithm to group the cyclase proteins into groups.
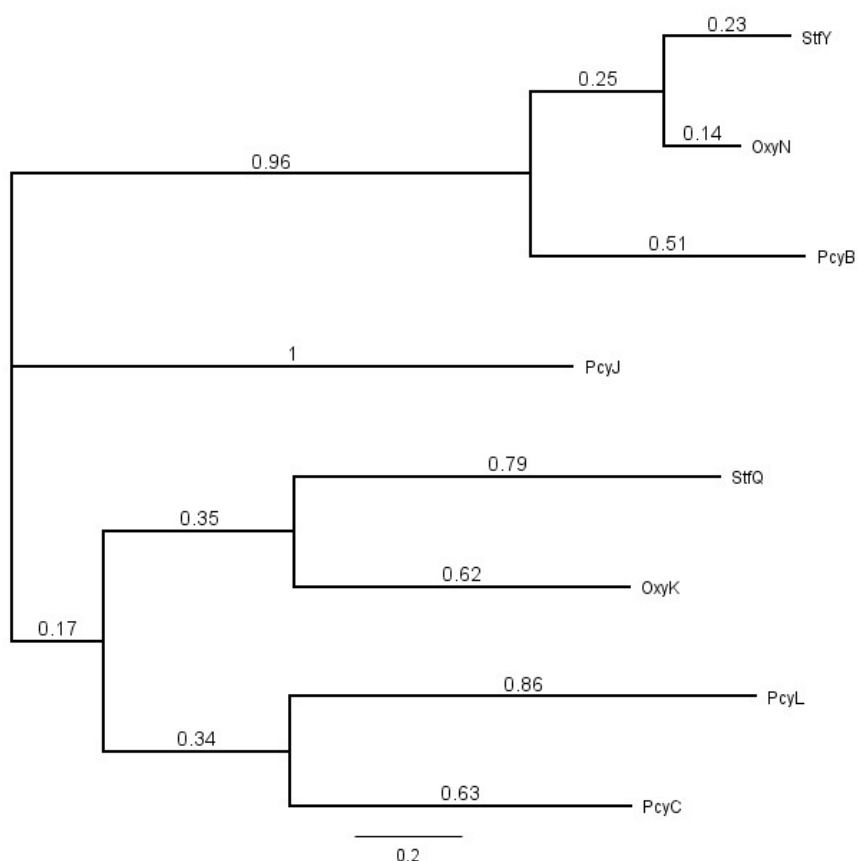
Figure S 12. *Neigbour joining tree of the cyclases in the Pyxidicycline, Steffimycin and Oxytetracycline pathways based on MUSCLE alignment of the respective proteins*

As one can see, PcyC and PcyL have homology to the first ring cyclase proteins OxyK and StfQ while the PcyB cyclase has homology to the second ring cyclases OxyN and StfY.[9,10] The specialized heterocycle forming protein PcyJ forms an outgroup in this cyclase *in silico* comparison. As we could show by knockout experiments, PcyL and PcyJ are involved in the cyclisation of the fourth nitrogen containing ring (D ring)

and since the third ring closure is spontaneous in these tetracycline like scaffolds,[9] this experiment strongly suggests PcyB to be the second ring (B ring) cyclase while PcyC is the first ring (A ring) cyclase.

S 2.3   Isolation and characterization of the pyxidicyclines

S 2.3.1   Bacterial fermentation procedures

Cultures for UPLC/high-resolution LC-MS analysis were grown in 300 ml shake flasks containing 50 ml of culture medium inoculated with 1 ml of pre culture. After inoculation the medium was supplemented with 2 % of sterile XAD-16 adsorber resin (Sigma Aldrich) suspension in water. Mutant strains are supplemented with concentrations of the respective antibiotics at concentrations as follows: Kanamycin 50 µg/ml, oxytetracyclin 12 µg/ml, ampicillin 100 µg/ml and apramycin 50 µg/ml. *Pyxidicoccus fallax* strain An d48 was grown in YM medium at 30 °C for 7-9 days using a orbitron shaker (Infors) at 160 rpm, likewise *Myxococcus xanthus* strain DK1622 was grown in CTT medium at 30°C for 5-7 days and *Stigmatella aurantiaca* DW4/3-1 was grown in TS-6 medium at 30°C for 11-14 days. Fermentation cultures of *Pyxidicoccus fallax* strain And48 *pcyA* :: pFPvan *pcyA*, *pcyJ* :: pFPtettn5 *pcyJ* used for isolation of the pyxidicyclines A and B were grown in 5 x 5 L shake flasks containing 2 L of YM medium each. Fermentation duration was 14 days post inoculation; inoculation was done with 50 ml of pre-culture and the medium was supplemented with 2 % of a sterile XAD-16 adsorber resin suspension in water. Fermentation cultures of *Pyxidicoccus fallax* strain And48 *pcyA* :: pSBtn5 *pcyA* used for isolation of the anthraquinone precursors anthraquinone 340, 324 and 296 were grown in 5x 5 L shake flasks containing 2 L of YM medium each. Fermentation duration was 10 days post inoculation; inoculation was done with 50 ml of pre-culture and the medium was supplemented with 2% of sterile XAD-16 adsorber resin suspension in water.

S 2.3.2   Preparation of small scale bacterial extracts for UHPLC-HRMS analysis

Pellet and cells were centrifuged at 6500 rcf in a 50 ml falcon centrifuge, the supernatant was decanted and the cell pellets were frozen at -20°C. The frozen cell pellet was transferred into a 100 ml erlenmeyer flask and a magnetic stirrer added. 50 ml of acetone (fluka analytical grade, redistilled in house) were added onto the pellet and the mixture was stirred for 60 minutes on a magnetic stirrer. The acetone extract was left to settle in order to sediment cell debris and XAD resin before a second extraction step. The supernatant was filtered with a 125 micron folded filter keeping cell pellet and XAD-16 resin in the Erlenmeyer flask for the second extraction step. The residual pellet and XAD-16 resin is extracted again with 30 ml of distilled acetone for 60 min on a magnetic stirrer and filtered through the same folded filter. The combined extracts are transferred into a 100 ml round bottom flask. The acetone was evaporated on

a rotary evaporator at 260 mbar and 40 °C water bath temperature. The residual water was evaporated at 20 mbar until the residue in the flask was completely dry. The residue was taken up in 550 µl of methanol (Chromasolv HPLC grade) and transferred into a 1.5 ml Eppendorf tube. This tube was centrifuged at 21500 rcf for 5 minutes to remove residual insolubilities such as salts, cell debris and XAD fragments.

S 2.3.3   Standardized UPLC-MS and UPLC-MS^2 Conditions

All measurements were performed on a Dionex Ultimate 3000 RSLC system using a Waters BEH C_{18} column (50 x 2.1 mm, 1.7 µm) equipped with a Waters VanGuard BEH C_{18} 1.7 µm guard column. Separation of 1 µl sample (Crude extract 1/5 dilutions in methanol, section S 3.2) was achieved by a linear gradient from (A) H_2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 600 µL/min and a column temperature of 45 °C. Gradient conditions were as follows: 0 – 0.5 min, 5% B; 0.5 – 18.5 min, 5 – 95% B; 18.5 – 20.5 min, 95% B; 20.5 – 21 min, 95 – 5% B; 21-22.5 min, 5% B. UV spectra were recorded by a DAD in the range from 200 to 600 nm. The LC flow was split to 75 µL/min before entering the Bruker Daltonics maXis 4G hr-qToF mass spectrometer using the Apollo II ESI source. Mass spectra were acquired in centroid mode ranging from 150 – 2500 m/z at a 2 Hz full scan rate. Mass spectrometry source parameters were set to 500V as end plate offset; 4000V as capillary voltage; nebulizer gas pressure 1 bar; dry gas flow of 5 l/min and a dry temperature of 200°C. Ion transfer and quadrupole settings were set to Funnel RF 350 Vpp; Multipole RF 400 Vpp as transfer settings and Ion energy of 5eV as well as a low mass cut of 300 m/z as Quadrupole settings. Collision cell was set to 5.0 eV and pre pulse storage time was set to 5 µs. Spectra acquisition rate was set to 2Hz. Calibration of the maXis4G qTOF spectrometer was achieved with sodium formate clusters before every injection to avoid mass drifts. All MS analyses were acquired in the presence of the lock masses $C_{12}H_{19}F_{12}N_3O_6P_3$, $C_{18}H_{19}O_6N_3P_3F_2$ and $C_{24}H_{19}F_{36}N_3O_6P_3$ which generate the [M+H]$^+$ Ions of 622.028960, 922.009798 and 1221.990638. The corresponding MS^2 method operating in automatic precursor selection mode picks up the two most intense precursors per cycle, applies smart exclusion after five spectra and performs CID and MS/MS spectra acquisition time ramping. CID Energy was ramped from 35 eV for 500 m/z to 45 eV for 1000 m/z and 60 eV for 2000 m/z. MS full scan acquisition rate was set to 2Hz and MS/MS spectra acquisition rates were ramped from one to four Hz for precursor Ion intensities of 10kcts to 1000kcts.

S 2.3.4   Statistical Data Treatment using Principal Component Analysis

One issue encountered while trying to identify the secondary metabolites produced by the *pcy* biosynthetic gene cluster was that homologous recombination apparently also affected the host strain in ways not necessarily linked to the locus of homologous recombination. While the metabolome detectable

with LC-MS changed only minimally upon targeted gene cluster disruption, introduction of the promotor upstream to the pcyA gene lead to diverse changes in the metabolome (main text Figure 4). It was therefore crucial to apply a statistical filter using PCA in order to pinpoint peaks truly unique to the 'activation' mutant and to discriminate those displaying only a fold change in abundance. To search for statistically relevant differences between the two pyxidicycline cluster inactivation mutants and the wild type as well as between the promotor exchange mutants and the inactivation mutants, 6 LC-HRMS measurements from extracts of three independent cultivations per bacterial strain are used for statistical data treatment. For preprocessing of MS data the molecular feature finder implemented in Bruker Compass Data Analysis 4.2 (Bruker, Bremen) was used with the compound detection parameters SN threshold 1; Correlation coefficient 0.9; minimum compound length 10 spectra and smoothing width of 3 spectra. Bucketing was done with Bruker Compass Profile Analysis 2.1 (Bruker, Bremen) with advanced bucketing and window parameters of 30s and 15 ppm. Bucket value was log transformed to avoid under-evaluating low intensity signals in the presence of high intensity signals [11]. The PCA t-Test function was used in order to separate medium derived MS features from the metabolome derived MS features, since the t-Test table that can be exported from Bruker Compass Profile Analysis 2.1 (Bruker, Bremen) contains information upon how many blanks and how many bacterial extracts contain said feature.

S 2.3.5   Purification protocols of all pyxidicycline related compounds

S 2.3.5.1   Isolation of the anthraquinone precursors 296, 324 and 340

The *Pyxidicoccus fallax* mutant strain *Pyxidicoccus fallax* And48 *pcyA* :: pSBtn5 *pcyA* is fermented in 50 ml YM medium supplemented with Kanamycin 50 µg/ml as a seed culture flasks on an Orbiton shaker at 160 rpm and 30°C. After the culture reaches OD of 0.8 after 5 to 7 days of fermentation it is used to inoculate 6 x 2L YM Kanamycin 50 µg/ml medium supplemented with 2 % XAD-16 resin suspension in sterilized water in 6 x 5L baffled shake flasks on an Orbiton shaker at 160 rpm and 30°C. Fermentation is complete after 14 days. Cells and XAD-16 resin are harvested by centrifugation on a Beckmann Avanti J-26 XP with the JLA 8.1 rotor at 6000 rcf. Combined resin and cells are extracted with 2 x 500 ml of technical grade methanol (Fluka) and 2 x 500 ml of technical grade acetone (Fluka). The extracts are combined and all solvent is evaporated on a rotary evaporator. Residue is taken up in 10 ml of analytical grade methanol (Fluka) for chromatographical separation. Purification was carried out on a Waters Autopurifier (Eschborn, Germany) high pressure gradient system, equipped with 2545 binary gradient module, SFO system fluidics organizer, 2767 sample manager and a 2998 photodiode array detector coupled to a 3100 single quadrupole mass spectrometer operated in positive ion mode. Source and voltage settings for the MS were as follows: mass range, m/z 300 - 1000; scan duration, 1 s; points per Dalton, 4; capillary

voltage, 3.5 kV; cone voltage, 30 V; extractor voltage 3 V; RF lens, 0.1 V; source temperature 120 C, desolvation temperature, 250 C; desolvation gas flow, 400 L/hr; cone gas flow, 50 L/hr; ion counting threshold, 30. Autopurification is done by setting the fraction collection trigger masses 297.1, 325.1 and 340.1 which correspond to the $[M+H]^+$ ions of the anthraquinones. Separation was carried out on Phenomenex Kinetex Biphenyl 5 µm, 250 x 20 mm column using MeOH + 0.1 % FA as B and H2O + 0.1% FA as A with a flow rate of 25 ml/min. Separation is started with a plateau at 95% A for 2 minutes followed by a ramp to 32% A during 4 minutes and a ramp to 2% A during 30 minutes. The A content is kept at 5% A for 2 minutes. The A content is ramped back to starting conditions during 30 seconds and the column is re equilibrated for 5 minutes. Further purification is done using a Dionex Ultimate 3000 SDLC low pressure gradient system on a Phenomenex Luna C-18, 5µm 250x10mm column with the eluents H2O + 0.1 % FA as A and MeOH + 0.1 % FA as B, a flow rate of 5 ml/min and a column thermostated at 30°C. The Anthraquinones are detected by UV absorption at 256 nm and 432 nm. The gradient that starts with a plateau at 95 % A for 2 minutes followed by a ramp to 59 % A during 6 minutes. Then A content is ramped to 30 % during 17 minutes and finally ramped to 5 % A during 1 minute. A content is kept at 5 % for 1 minute and then ramped back to 95 % during 30 seconds. The column is reequilibrated at 95 % A for 5 minutes. The pure compounds are subsequently dried by lyophilization. All three anthraquinones are obtained as dark yellow substances that possess slightly varying fluorescent yellow colors in solution depending on the solvent. Upon deprotonation, the color of all three compounds changes to dark red. The compounds are analyzed by LCUV and LC-HRMS before NMR spectra are acquired.

Figure S 13. UV absorption spectra of the anthraquinone metabolites acquired with the Dionex Ultimate 3000 diode array detector during an LC run using H_2O + 0.1% FA and ACN + 0.1% FA as eluents.

Figure S 14. *LC-UV Chromatograms of the three isolated Anthraquinone metabolites for purity analysis by liquid chromatography*
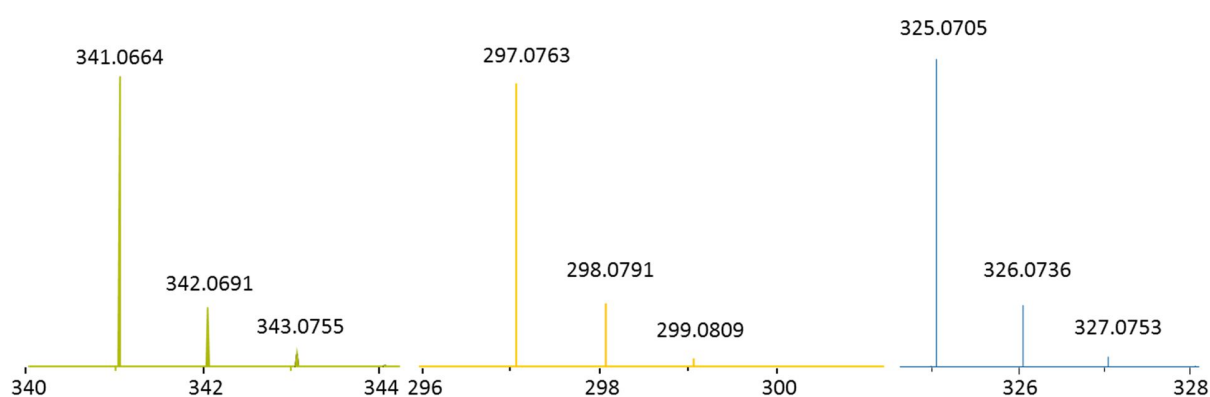
Figure S 15. *High resolution masses of anthraquinone 340,296 and 325 [M+H]⁺ in m/z units acquired on our maXis4G qTOF high resolution mass spectrometer*

S 2.3.5.2   Isolation of Pyxidicycline A and B

The *Pyxidicoccus fallax* mutant strain *Pyxidicoccus fallax* And48 *pcyA* :: pFPVan *pcyA, pcyJ* :: pFPtettn5 *pcyJ* (An d48 *pcy*) is fermented in 50 ml YM medium supplemented with Kanamycin 50 µg/ml as a seed culture flasks on an Orbiton shaker at 160 rpm and 30°C. After the culture reaches OD of 0.8 after 5 to 7 days of fermentation it is used to inoculate 6 x 2L YM Kanamycin 50 µg/ml medium supplemented with 2 % XAD-16 resin suspension in sterilized water in 6 x 5L baffled shake flasks on an Orbiton shaker at 160 rpm and 30°C. Simultaneously with inoculation the cultures are supplemented with 5mM Sodium vanillate

solution to induce pyxidicycline production. Fermentation is complete after 17 days. Cells and XAD-16 resin are harvested by centrifugation on a Beckmann Avanti J-26 XP with the JLA 8.1 rotor at 6000 rcf. Combined resin and cells are extracted with 2 x 500 ml of technical grade methanol (Fluka) and 2 x 500 ml of technical grade acetone (Fluka). The extracts are combined and all solvent is evaporated on a rotary evaporator. The residue is now partitioned between saturated brine and chloroform (2 extractions of 500 ml brine with 400 ml of chloroform) in a 2L separating funnel leading to close to complete transfer of the pyxidicyclines A and B into the chloroform phase. The chloroform phase is subsequently dried on a rotary evaporator. The residue is resuspended in 50 ml of analytical grade Acetonitrile and transferred completely into a 50 ml falcon tube (Sarstedt). After vigorous vortexing of the suspension the falcon tube is centrifuged for 15 min at 7800 rcf to pellet the pyxidicyclines that are close to insoluble in acetonitrile. The dark red residue is taken up in 1 ml of analytical grade formic acid (99 %, Sigma Aldrich) for chromatographic separation of pyxidicycline A and B. Purification is done using a Dionex Ultimate 3000 SDLC low pressure gradient system on an Agilent Zorbax XDB-C8, 5µm 250x10mm column with the eluents H_2O + 0.5% FA as A and MeOH + 0.5% FA as B, a flow rate of 5 ml/min and a column thermostated at 30°C. The pyxidicyclines are detected by UV absorption at 474 nm (pyxidicycline B) and 498 nm (pyxidicycline A). The pyxidicyclines are further purified with the gradient that starts with a plateau at 95% A for 2 minutes followed by a ramp to 74% A during 3 minutes. Then A content is ramped to 56% during 32 minutes and finally ramped to 5% A during 30 seconds. A content is kept stable at 5% for 1 minute and then ramped back to 95% A during 30 seconds. The column is reequilibrated at 95% A for 5 minutes. The pure compounds are subsequently dried by lyophilization and obtaines as the corresponding formate esters. The pyxidicyclines are obtained as dark red substances that possess slightly varying fluorescent orange colors in solution depending on the solvent. Upon deprotonation, the color of both pyxidicycline compounds changes to dark blue. In total 5mg of pyxidicycline A and 4 mg of pyxidicycline B are obtained from the fermentation culture as formate esters. It is worth noting that the actual theoretical yield of pyxidicyclines seems to be at least 50 % higher as on the one hand elution of the pyxidicyclines from the XAD-16 resin is incomplete and on the other hand peak shape issues with the compounds in reverse phase chromatography due to low insolubility in all tested LC solvents lead to low recovery rates after chromatographic separation. Both compounds are very soluble and stable in formic acid while being insoluble in most common organic solvents and instable in DMSO and DMF.
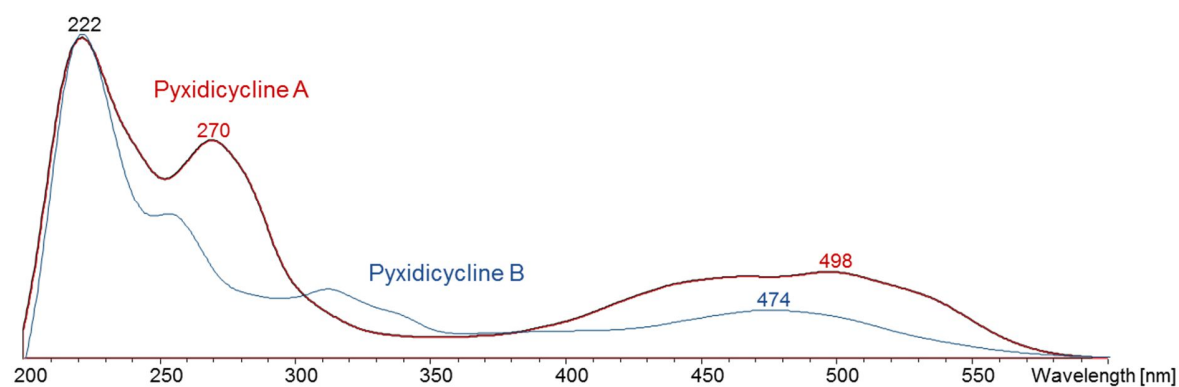
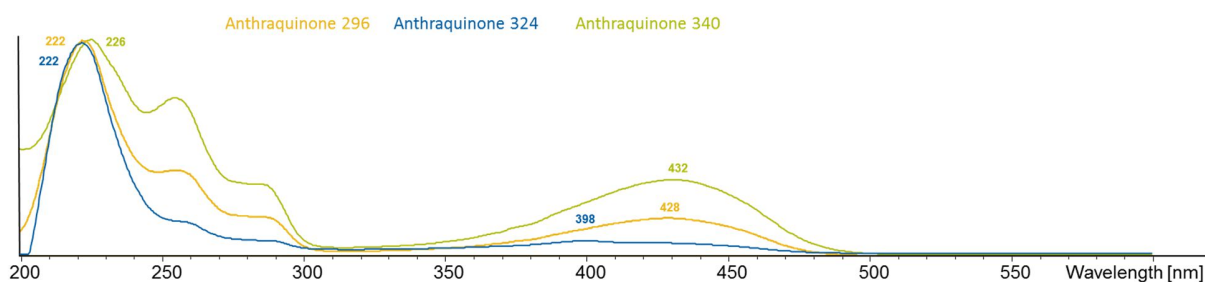***Figure S 16.*** *UV absorption spectra of pyxidicycline A and B acquired with the Dionex Ultimate 3000 diode array detector during an LC run using H_2O + 0.1% FA and ACN + 0.1% FA as eluents.*

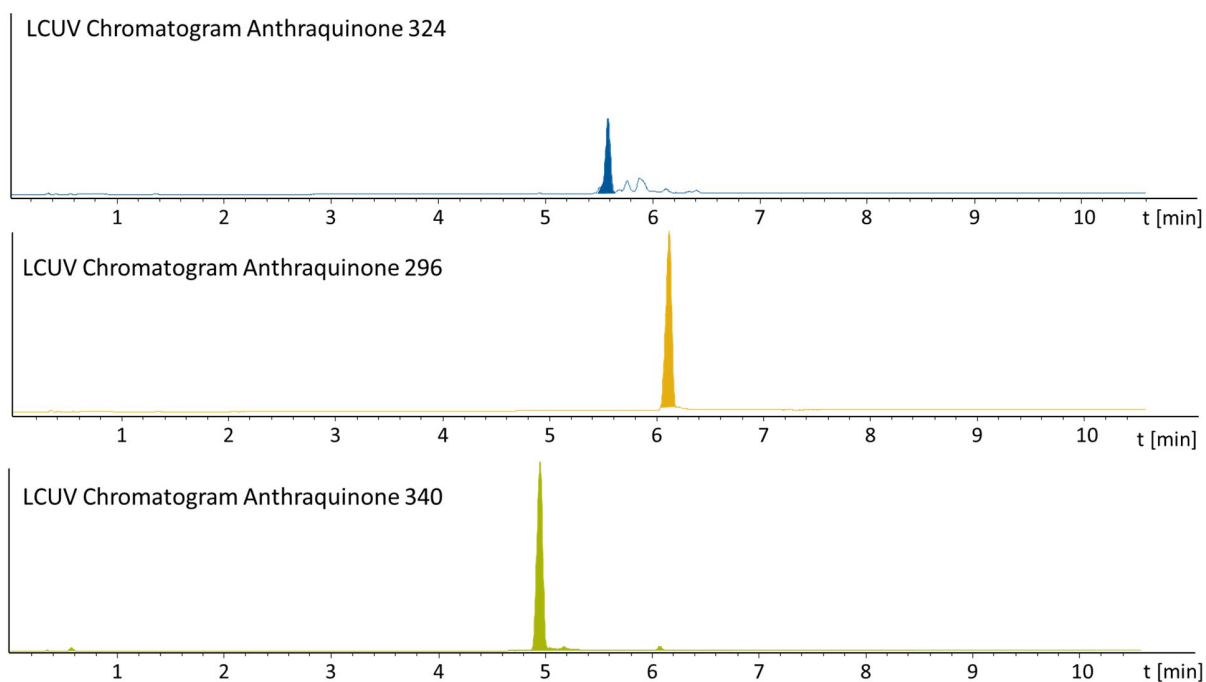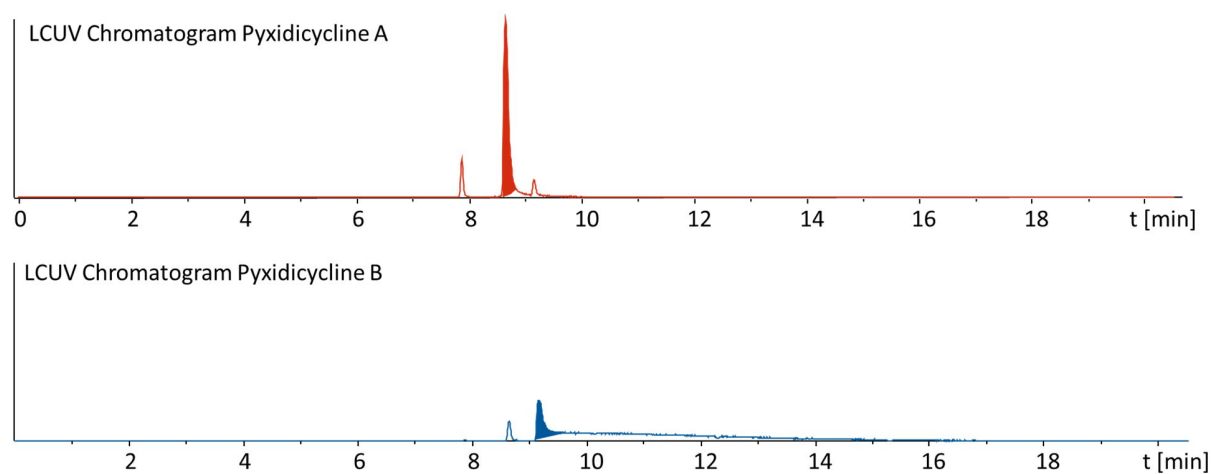***Figure S 17.*** *LCUV Chromatograms of the two Pyxidicyclines for purity analysis by liquid chromatography, the first peak belonging to the formate ester, the second to the free pyxidicyclines*
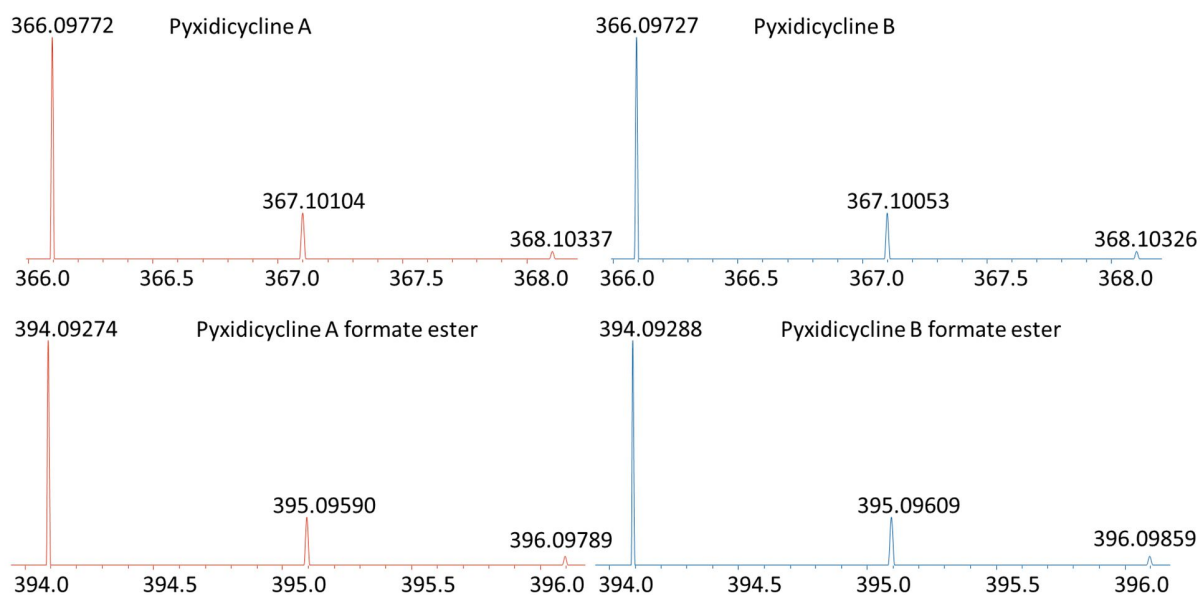
Figure S 18. *High resolution masses of pyxidicycline A and B [M+H]$^+$ in m/z units as well as the corresponding formate esters acquired on our maXis4G qTOF high resolution mass spectrometer*

S 2.3.6   Crystallization parameters for the pyxidicycline A and B

The pure pyxidicyclines A and B are taken up in a minimum of formic acid to an approximate mass concentration of 5 mg/ml and filled into a 2 ml glass vial. The solutions are sonicated for a few minutes in an ultrasound bath until no visible residues remain. The solutions are subsequently placed into the refrigerator with an unscrewed cap. After two days one can observe crystallization of pyxidicycline B as dark red needle shaped crystals. After one week one starts to observe crystallization of pyxidicycline A as dark red needle shaped crystals.

S 2.3.7   Crystal structure parameters obtained through single-crystal X-ray diffraction experiments

The compounds crystallized from formic acid solution as formic acid esters that formed needle shaped dark red crystals. The data were collected at low temperature on a BrukerAXS X8Apex CCD diffractometer operating with graphite-monochromatized Mo Kα radiation. Frames of 0.5° oscillation were exposed; deriving reflections in the θ range of 2 to 27° with a completeness of ~99%. Structure solving and full least-squares refinement with anisotropic thermal parameters of all non-hydrogen atoms were performed using SHELX [12]. Crystal structure parameters for pyxidicycline B: orthorhombic, Pca2$_1$, a=7.0231(8) b=17.534(2) c=14.682(2)Å ; and for pyxidicycline A: triclinic, P-1, a=7.090(1) b=7.910(2) 17.330(3)Å, α=89.56(1) β=84.80(1), γ=74.66(1). All relevant data concerning the crystal structures can be found in this section. Crystallographic data for the structure have been deposited with the Cambridge

Crystallographic Data Centre, CCDC, 12 Union Road, Cambridge CB21EZ, UK. Copies of the data can be obtained free of charge on quoting the depository number CCDC 1831467 and CCDC 1831468. www.ccdc.cam.ac.uk/data_request/cif

S 2.3.8   NMR based structure elucidation of the anthraquinone precursors

The chemical structures of the anthraquinone secondary metabolite precursors were determined via multidimensional NMR analysis. ^{1}H NMR, ^{13}C NMR and 2D spectra were recorded at 700 MHz (^{1}H)/175 MHz (^{13}C) conducting an Ascend 700 spectrometer using a cryogenically cooled triple resonance probe (Bruker Biospin, Rheinstetten, Germany). Samples were dissolved in DMSO-d$_6$, Acetone-d$_6$ and Chrloroform-d. Chemical shifts are reported in ppm relative to TMS, the solvent was used as the internal standard. NMR data are presented in tabulated form by ascending ^{13}C shifts.
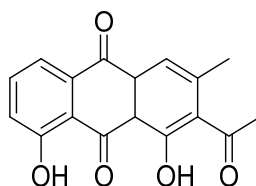
Figure S 19. *Structure formula of anthraquinone 296*

Table S 17. Tabulated NMR data of the anthraquinone 296 in DMSO-d$_6$

^{1}H	J, Multiplicity, Nb. of H	^{13}C	H-H Cosy	HMBC	ROESY
2.38	s, 3H	20.5	7.66	114.3; 122.5; 136.4; 105.8; 159.6; 203.2	7.66
2.59	s, 3H	32.1	7.66	203.1; 136.4	-
-	-	114.3	-	-	-
-	-	116.0	-	-	-
7.82	1.15; 7.45, dd, 2H	120.5	7.30; 7.69; 7.66	116.0; 125.2; 181.6	-
7.66	7.45; 8.48, dd, 2H	137.6	2.38, 7.82	114.3; 136.5; 181.6; 203.1; 20.5	2.38
7.3	1.2; 8.48, dd, 2H	125.2	7.69, 7.82	116.0; 120.6; 136.4	-
-	-	133.3	-	-	-
-	-	133.6	-	-	-
-	-	136.4	-	-	-
7.69	s, 1H	122.5	7.30; 7.69; 7.82	162.8; 133.6	-
-	-	145.8	-	-	-
-	-	159.6	-	-	-
-	-	162.8	-	-	-
-	-	181.6	-	-	-
-	-	192.8	-	-	-
-	-	203.1	-	-	-
11.96	s, 1H	-	7.69	116.0; 125.2; 162.8; 120.0; 192.8	-
12.34	s, 1H	-	7.66; 2.59	114.3; 136.4; 159.6; 20.5; 192.8	-

2D NMR data shown is in good agreement with published structure of protetrone [13].
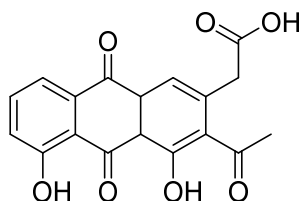
Figure S 20. Structure formula of anthraquinone 340

Table S 18. Tabulated NMR data of the anthraquinone 340 in Acetone-d$_6$

^{1}H	J, Multiplicity, Nb. of H	^{13}C
2.61	s, 3H	31.7
3.93	s, 2H	38.7
7.82	1.4; 7.0, dd, 1H	120.2
7.79	7.0; 8.0,dd, 1H	122.9
7.4	1.0;8.0, dd, 1H	125.2
7.87	s, 1H	138.4

Acquired NMR data in good agreement with NMR data published by Zhang et al. [14].
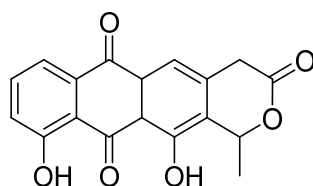
Figure S 21. *Structure formula of anthraquinone 324*

Table S 19. *Tabulated NMR data of the anthraquinone 324 in Chloroform-d*

^{1}H	J, Multiplicity, Nb. of H	^{13}C	H-H Cosy	HMBC
1.66	d	20.8	5.98	73.7, 130.2
3.86 + 3.82	2 x s	34.6	7.63	118.8, 130.4, 139.3, 167.9
5.98	6.90, q, 1H	73.7	1.66, 3.82	20.7, 130.4, 139.3, 157.6, 167.9
-		114.6		
-		115.6		
7.63	s, 1H	118.6	3.86, 3.82	34.6, 114.6, 130.2,181.5,197.2
7.83	0.9; 8.5, dd, 1H	120.4	7.69,7.31	115.6, 125,4, 162.8, 181.5,192.7
7.31	1; 8.5, dd, 1H	125.1	7.69, 7.83	115.6, 120.4, 162.8, 192.7
-		130.2		
-		133.2		
7.69	7.4; 8.5, dd, 1H	137.6	7.31, 7.83,11.94	115.6, 133.2, 162.8, 181.5
-	-	139.3	-	-
-	-	157.6	-	-
-	-	167.9	-	-
-	-	162.8	-	-
-	-	181.5	-	-
-	-	192.7	-	-
-	-	197.2	-	-
11.94	s, 1H	-	7.69	115.6, 125.0, 137.7, 162.8, 192.7
12.46	s, 1H	-	-	114.6, 130.2, 139.3, 157.6, 192.7

2D NMR data obtained in from anthraquinone 324 in good agreement with published NMR data for ekatetrone [15].

S 2.3.9   Tandem MS based structure assignment of anthraquinone 340

Due to poor stability of anthraquinone 340 in most NMR solvents, which led to decarboxylation of the compound to anthraquinone 296 via decarboxylation, only a ^{1}H and HSQC spectrum were recorded for this compound. To ascertain that the seen anthraquinone possesses the proposed structure, the first fragments of the LC-MS^2 spectrum recorded on the HPLC coupled maXis 4G qTOF spectrometer (section S 3.3) are interpreted.
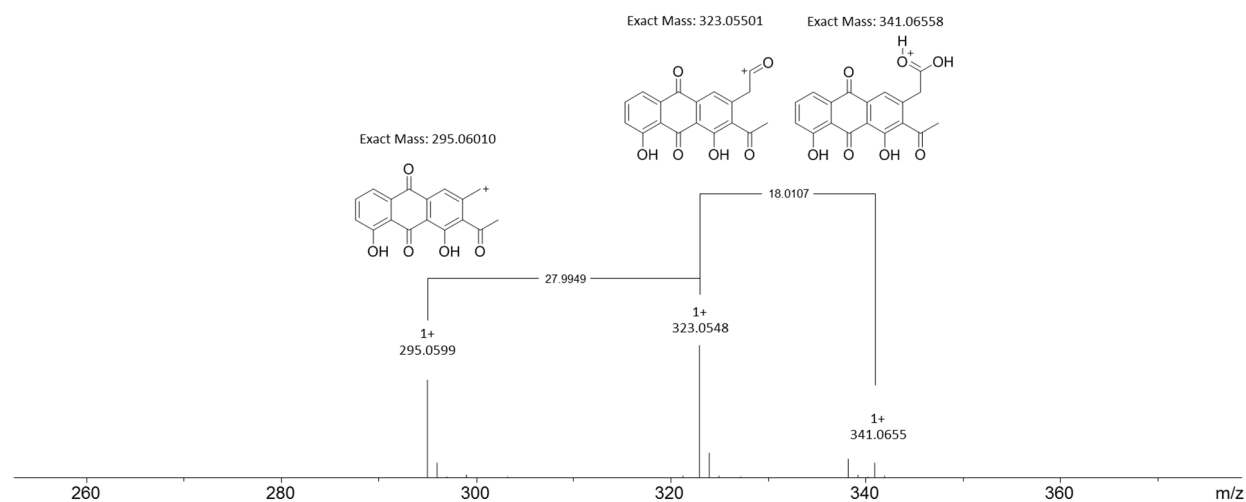
Figure S 22. *MS² spectrum of anthraquinone 340 with proposed gas phase dissociation reactions acquired during an LC-MS² run of* Pyxidicoccus fallax *And48* pcyA::pSBtn5 pcyA *extract on the HPLC coupled maXis 4G qTOF spectrometer*

One can clearly see the MS² fragments fitting the proposed structure that is also in accord with the reference NMR data. Furthermore, as this scaffold would indeed favor decarboxylation in solution to anthraquinone 296 as it is observed during longer NMR experiments one can safely assume the anthraquinone 340 to have the presented structure.

S 2.3.10 Tandem MS based assignment of the incorporated deuterium atoms into Pyxidicycline A

10 mmol 1,2 Ethanolamine $^{13}C_2$, 2,2-Glycine-d_2 and 2,3,3-serine-d3 are fed into a 25 ml CTT medium culture of *M. xanthus* attB :: attP pClY pyxidicycline cluster in order to observe isotopic labelling of the pyxidicyclines. Extracts were prepared according to standard protocol (section S 3.2) and analyzed via standardized UPLC-MS and UPLC-MS² (section S 3.3). Only feeding of 3,3,2-serine-d_3 into *M. xanthus* DK1622 attB::attP pClY pyxidicycline cluster reveals strong labelling of the pyxidicyclines. To prove the incorporation into the side chain of the pyxidicyclines as it was hypothesized in the biosynthesis proposal backed up with labelling experiments the labelled compound is subjected to tandem MS measurements.
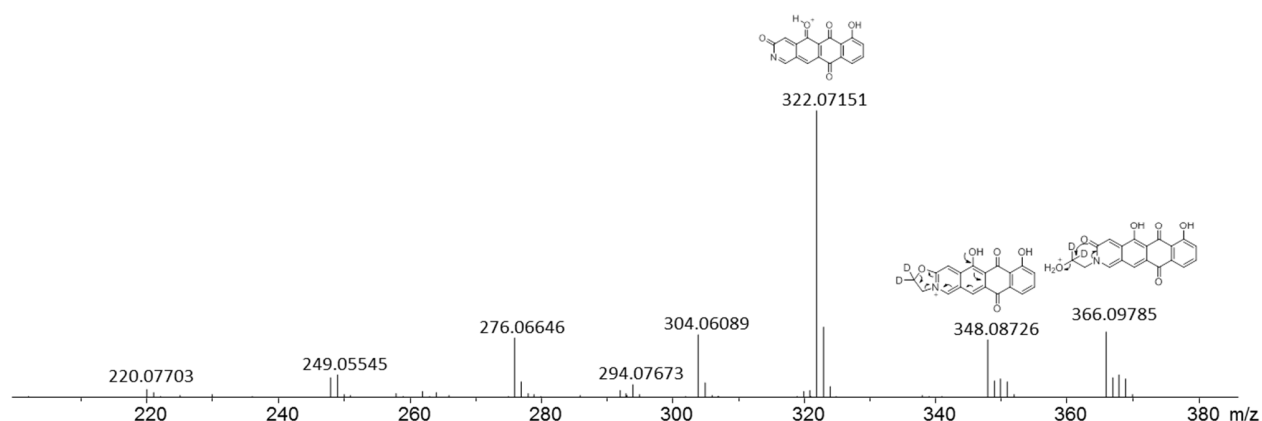
Figure S 23. *LC-MS² CID-fragmentation pattern of d₂ labelled pyxidicycline A from M. xanthus DK1622 attB :: attP pClY pyxidicycline with proposed gas phase reaction acquired on the HPLC coupled maXis 4G qTOF spectrometer*

As the molecule fragment with the mass of 322.0715 Daltons lost the deuterium atoms as visible by the isotope pattern one can assume that both deuterium atoms are present on this 2-hydroxyethyl side chain. This fits well with the molecule's structure as one would assume the linear 2-hydroxyethyl side chain to be the most labile part in CID based MS² experiments that likely fragments according to the gas phase reactions proposed in Figure S 23. One can therefore experimentally prove the incorporation of serine into the structure.

S 2.3.11   NMR based structure elucidation of the pyxidicyclines

As the pyxidicyclines are only soluble in DMSO and DMF that both unfortunately lead to degradation of the compound over time, NMR data is acquired in formic acid-d₂. The pyxidicyclines readily and reversibly form formate esters that are soluble in formic acid. ¹H NMR, ¹³C NMR and 2D spectra were recorded at 700 MHz (¹H)/175 MHz (¹³C) conducting an Ascend 700 spectrometer using a cryogenically cooled triple resonance probe (Bruker Biospin, Rheinstetten, Germany). Samples were dissolved in formic acid-d₂ Chemical shifts are reported in ppm relative to TMS, the solvent was used as the internal standard. NMR reference information for formic acid is taken from Roy Hoffmann.[16] Carbon numbering for NMR signal assignment purposes is done in a biosynthetic logic meaning that the carbon first integrated into the structure obtains the first number.
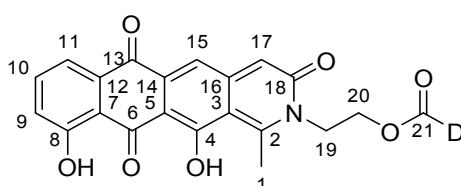
Figure S 24. *Carbon numbering for NMR assignment purposes for pyxidicycline A formate ester*

Table S 20. *Tabulated NMR data of pyxidicycline A formate ester acquired in formic acid-d₂. Carbon numbering is done according to the scheme in Figure S 24.*

C Number	1H	J, Multiplicity, Nb of H	13C	H-H Cosy	HMBC
1	3.57	s, 3H	21.4	7.56	108.5, 116.2, 157.5, 168.8
2	-	-	125.5	-	-
3	-	-	116.2	-	-
4	-	-	168.8	-	-
5	-	-	109.3	-	-
6	-	-	190.6	-	-
7	-	-	116	-	-
8	-	-	162.7	-	-
9	7.41	n.d.; 8.3, dd, 1H	125.7	7.81, 7.86	116.0, 120.6, 162.0
10	7.81	7.3; 8.3, dd, 1H	137.8	7.41,7.86	116.0, 133.6, 162.0
11	7.86	n.d.; 7.3, dd, 1H	120.6	7.41, 7.81	116.0, 125.7, 137.8, 182.2, 190.6
12	-	-	133.6	-	-
13	-	-	182.2	-	-
14	-	-	126.6	-	-
15	8.05	s, 1H	118.4	3.57, 7.56	108.5, 109.3, 116.2, 145.5, 182.2, 190.6
16	-	-	145.4	-	-
17	7.56	s, 1H	108.5	8.05	116.2, 125.5, 126.6 157.5, 162.5, 168.8
18	-	-	157.5	-	-
19	5.14	5.30, t, 2H	47.4	4.81	60.8, 157.5, 162.5
20	4.81	5.30, t ,2H	60.8	5.14	47.4,162.5,163.3
FA C21	-	-	163.3	-	-
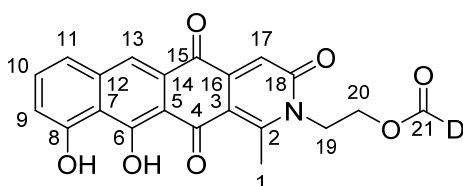
Figure S 25. *Carbon numbering for NMR assignment purposes for pyxidicycline B formate ester*

Table S 21. *Tabulated NMR data of pyxidicycline B formate ester acquired in formic acid-d₂. Carbon numbering is done according to the scheme in Figure S 25.*

C Number	1H	J, Multiplicity, Nb of H	13C	H-H Cosy	HMBC	ROESY
1	3.23	s, 3H	19.1	7.38	113.2, 114.8, 142.2, 159.8, 164.5, 186.4	4.74, 4.65
2	-	-	159.8	-	-	-
3	-	-	113.2	-	-	-
4	-	-	186.4	-	-	-
5	-	-	124.1	-	-	-
6	-	-	165.1	-	-	-
7	-	-	108.7	-	-	-
8	-	-	157.2	-	-	-
9	7.1	7.9, d, 1H	115.8	7.54, 7.64	116.0, 122.7, 157.2	-
10	7.64	7.9; 8.0, dd, 1H	133.5	7.10, 7.54	115.8, 122.7, 137.2, 157.2	-
11	7.54	8.0, d, 1H	122.7	7.10, 7.64	115.8, 123.0, 137.2, 157.2, 165.1	-
12	-	-	137.2	-	-	-
13	8.15	s, 1H	123	7.53	108.7, 115.8, 122.7, 124,1, 137.2, 157.2, 165.1, 181.3, 186.4	-
14	-	-	n.d.	-	-	-
15	-	-	181.2	-	-	-
16	-	-	142.2	-	-	-
17	7.38	s, 1H	114.8	3.23	113.2, 142.2, 164.5, 181.6	-
18	-	-	164.5	-	-	-
19	4.74	5.52, t, 2H	44.5	4.65	61.1, 159.8, 164.5	3.23
20	4.65	5.52, t, 2H	61.2	4.74	44.5, 163.0	3.23
FA C21	-	-	163	-	-	-

S 2.3.12 Comparative evaluation of the different strains pyxidicycline and pyxidicycline precursor production potential

Table S 22 shows the relative and absolute peak areas measured on the UHPLC coupled maXis 4G spectrometer system by injection of 0.5 µL of crude extract of the respective species grown and extracted according to standard protocol described in section 3.1 to 3.3. The last four columns contain quotients to illustrate the effect of expression in different hosts or gene deletions upon pyxidicycline expression.

Table S 22. *Comparative peak area analysis measured in [cts x min] of the pyxidicyclines and its precursors in all generated producer strains measured by integration the [M+H]⁺ EIC in the standard LC-MS chromatogram setting (section S 3.3). The two anthraquinone metabolites are measured as LC-MS EIC peak area at 341.06 [M+H]⁺, 325.07 [M+H]⁺ and 297.075 [M+H]⁺, the two pyxidicyclines are measured as LCMS EIC peak areas at 366.10 [M+H]⁺.*

Producer strain	Area Pyx A	Area Pyx B	Area AQ 340	Area AQ 324	Area AQ 296	Area Pyx A / Area AQ 340	Area Pyx B / Area AQ 340	Area Pyx A / Area AQ 296	Area Pyx B / Area AQ 296	Area Pyx A / Area Pyxi B
P. fallax And48 pcyA :: pSBtn5 *pcyA*	-	-	1135792	119418	226789	-	-	-	-	-
P. fallax And48 pcyA :: pFPVan *pcyA*	-	-	129491	13364	151052	-	-	-	-	-
P. fallax And48 pcyA :: pFPVan *pcyA*; pcyJ :: pFPtettn5 *pcyJ*	1905806	1859927	15991	-	1119844	119.18	116.31	1.70	1.66	1.02
M. xanthus DK1622 attB :: attP Pyxidicycline Cluster	1157822	2240664	126015	-	540949	9.19	17.78	2.14	4.14	0.52
M. xanthus DK1622 attB :: attP Pyxidicycline Cluster Δ *pcyJ*	-	-	1173697	-	1580399	-	-	-	-	-
M. xanthus DK1622 attB :: attP Pyxidicycline Cluster Δ *pcyK*	170034	137566	1156196	-	2065290	0.15	0.12	0.08	0.07	1.24
M. xanthus DK1622 attB :: attP Pyxidicycline Cluster Δ *pcyL*	153049	132066	1202235	-	2301855	0.13	0.11	0.07	0.06	1.16
M. xanthus DK1622 attB :: attP Pyxidicycline Cluster T7A1 *pcyO*	1400366	3180800	365870	-	399734	3.83	8.69	3.50	7.96	0.44
S. aurantiaca DW 4/3-1 attB :: attP Pyxidicycline Cluster	75069	290585	102942	-	75896	0.73	2.82	1.36	3.83	0.26

S 2.4   Bioactivity profiling of the anthraquinones and its precursors

S 2.4.1   Cell based biological assays to valuate MIC values on bacteria and cancer cell lines

S 2.4.1.1   Antimicrobial Assay

All microorganisms were handled according to standard procedures and were obtained from the German Collection of Microorganisms and Cell Cultures (*Deutsche Sammlung für Mikroorganismen und Zellkulturen*, DSMZ) or were part of our internal strain collection. For microdilution assays, overnight cultures of Gram-positive bacteria in Müller-Hinton broth (0.2 % (w/v) beef infusion, 0.15 % (w/v) corn starch, 1.75 % (w/v) casein peptone; pH 7.4) were diluted in the growth medium to achieve a final inoculum of ca. 10^6 cfu ml^{-1}. Serial dilutions of pyxidicyclines and the anthraquinone precursors were prepared from freshly prepared DMSO or Methanol stocks in sterile 96-well plates. The cell suspension was added and microorganisms were grown on a microplate shaker (750 rpm, 37°C and 16 h). Growth

inhibition was assessed by visual inspection and given MIC (minimum inhibitory concentration) values are the lowest concentration of antibiotic at which no visible growth was observed.

S 2.4.1.2   **Cytotoxic Activity**

Cell lines were obtained from the German Collection of Microorganisms and Cell Cultures (Deutsche Sammlung für Mikroorganismen und Zellkulturen, DSMZ) or were part of our internal collection and were cultured under conditions recommended by the depositor. Cells were seeded at 6 x 10^3 cells per well of 96-well plates in 180 µl complete medium and treated with pyxidicyclines in serial dilution after 2 h of equilibration. Each compound was tested in duplicate as well as the internal solvent control. After 5 d incubation, 20 µl of 5 mg ml^{-1} MTT (thiazolyl blue tetrazolium bromide) in PBS was added per well and it was further incubated for 2 h at 37°C. The medium was discarded and cells were washed with 100 µl PBS before adding 100 µl 2-propanol/10 N HCl (250:1) in order to dissolve formazan granules. The absorbance at 570 nm was measured using a microplate reader (Tecan Infinite M200Pro), and cell viability was expressed as percentage relative to the respective methanol control. IC$_{50}$ values were determined by sigmoidal curve fitting.

S 2.4.2   **Topoisomerase inhibition assays performed to determine the target of the pyxidicyclines**

All topoisomerase related inhibition assays were performed using topoisomerase inhibition evaluation kits for *E. coli* topoisomerase II, *E. coli* topoisomerase IV and human topoisomerase I supplied by Inspiralis (Inspiralis ltd., Norwich, U.K.). The kits were used according to the manuals. All compounds were tested as freshly prepared DMSO solutions. Pyxidicyclines were assayed to be stable in DMSO for about 12 hours. Topoisomerase inhibition assays were evaluated by agarose gel electrophoresis using 0.8 % agarose gels in 1x TAE buffer. Total volume of the assay reaction was 30 µL. All assays are carried out in a 96 well micro titer plate.

Inspiralis *E. coli* DNA gyrase inhibition assay
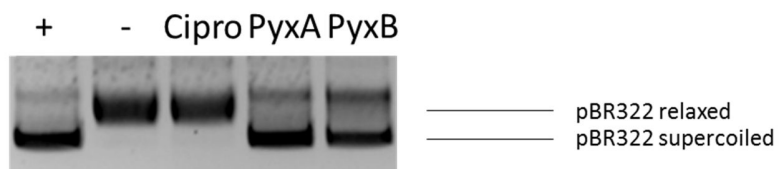
positive control, negative control, Ciprofloxacin 60μg/ml
Pyxidicycline A 50μg/ml, Pyxidicycline B 50μg/ml

Figure S 26. E. coli *DNA gyrase inhibition assay, positive control is 0.5% DMSO, negative control is assay without enzyme; Ciprofloxacin, pyxidicycline A and pyxidicycline B are supplied in DMSO to a total concentration of 0.5% DMSO and the stated end concentrations in the assay*

Inspiralis *E. coli* topoisomerase IV inhibition Assay
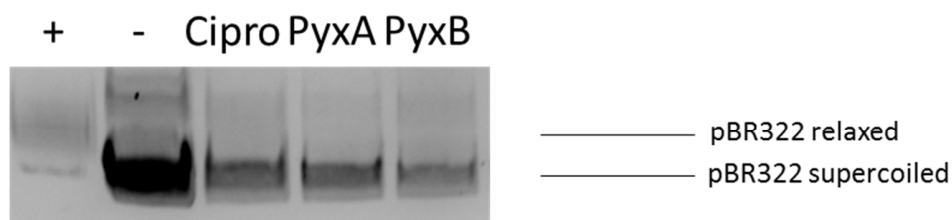
+   -   Cipro PyxA PyxB

————— pBR322 relaxed
————— pBR322 supercoiled

positive control, negative control, Ciprofloxacin 60µg/ml,
Pyxidicycline A 50µg/ml, Pyxidicycline B 50µg/ml

E. coli topoisomerase IV
IC50 Pyxidicycline A

pBR322 relaxed   ———
pBR322 supercoiled ———

c [µg/ml]   50   25   12.5 6.25 3.2   1.6   0.8   0.4

E. coli topoisomerase IV
IC50 Pyxidicycline B

pBR322 relaxed   ———
pBR322 supercoiled ———

c [µg/ml]   50   25   12.5 6.25 3.2   1.6   0.8   0.4
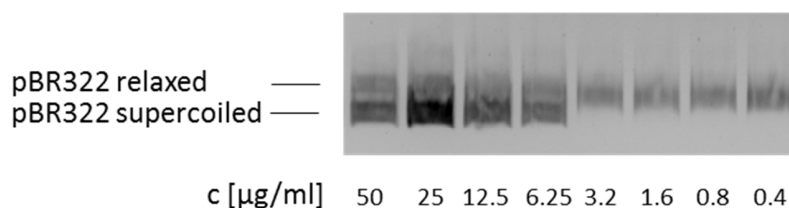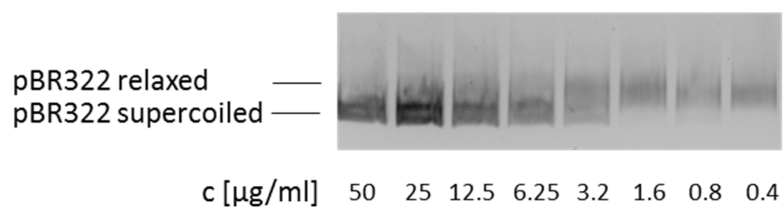
Figure S 27. E. coli *DNA topoisomerase IV inhibition assay, positive control is 0.5% DMSO, negative control is assay without enzyme; Ciprofloxacin, pyxidicycline A and pyxidicycline B are supplied in DMSO to a total concentration of 0.5% DMSO and the stated end concentrations in the assay*

Inspiralis Human topoisomerase I inhibition assay
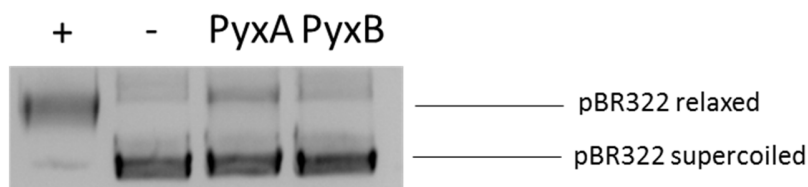
positive control, negative control
Pyxidicycline A 50µg/ml, Pyxidicycline B 50µg/ml

Human topoisomerase I
IC50 Pyxidicycline A

| c [µg/ml] | 50 | 25 | 12.5 | 6.25 | 3.2 | 1.6 | 0.8 | 0.4 | 0.2 |

Human topoisomerase I
IC50 Pyxidicycline B

| c [µg/ml] | 50 | 25 | 12.5 | 6.25 | 3.2 | 1.6 | 0.8 | 0.4 | 0.2 | 0.1 | 0.05 |
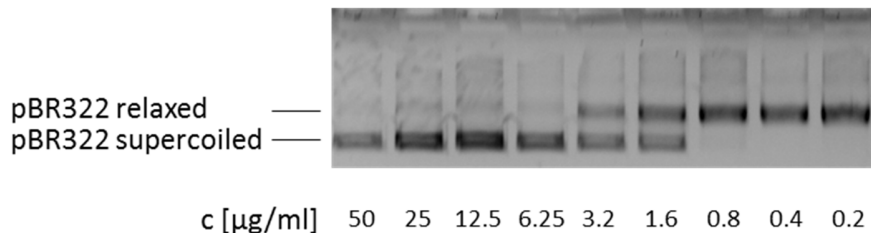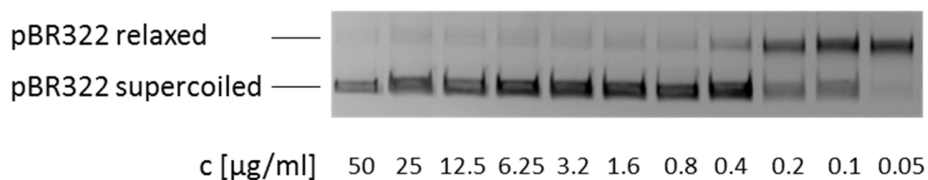
Figure S 28. *Human DNA topoisomerase I inhibition assay, positive control is 0.5% DMSO, negative control is assay without enzyme, pyxidicycline A and pyxidicycline B are supplied in DMSO to a total concentration of 0.5% DMSO and the stated end concentrations in the assay*

S 2.5 References

[1] P. Youderian, N. Burke, D. J. White, P. L. Hartzell, *Mol. Microbiol.* **2003**, *49*, 555.

[2] O. Bilyk, O. N. Sekurova, S. B. Zotchev, A. Luzhetskyy, *PLoS ONE* **2016**, *11*, e0158682.

[3] V. Magrini, C. Creighton, P. Youderian, *J. Bacteriol.* **1999**, *181*, 4050.

[4] S. Shah, J. G. Heddle, *Appl. Biochem. Biotechnol.* **2014**, *98*, 9545.

[5] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, *Nat. Protoc.* **2015**, *10*, 845.

[6] J. Dreier, C. Khosla, *Biochemistry* **2000**, *39*, 2088.

[7] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales et al., *Nucleic acids research* **2017**, *45*, D200-D203.

[8] R. C. Edgar, *Nucleic Acids Res.* **2004**, *32*, 1792.

[9] L. B. Pickens, Y. Tang, *J. Biol. Chem.* **2010**, *285*, 27509.

[10] S. Gullón, C. Olano, M. S. Abdelfattah, A. F. Braña, J. Rohr, C. Méndez, J. A. Salas, *Appl. Environ. Microbiol.* **2006**, *72*, 4172.

[11] van den Berg,R. A., H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, van der Werf,M. J., *BMC Gen.* **2006**, *7*, 142.

[12] G. M. Sheldrick, *Acta Crystallogr. Sect. A* **2008**, *64*, 112.

[13] J. R. D. McCormick, E. R. Jensen, *Adv. Ceram. Mater.* **1968**, *90*, 7126.

[14] W. Zhang, Y. Li, Y. Tang, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 20683.

[15] V. Prikrylova, M. Podojil, P. Sedmera, J. Vokoun, Z. Vanek, C. H. Hassal, *J Antibiot (Tokyo)* **1978**, *31*, 855.

[16] R. E. Hoffman, *Magnetic resonance in chemistry : MRC* **2006**, *44*, 606.

Chapter 3

Novel β-methoxymethacrylate natural products uncovered by statistics-based mining of the *Myxococcus fulvus* secondary metabolome

Fabian Panter[a,b], Daniel Krug[a,b] and Rolf Müller*[a,b]

Affiliation

[a]Department Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmaceutical Biotechnology, Saarland University, Campus E8.1, 66123 Saarbrücken, Germany

[b] German Centre for Infection Research (DZIF), Partner Site Hannover–Braunschweig, Hannover–Braunschweig, Germany

Contributions and Acknowledgements

Author's effort:

The author significantly contributed to the conception of this study, designed and performed experiments, evaluated and interpreted resulting data. The laboratory and *in silico* work regarding gene cluster analysis, isolation and full structure elucidation of the fulvuthiacenes and structure validation of the created myxothiazol derivatives, as well as the heterologous expression of *ftaG+ftaH* and *ftaM+ftaN* were performed by the author. The author further contributed to the study creating the fulvuthiacene inactivation mutant. Additionally, the author contributed significantly to conceiving and writing this manuscript.

Contributions by others:

Daniel Krug contributed to conception and supervision of this study, contributed to conceiving, writing and editing of the manuscript. Rolf Müller contributed by supervision of the project as well as editing and proofreading of the manuscript.

3.1   Abstract

    This study reports the uncovering of new myxobacterial natural products through comprehensive analysis of the *Myxococcus fulvus* secondary metabolome. Statistics-based mining of mass spectrometry data paved the way for full structure elucidation of two new secondary metabolites named fulvuthiacene A and B, and investigation of the underlying biosynthetic pathway revealed an evolutionary link between the fulvuthiacene hybrid polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) gene cluster and the related myxothiazol and melithiazol assembly lines. Detailed characterization of the post-PKS modification enzyme cascade responsible for the fulvuthiacenes' terminal β-methoxy-methylacrylate moiety was pursued by heterologous expression of these enzymes in the myxothiazol producer *S. aurantiaca* DW4/3-1. The discovery of fulvuthiacenes provides new insights into the overall structure-activity relationship picture for the β-methoxyacrylate class of respiratory chain inhibitors and might thus serve as starting point for the development of next-generation β-methoxymethacrylate fungicides.

3.2   Introduction

    Natural products discovery from microbes is currently undergoing a paradigm change, as genomics continues to reveal an impressive genome-encoded capacity for the production of secondary metabolites, outweighing the numbers of known compounds for many species. In parallel advances in analytical methods and instrumentation allow previously unaccomplished insights into the depth of microbial secondary metabolomes.[1] These developments in turn demand for improved data treatment techniques and new data mining approaches in order to transform the steadily increasing genomic and metabolomic data collections into the true chemical and biological characterization of novel natural products. Whereas screening campaigns in the past mainly uncovered compounds produced in large amounts or featuring abundant UV and/or mass spectrometry (MS) signals, there is now an urgent need for enabling methods to highlight also lower-abundance secondary metabolites including those that lack prominent analytical properties. Evidently, production titer or ease of detection is no *a priori* indicator for the value of newly discovered compounds in terms of structural novelty or biological activity. Thus, to overcome this conceptual bias, a number of studies aimed at establishing improved or alternative methods to mine the 'hidden' components of compound-rich microbial metabolomes have been conducted recently.[2–5] The approaches used can be roughly categorized as "gene-to-compound" or "compound first", whereas these genome-mining techniques are often paired with mass spectrometry workflows based on comprehensive metabolite profiling in order to highlight compounds of interest.[5,6] Nevertheless, a staggering discrepancy

remains for many microbes between the characterized secondary metabolome and the potential for secondary metabolite production as predicted by genome-based analysis, routinely accessible through bioinformatics tools such as antiSMASH.[7–9] In this study we describe an analytical procedure to uncover new secondary metabolites from myxobacteria, which are often masked in the complex bacterial extracts due to low productivity, or represented by low-intensity MS signals. Key features of the devised workflow include improved fragmentation coverage for such inconspicuous analytes in tandem-MS measurements through statistics-based filtering of relevant MS signals attributable to bacterial metabolism. This is performed prior to the evaluation of candidate compounds in terms of structural similarity through spectral networking.[2,3,10] Application of our method to Myxococcus fulvus MCy9280 revealed the fulvuthiacenes as a new family of secondary metabolites characterized by an intriguing terminal β-methoxy-methylacrylate moiety. In addition, following the identification of the fulvuthiacene biosynthesis gene cluster we complemented our compound discovery with a detailed investigation of the underlying biochemistry.

3.3   Results and Discussion

Myxobacteria frequently expose a striking discrepancy between the number of biosynthetic gene clusters detectable in their genomes and the number of known secondary metabolites, a finding well exemplified by *M. fulvus* MCy9280.[8] Previously, only stipiamide (phenalamide) and rhizopodin were detected from MCy9280 although the strains' genome encodes more than 20 NRPS, PKS or NRPS-PKS hybrid pathways.[8,11] This strain was thus considered a worthy subject for our endeavors to reduce the gap between genome-deduced potential and secondary metabolite 'reality' through a MS-based secondary metabolite identification workflow.

3.3.1   Statistics-based metabolome mining to reveal secondary metabolite families

Microbial secondary metabolites often appear as families of multiple derivatives produced by the same biosynthetic pathway, as it is seen e.g. for myxothiazols and melithiazols - just like with most other natural product classes.[12,13] If these derivatives share the same backbone structure with slightly different decorations, one can presume the presence of novel secondary metabolites by observation of grouping patterns during similarity-driven clustering of tandem mass spectrometry (MS²) spectra acquired from a bacterial crude extract. In recent years, different methods to cluster mass spectra according to similarity such as spectral networking, fingerprinting and *in silico* fragmentation methods have been introduced, where particularly the former has been already used for the disclosure of novel secondary metabolite

scaffolds.[3,4,14] These methods are typically applied to untargeted MS² data. Nevertheless, this common practice comes with severe limitations.[15] Since biological samples such as bacterial extracts often contain abundant medium derived MS signals, picking relevant features for MS² fragmentation using intensity based automatic precursor selection is not sufficient. Therefore, as a preselection step to optimize the acquisition for a maximum number of meaningful MS² spectra with regard to the subsequent search for secondary metabolite families, we intended to create precursor lists that selectively contain mass spectrometric features related to compounds produced by the myxobacterium. To account for this requirement, the myxobacterium *Myxococcus fulvus* MCy9280 investigated in this work was grown in triplicates in VYGS medium and bacterial extracts as well as triplicates of VYGS medium "blanks" were prepared according to a standardized protocol (section S 3). Both sets of extracts were analyzed twice with an ultra-performance liquid chromatography (UPLC) coupled high-resolution time of flight (hrToF) spectrometer in order to obtain a total of twelve LC-MS data sets (section S 3). These data sets were further processed by molecular feature detection and subjected to statistical filtering through principal component analysis (PCA), which allows for differentiation between the approximate bacterial metabolome and media compounds. The tabulated data generated from this secondary metabolome profile analysis were subsequently processed to retain only those molecular features occurring in all replicate bacterial extracts but in none of the blank extracts, provided that they exceed a precursor ion intensity threshold. This double filtering procedure ensures that precursor ions selected for MS² fragmentation are associated with the myxobacterial bacterial metabolome.[2] At the same time, candidate signal intensity offers a realistic chance for producing meaningful MS² data. The set of candidate features obtained by this method was transferred to a scheduled precursor list (SPL) for targeted MS² experiments that specifically cover the supposable myxobacterial metabolome, leaving out medium-derived MS signals. The resulting scheduled precursor list-based MS² chromatograms were then fed into Global Natural Products Social molecular networking (GNPS) (section S 3).[3] The resulting spectral networks were visualized with cytoscape and all single nodes possessing no spectral network neighbors were removed (Figure 1, section S 3) in line with the presented rationale that secondary metabolites mostly occur as compound families. In principle, spectral networking of MS² data using automatic precursor selection (auto-MS²) already shows the basic capacity of the workflow to highlight novel secondary metabolite scaffolds from myxobacteria (Figure 1, section S 3). However, the biggest spectral network obtained from that data contains mainly medium-derived peptides (section S 3.1). Thus, a significant amount of auto-MS² spectra acquisition time is unnecessarily dedicated to recording medium-related MS² spectra, whilst missing co-eluting secondary metabolites.
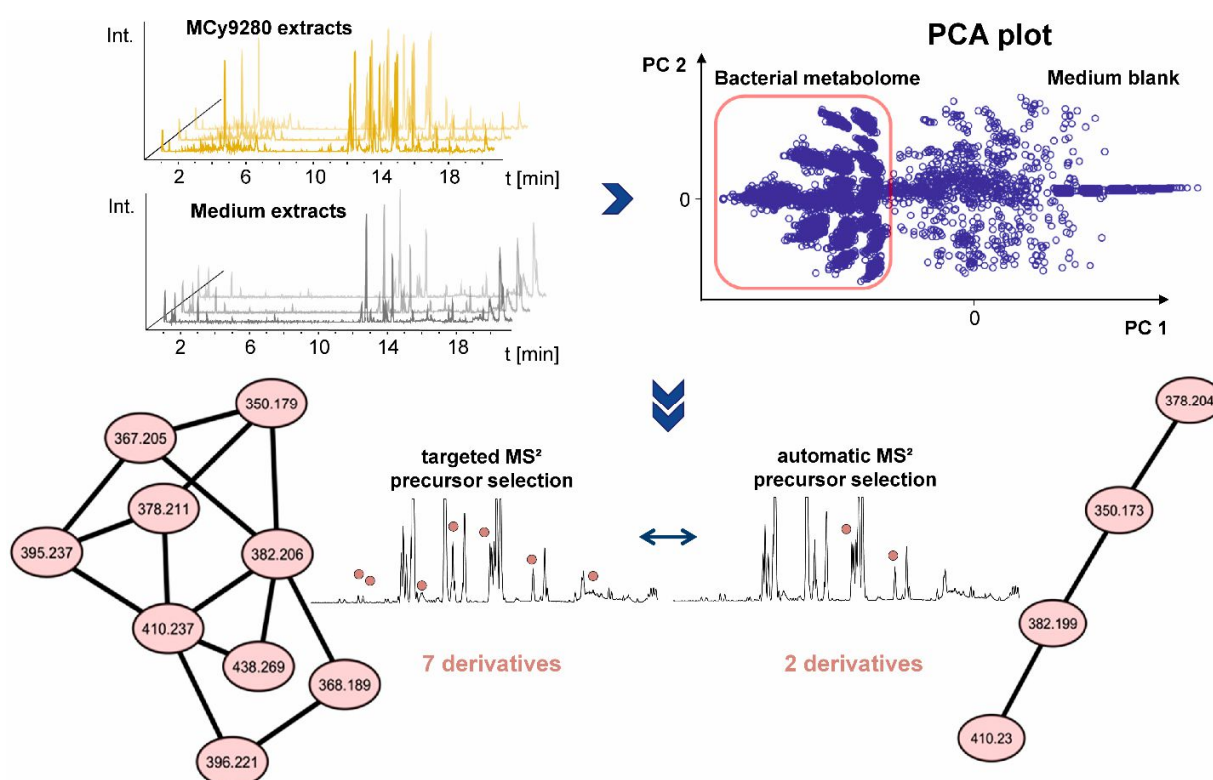
Figure 1. *LC-MS² data acquisition, principal component (PC) based MS feature discrimination workflow and comparison between the spectral networks obtained from automatic precursor selection MS² (lower right) and SPL-derived MS² precursor selection (lower left).*

Furthermore, the auto-MS² setting makes it impossible to discern medium derived molecular features, which are regarded as not interesting, from the molecular features representing the bacterial metabolome. Consequently, data-independent MS² acquisition is inherently suboptimal. Nevertheless, in the case of MCy9280 extracts, this approach reveals distinct spectral networks, like the example spanning four MS² spectra highlighted in Figure 1 (lower right). In order to ensure that this networking cluster was not constructed from medium derived MS signals, specified precursor list derived MS² data were acquired and clustered accordingly. The previously highlighted spectral network is still present, indicating that it is indeed composed of bacterial secondary metabolite features, but importantly the second version of this network comprised significantly more precursor ions (Figure 1 lower left, section S 3). This finding is readily explained by the specified precursor list MS² method, which is able to pick precursors of lower intensity that are missed in the automatic precursor selection approach. The observed difference also confirms that metabolome-associated features do often exhibit lower intensity and would thus be systematically disregarded by automatic precursor selection, because of co-eluting higher-intensity matrix features in the LC-MS run.[2] Manual inspection of the precursors forming the target spectral network cluster above revealed the candidate compounds in Table 1, whereas it is worth noting that multiple features in the

spectral network (Figure 1, section S 3) may represent the same compound. The entity later characterized as fulvuthiacene A (**1**) for example is responsible for precursor ions 382.197 [M+H]$^+$ and 350.157 [M-MeOH+H]$^+$ whose MS^2 spectra obviously cluster. This explains the discrepancy between the size of the molecular network in Figure 1 and the number of found fulvuthiacene derivatives in Table 1.

Table 1. Members of the new fulvuthiacene compound family detected through MS^2 analysis in this study.

Name	Rt [min]	m/z [M+H]$^+$	calcd. sum formula
fulvuthiacene A (**1**)	14.3	382.197	$C_{20}H_{31}NO_4S$
fulvuthiacene B (**2**)	15.9	410.236	$C_{22}H_{35}NO_4S$
fulvuthiacene C (**3**)	17.2	438.267	$C_{24}H_{39}NO_4S$
desmethyl-fulvuthiacene A (**6**)	12.3	368.189	$C_{19}H_{29}NO_4S$
desmethyl-fulvuthiacene B (**7**)	14.0	396.221	$C_{21}H_{33}NO_4S$
prefulvuthiacene A (**4**)	11.3	367.205	$C_{19}H_{30}N_2O_3S$
prefulvuthiacene B (**5**)	13.2	395.236	$C_{21}H_{34}N_2O_3S$

3.3.2 Structure elucidation

The predicted sum formulae (Table 1) as well as a methanol loss as an in-source fragmentation reaction observed with all members of the highlighted spectral network suggested these molecular features to represent a new class of myxobacterial secondary metabolites. We therefore set out to isolate these compounds from a 12-liter fermentation of MCy9280 in VYGS medium. Compounds corresponding to the target masses of 382.197 and 410.236 [M+H]$^+$ were purified by a series of liquid-liquid extractions followed by HPLC separation. The two compounds fulvuthiacene A (**1**) and fulvuthiacene B (**2**) were obtained as white solids and subsequently structurally elucidated using NMR data (section S 4). Whereas the yield of additional minor derivatives of the fulvuthiacenes was by far not sufficient to allow structure elucidation by NMR experiments, we propose the structure of fulvuthiacene C (**3**) based on the MS^2 fragmentation patterns (Figure 2). Compounds **1** and **2**, which were structurally elucidated using multidimensional NMR spectroscopy, fragment similarly to the putative precursor ion of **3**. Furthermore, the three largest and most intense product ions are shifted by the mass difference of C_2H_4, which is indicative of an additional C_2H_4 unit in a fatty acyl side chain.
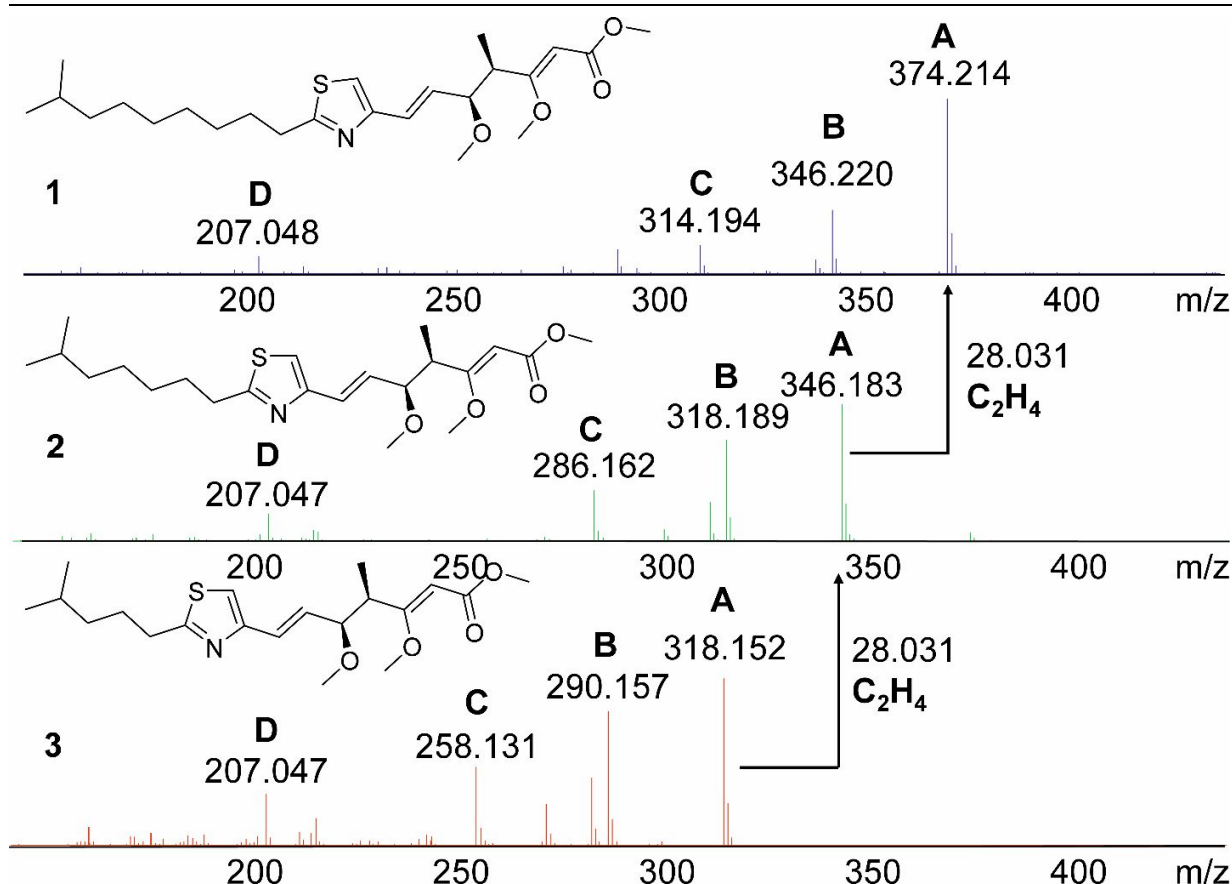
Figure 2. Structures of fulvuthiacenes A (**1**) and B (**2**) assigned by NMR and fulvuthiacene C (**3**) assigned by MS² spectra comparison. The fragmentation patterns show the [M - 2MeOH + H]⁺ ions (**A**), [M - 3MeOH + H]⁺ ions (**B**), [M - 3MeOH - CO + H]⁺ ions (**C**) and a $C_{11}H_{11}O_2S^+$ product ion with an m/z ratio of 207.047 Da that is always observed for the fulvuthiacenes.

This additional mass shift is equally seen between **2** and the putatively identified **3,** indicating again the observed mass shift difference of exactly 28.031 - which represents C_2H_4 - very likely to stem from the incorporation of an additional malonyl-CoA during biosynthesis of the side chain. Additionally, all three structures show a fragment ion with an exact mass of 207.048, which corresponds to $C_{11}H_{11}O_2S$ (Figure 2). Although the exact structure of this fragment remains elusive, the fact that it contains sulfur and two oxygen atoms proves its origin from the conserved central part of the scaffold. The proposed structure of the putatively identified **3** (Figure 2) also agrees with biosynthetic considerations (Figure 3). The other MS² spectra clustered in the spectral network of the fulvuthiacenes are likely biosynthesis intermediates and thus assigned to putative structures according to the biosynthesis scheme, also taking into account enzymatic steps responsible for post assembly line modifications (Figure 3). Unfortunately, these minor derivatives are produced in minute amounts rendering confirmation by NMR spectroscopy impossible.

3.3.3   Fulvuthiacene biosynthesis

With the new fulvuthiacene structure at hand, we hypothesized this compound family to be produced by a modular type 1 polyketide synthase (PKS) non-ribosomal peptide synthetase (NRPS) hybrid pathway. To elucidate the biosynthetic origin of these compounds, genomic DNA of the strain MCy9280 was sequenced and annotated by antiSMASH to predict secondary metabolite biosynthesis pathways encoded in the genome.[7] A candidate biosynthetic gene cluster (BGC) was identified, which can plausibly encode the fulvuthiacene assembly line according to retro-biosynthetic reasoning. As this putative BGC sequence contained two gaps, these gaps were PCR-amplified, subcloned and resequenced. The cluster was then assembled in full and biosynthetic domains were annotated. The *fta* gene cluster shares significant similarity on protein level to the myxothiazol and melithiazol assembly lines, but there are several important differences (Figure 7, section S 2, GenBank accession number: MH069655).[16,17]
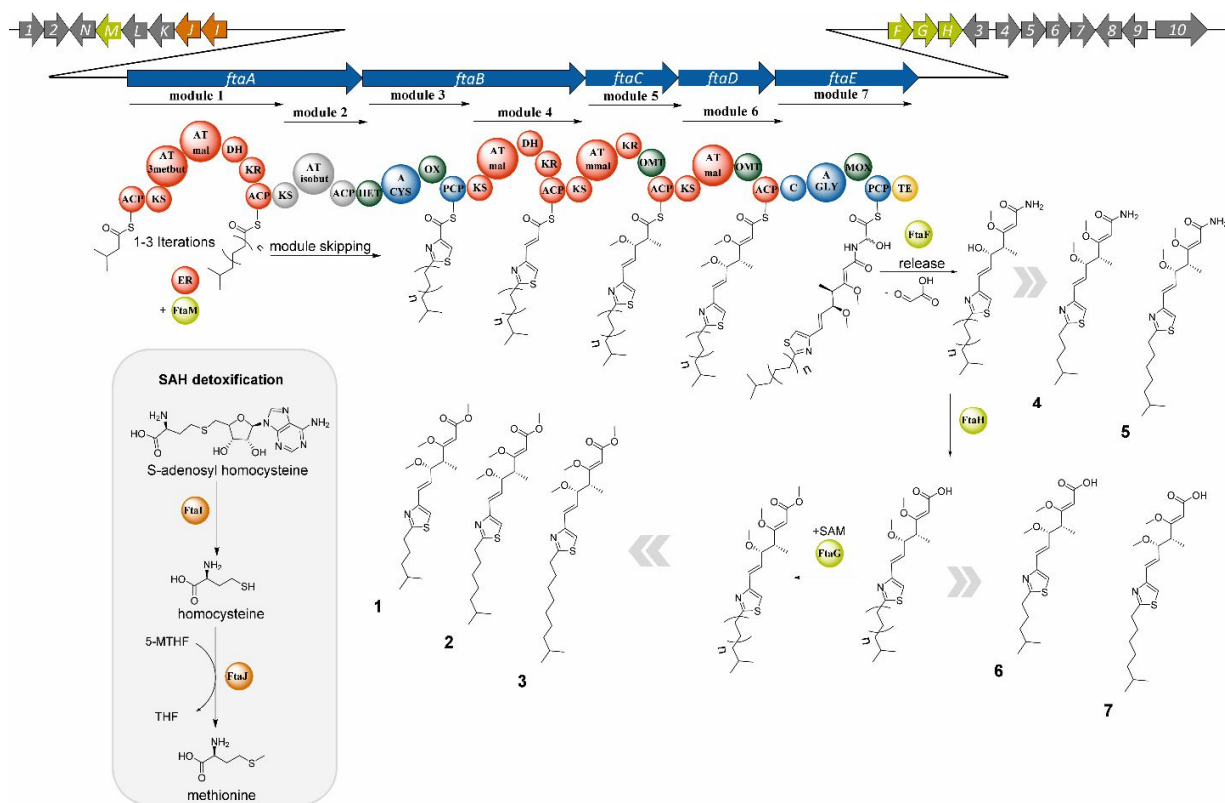
Figure 3. *Gene cluster organization and biosynthesis proposal for the fulvuthiacenes. NRPS and PKS assembly line genes as blue arrows; tailoring enzyme genes as light green arrows, proteins as light green balls; methionine salvage pathway genes as orange arrows, proteins as orange balls; PKS domains (red), NRPS domains (blue), inactive domains (gray), thioesterase (TE)(yellow); ACP acyl carrier protein; AT acyl transferase; A adenylation domain ; C condentation domain; DH dehydratase; ER enoyl reductase; HET heterocyclisation domain; KR ketoreductase; KS ketosynthase; MOX monooxygenase; OMT O-methyltransferase; OX oxidation; PCP peptidyl carrier protein. Fulvuthiacene A-C, **1-3;** prefulvuthicene A and B, **4** and **5**; desmethyl-fulvuthiacene A and B, **6** and **7**.*

Domain predictions by antiSMASH as well as the predicted gene functions of putative tailoring enzymes obtained by blast searches against the NCBI non-redundant protein data base (nr) (Table S 8) were used to deduce a biosynthetic model for fulvuthiacene biosynthesis.[7] Fulvuthiacene biosynthesis exhibits some peculiarities, such as the first module, which is able to incorporate multiple malonyl-CoA extender units by programmed iteration. This reaction sequence forms a branched chain fatty acyl residue of variable length starting with 3-methylbutyryl-CoA. This relaxed selectivity is reflected by the different fulvuthiacene derivatives detected, which either differ in their post assembly line processing state or in the number of iterations performed by the first module. This first module also loads the 3-methylbutyryl-CoA starter unit indicated by the presence of two copies of the acyl transferase domain (AT). HMM predictions according to Minowa et al. suggested the first AT is 3-methylbutyryl-CoA specific, while the second AT had a predicted substrate specificity for malonyl-CoA.[18] Moreover, this iterative PKS type I module does not feature an enoyl reductase domain (ER), whilst producing completely saturated fatty acyl chains. Enoyl reduction is thus likely catalyzed by FtaM as an in *trans* acting enoyl reductase similar to LovC that catalyzes enoyl reduction in the lovastatin biosynthesis.[19] Although primary sequence homology of the proteins LovC and FtaM is low (12 % by MUSCLE alignment), they are both composed of the same domain architecture. This architecture comprises a crotonase fold and a NAD cofactor binding pocket, both of which are crucial for enoyl reduction (section S 2.3). Module 2 is plausibly skipped during the biosynthesis as a large part of the AT domain in this module is missing, rendering it inactive (section S 2.2, Figure S 8). The following heterocycle forming module incorporates cysteine and synthesizes a thiazoline moiety further oxidized by the oxidation domain in module 3, creating the thiazole ring. Further assembly of the compound, chain release by the TE-domain and the first modification steps seem to happen congruently to melithiazol biosynthesis.[20,21] The oxidative glycine degradation, deamination and methyl transfer pathway as depicted in Figure 3 has already been successfully elucidated for the melithiazol biosynthesis and successfully transferred to myxothiazol biosynthesis.[20] FtaF, FtaH and FtaG encode enzymes of the fulvuthiacene pathway showng high similarity to the corresponding biosynthetic enzymes in the melithiazol and myxothiazol clusters and are therefore convincing candidates to produce the methyl ester intermediate shown in the biosynthesis scheme (Figure 7, Figure S 14, S 15 and S 16). A peculiar trait of fulvuthiacene biosynthesis is encoded by genes FtaI and FtaJ: these two genes encode an extra copy of the S-adenosyl homocysteine (SAH) detoxification and methionine-recycling pathway, by specifically encoding an extra copy of the S-adenosylhomocysteinase (FtaI) and an extra copy of the 5-methyltetrahydrofolate cyclohydrolase (FtaJ) (Table S 8). These enzymes appear useful for fulvuthiacene production, since three methyl transfer steps per molecule are required.
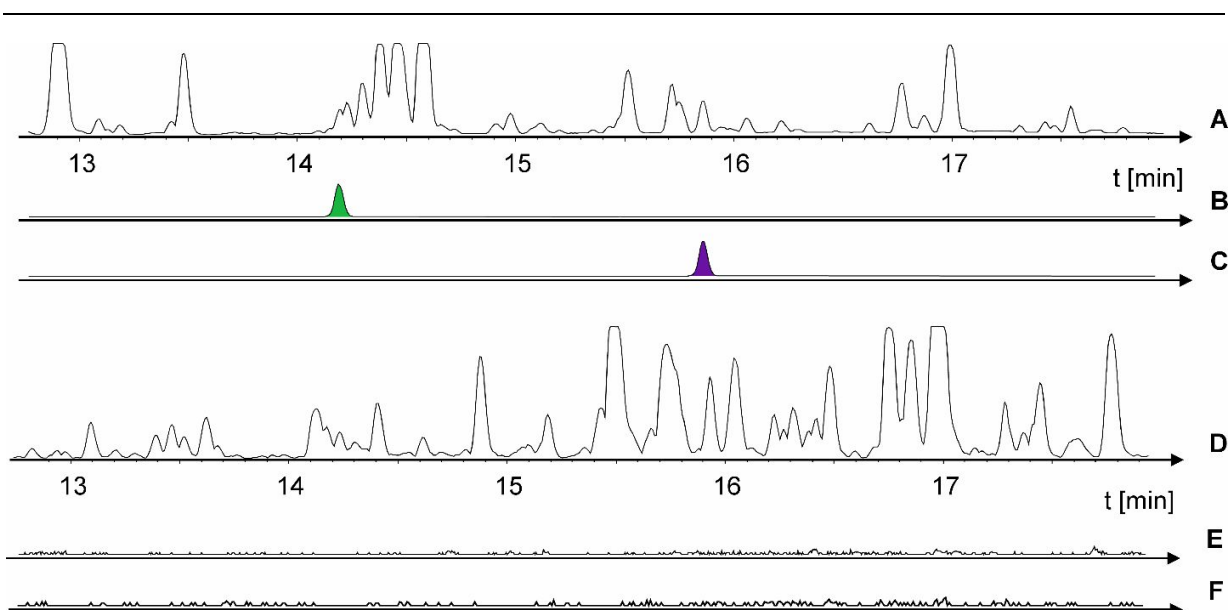
Figure 4. *UHPLC MS extracted ion chromatogram (EIC) traces of fulvuthiacene A (**B,E**; green) and fulvuthiacene B (**C,F**, purple) as [M-MeOH+H]+ ions (350.1783 Da and 378.2096 Da) in the wild type (**A-C**) as well as the* ftaD *single crossover inactivation mutant (**D-E**) crude extracts. Grey traces represent the corresponding base peak chromatograms (BPC). E and F are magnified 100x.*

Consequently, cells needs to ensure SAH recycling, as intracellular SAH accumulation is lethal and cells also need to refill the S-adenosyl methionine (SAM) pool quickly.[22]

In order to ultimately prove the compound to gene cluster assignment, the fulvuthiacene biosynthesis gene cluster was inactivated via single crossover plasmid insertion. Therefore, a pCR2.1 vector carrying a 1kb homology fragment to the *ftaD* gene was integrated into the *ftaD* gene on the MCy9280 chromosome and integration was verified via PCR. The mutant strain was grown in VYGS medium supplemented with 50 µg/ml kanamycin and fulvuthiacene production was monitored in the bacterial crude extract (section S 1.9). As a result of single crossover inactivation of the *ftaD* gene, the production of **1** and **2** was abolished. Additionally, no other fulvuthiacene derivatives could be detected anymore. This finding confirmed the *fta* cluster as the BGC responsible for biosynthesis of all fulvuthiacene derivatives.

3.3.4   Determination of the fulvuthiacenes' absolute configuration

The absolute configuration of **1** and **2** could not be solved using solely NMR spectroscopy. With the fulvuthiacene BGC confirmed, we approached this problem *in silico*. Therefore, an established tool to predict the stereochemistry outcome afforded by KR domains called profileHMM was used to predict fulvuthiacene configuration as produced by module 5.[23] The stereochemistry prediction for the methoxy group as well as the neighboring methyl branch are shown in Figure 8 (section S 2.4).[23] Convincingly, this is the same configuration as found in the corresponding PKS part of the myxothiazol natural products.[24] Since the configuration of myxothiazol was elucidated via an ozonolysis reaction with subsequent

derivatization, we pursued a similar approach for the fulvuthiacenes. In contrast to myxothiazols that are produced in high yields, fulvuthiacene could not be isolated at increased scale. Therefore, we used a small-scale reaction and derivatization cascade that would forcibly create diastereomers out of all four hypothetical stereoisomers (Scheme 1). This allows direct comparison of the stereo centers in fulvuthiacene and myxothiazol via LC-MS, resembling amino acid conformation determination in Marfey's analysis (section S 4.2).[25]

Myxothiazol A (8)   **Fulvuthiacene A (1)**

1) O_3, MeOH, -78 °C to RT
2) H_2O_2, H_2O 1.5h

10 eq. L-Phenylalanin-O-MeEster
HATU, DIPEA
DMF RT ON
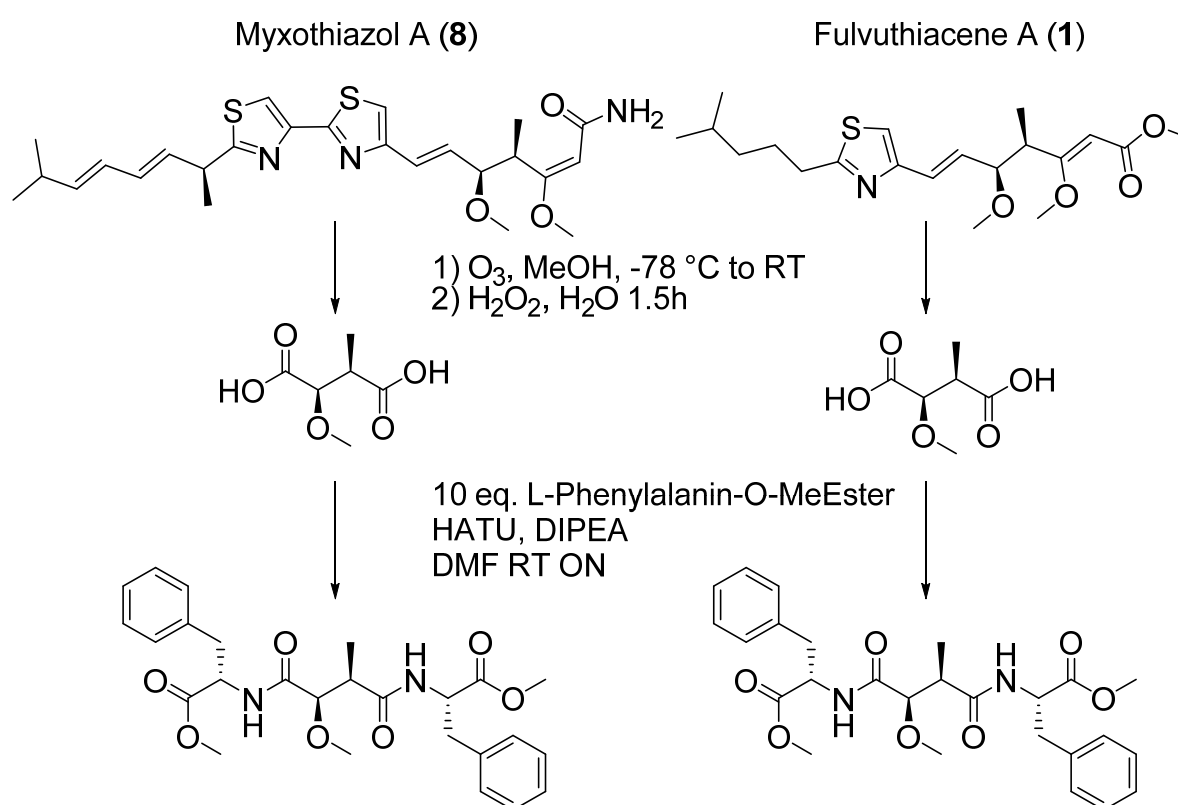
Scheme 1. Reaction and derivatisation scheme for the elucidation of the fulvuthiacene's absolute stereochemistry by LC-MS.

The structural elements excised from myxothiazol and fulvuthiacene by ozonolysis and derivatized as described in Scheme 1 have the same configuration as they elute at the same retention time. Subsequent MS² analysis of given reaction product confirms the reaction product's structure as seen in Figure 5. On a side note, a small peak close to the main peak for the derivatization of the myxothiazols corresponds to a small degree of epimerisation of the ozonolysis product during the HATU mediated peptide coupling.[26] As this peak is separated from the molecule with the correct stereochemistry we confirm that the stereoisomers separate sufficiently, allowing assignment of the fulvuthiacenes' configuration via LC-MS.

Figure 5. *EIC traces of 485.228 [M+H]+ representing the ozonolysis and derivatization reaction product of Myxothiazol A (depicted in blue) and Fulvuthiacen A (depicted in dark red), its structure formula and MS² analysis of the detected ozonolysis product.*

3.3.5 Insights into β-methoxymethacrylate formation

To complete our investigation of fulvuthiacene biosynthesis we sought to shed light onto formation of the terminal β-methoxymethacrylate unit. From previous work on the melithiazol pathway *in vivo* and *in vitro,* it was likely that the fulvuthiacenes' terminal β-methoxymethacrylate moiety would be formed using a route analogous to melithiazol maturation.[20] This is further supported by the detection of the corresponding intermediates, the prefulvuthiacenes and desmethyl-fulvuthiacenes in our SPL-based spectral networking approach (Figure 1, Table 1) representing the corresponding intermediates. In both cases, the assembly line finishes with the addition of a glycine, subsequent alpha oxidation and release of a terminal amide from the hemiaminal structure (Figure 3). [20,21] The resulting amide (see Figure 2) that can putatively be detected as prefulvuthiacene A (**4**) and B (**5**) by mass spectrometry (Table 1) is desaminated and methylated by the amidohydrolase FtaH and one of the SAM dependent methyl transferase FtaG or FtaN. As MCy9280 proved not amenable to genetic manipulation beyond the basic

single-crossover homologous recombination protocol, we chose to work with the myxothiazol producer *S. aurantiaca* DW/3-1 to underpin the reaction sequence that lead to β-methoxymethacrylate biosynthesis.[16] Myxothiazol biosynthesis is analogous to the biosynthesis of melithiazol and fulvuthiacene but it lacks the post PKS modification steps transforming the compounds' terminal amide into a methoxy group (Figure 7).[16] By transfer of the maturation enzymes FtaG and FtaH into a myxovthiazol producer we should be able to transform the amide moiety in myxothiazol A into a methyl ester. Additionally, with respect to the biosynthesis model in Figure 3, we lack definite proof for *in trans* enoyl reduction in module 1, although there is no ER domain on the *ftaA* protein (Figure 3). The best candidate for this enzyme is FtaM, a protein with a crotonase fold and an NAD binding pocket, showing homology to the FadJ subunit as determined by NCBI blast and conserved domain search.[27] We therefore created a plasmid that comprises FtaG and FtaH under a tn5 promotor as well as FtaM and FtaN under a vanillate promotor and repressor system (section S 1.15) to explore the influence of FtaG and FtaH as well as FtaM and FtaN on myxothiazol biosynthesis. After transformation of this plasmid into *S. aurantiac*a DW4/3-1 we observed changes in the myxothiazol production as shown in Figure 6 (section S 1.9).
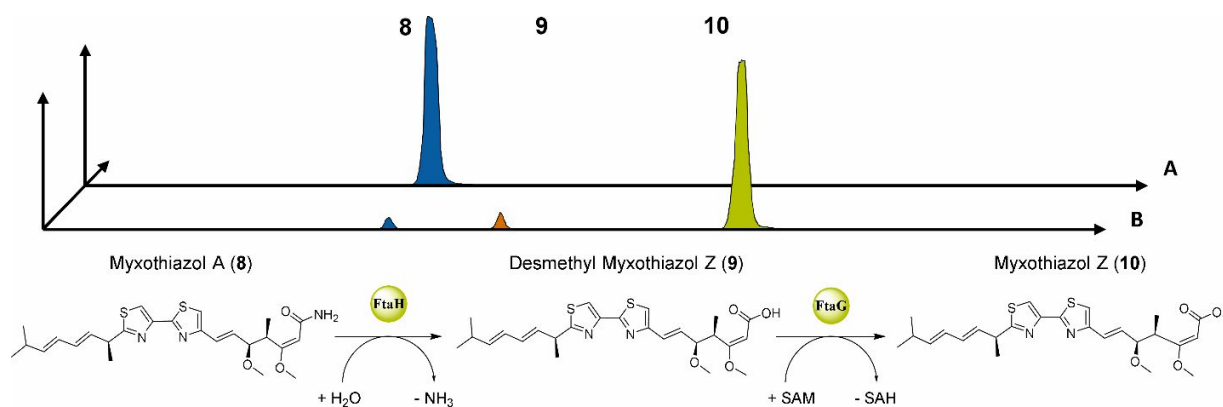
Figure 6. Production scheme for the secondary metabolites produced by S. aurantiaca DW4/3-1 WT (A) and DW4/3-1 attB :: attP pFPVan FtaMN tn5 GH (B). Myxothiazol A (8) is depicted in blue, myxothiazol Z (10) [20] is depicted in green and desmethyl-myxothiazol Z (9) [17] is depicted in orange.

The transformation of **8** to **10** with a 'hybrid' myxothiazol cluster does not only work in the previously reported melithiazol case but also with the respective fulvuthiacene pathway enzymes named FtaG and FtaH.[20] The induction of the enzymes FtaM and FtaN putatively involved in the biosynthesis do not afford further changes in the chromatogram, meaning that the transformation of **8** to **10** is still almost complete while no further processing of **10** takes place. We can therefore conclude that FtaG and not FtaN is responsible for terminal methyl ester formation. Also, this reaction scheme shows that the desmethyl-fulvuthiacenes as well as the prefulvuthiacenes likely possess the structures assigned in the biosynthesis diagram as they are intermediates produced along the fulvuthiacene assembly process (Table 1). Since

NMR data for **9** clearly show that this compound represents the free acid (section S 4.1.2 and 4.1.3) it is very likely that desmethyl-fulvuthiacene A (**6**) and B (**7**) have the same structure, although this couldn't be confirmed by NMR due to the extremely low yields. Finally, heterologous transfer of the *ftaM* gene from the fulvuthiacene pathway to the myxothiazol pathway does not lead to the production of myxothiazols featuring a saturated sidechain. As module 1 on the *ftaA* gene differs quite significantly from *mtaB* module 1 (Figure 7, section S 2.2) FtaM is likely unable to dock correctly to *mtaB* module 1, therefore preventing the formation of myxothiazols lacking the double bonds in their fatty acyl sidechain. To further investigate this enoyl reduction reaction one would need a heterologous transfer of the enire fulvuthiacene cluster into a heterologous host that has a better accessibility for genetic manipulation than MCy9280, which would allow gene deletion experiments.

3.3.6   Comparison between fulvuthiacene, myxothiazol and melithiazol BGCs

As already pointed out, there is a striking degree of similarity between the BGCs for myxothiazol, melithiazol and fulvuthiacene, which is also reflected by similar biosynthesis and maturation steps for these compounds. Protein sequences for corresponding functional units in the respective BGCs as apparent in Figure 7 were aligned to obtain insights into the divergent evolution process that created these different clusters. Firstly, the last three assembly line proteins FtaC, D and E in the fulvuthiacene case, MelE, F and G in the melithiazol case and MtaE, F and G in the myxothiazol case share high similarity even though the proteins stem from different myxobacterial species (Figure 7, Figure S 7 to S 9). One of the major differences in the fulvuthiacene biosynthesis compared to the myxothiazol and melithiazol biosynthesis is encoded in the proteins FtaB and MelC/MelD or MtaC/MtaD respectively. While both the myxothiazol and the melithiazol assembly line contain two heterocycle forming modules, fulvuthiacene contains only one of these. Additionally, the single remaining thiazole forming module encoded on *ftaB* resembles *melC/mtaC* upstream of the A domain and *melD* and *mtaD* downstream of the A domain (Figure S 8 and S 9). It is therefore likely that the fulvuthiacene cluster lost one of the thiazole forming units by homologous recombination of a *melC/mtaC* like gene with a *melD/mtaD* like protein at the site of the A domain. Another interesting part of the fulvuthiacene biosynthesis is module 2 on the *ftaA* gene. This module resembles module 1 from the melithiazol BGC with the only exception that parts of the AT domain in *ftaA* module 2 are deleted. This explains the inactivity of module 2 as it does not accomplish incorporation of isobutyl-CoA as an alternative starter unit, unlike predicted by hidden markov models (HMM). In contrast, the first module of *ftaA* shows high similarity to the first module on *mtaB* both of which are combined loading and elongation modules. This is peculiar as both modules produce rather

different polyketide sidechains. Not only does the *ftaA* module 1 act iteratively while *mtaB* module 1 catalyzes only one chain elongation step, but *ftaA* module 1 also forms a completely saturated fatty acid chain despite lacking an ER domain. This is most likely the effect of a protein encoded on *ftaM* that possesses a crotonase fold and a NAD binding pocket and has no homologs in the myxothiazol cluster. One peculiarity of all three clusters is the protein named FtaF, MelH and MtaH, respectively. Although it has been shown that this fumarylacetoacetate hydrolase like protein (Figure S 14) is not necessary for myxothiazol biosynthesis it seems likely that this hydrolase type protein facilitates removal of the product and the oxaloacetate that is released (Figure 3).[28] Finally, as proven by heterologous expression in this work, FtaH and FtaG form the methyl ester analogous to the methyl ester formation in melithiazol catalyzed by the proteins MelJ and MelK.
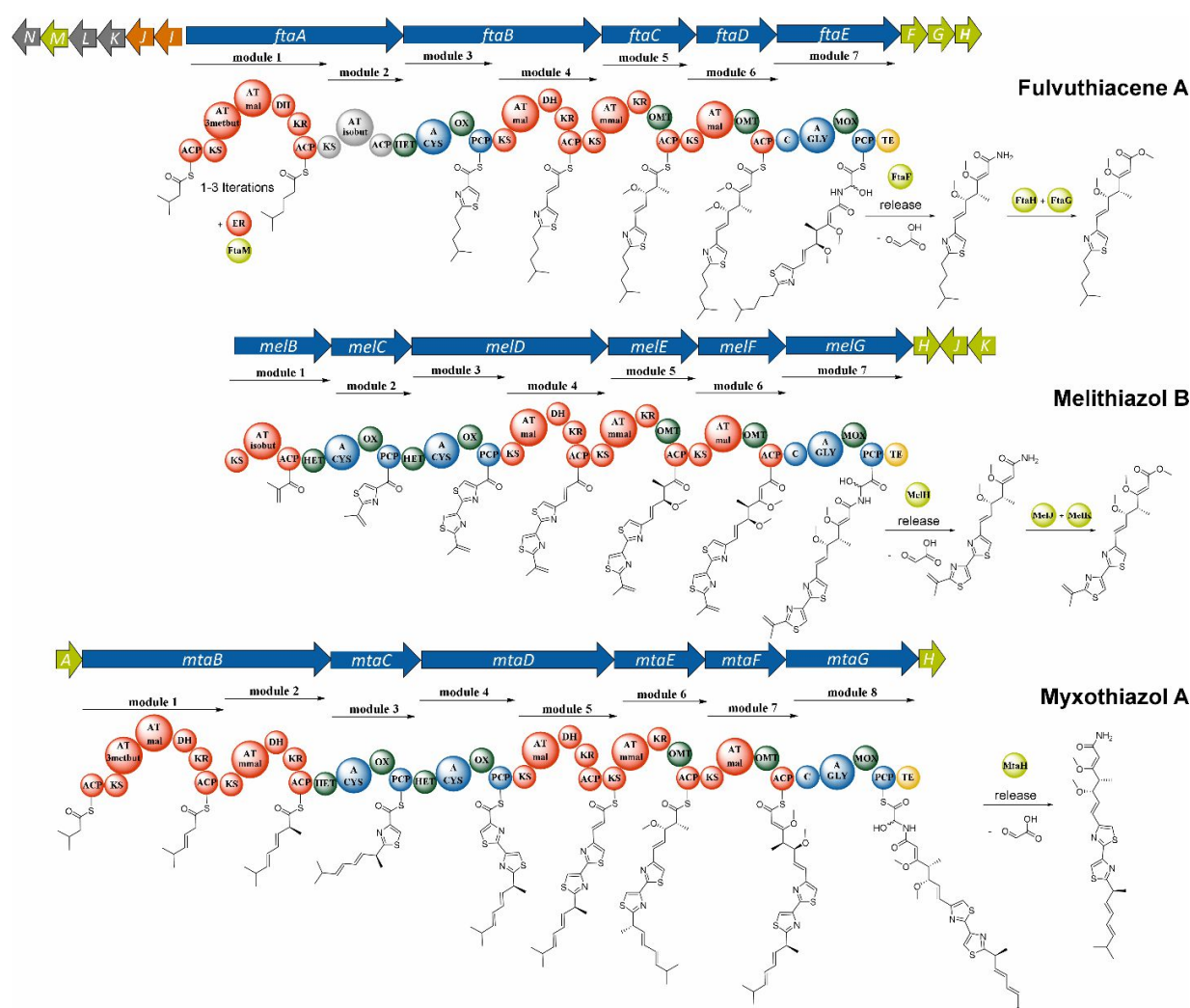
Figure 7. *Comparative analysis of the myxothiazol, melithiazol and fulvuthiacene BGCs and the respective biosynthesis steps* [16,17]

3.3.7 Structure-acitivity relationships within the β-methoxyacrylate class of respiratory chain inhibitors

The fulvuthiacene scaffold contains a β-methoxyacrylate motif, a pharmacophore seen in many fungicides targeting the bc-1 subunit of the respiratory chain.[29] As an example, strobilurine has long been used successfully in agriculture as an antifungal agent for crop protection. Interestingly, the fulvuthiacenes do not display any apparent fungicidal activity or cytotoxicity against *M. hiemalis*, *C. albicans* or CHO and HCT116 cells (up to 64 µg/ml), which is remarkable considering their structural relatedness to the strongly antifungal and cytotoxic myxothiazols and melithiazols.[29] While the β-methoxyacrylate has long been described as the pharmacophore for respiratory chain inhibition at the bc-1 subunit, the biological inactivity of the fulvuthiacenes shows that there are missing parts in this rather simplistic structure-activity relationship (SAR).[29] While extending the β-methoxymethacrylate core motif in C-2 position - as in the strobilurins, oudemansins and cyrmenins - tolerates a wide variety of residues without significant activity losses, the SAR for C-3 extended congeners is more narrow (Figure 8 A).[29,30]
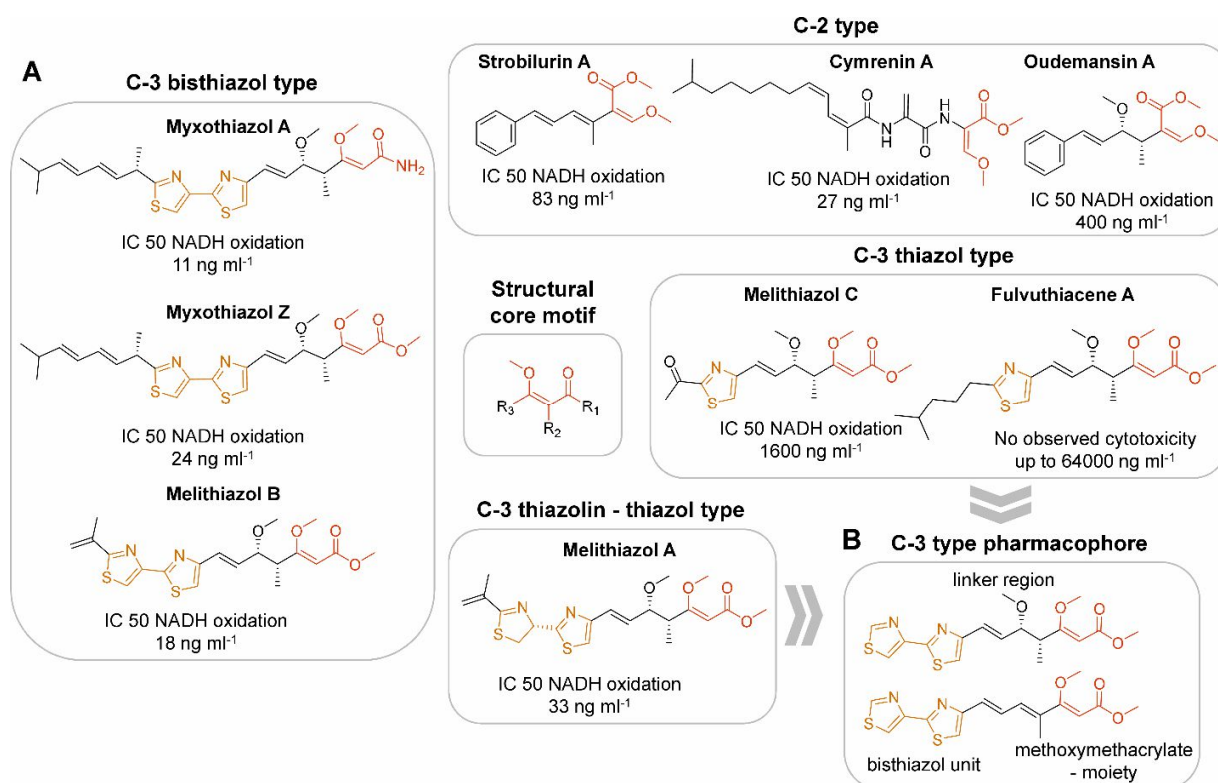
Figure 8. A) *Naturally occurring methoxy-methacrylate type respiratory chain inhibitor classes displayed revealing the main features of the β-methoxymethacrylate structure activity relationship (SAR)[12,31] and* ***B)*** *Pharmacophore of myxothiazol type C-3 thiazol type respiratory chain inhibitors and the putative pharmacophore of myxothiazol type C-3 thiazol type respiratory chain inhibitors considering the oudemansin – strobilurine relationship*

The β-methoxyacrylamide type bisthiazol compound myxothiazol A and the β-methoxymethacrylate type bisthiazol compounds myxothiazol Z and melithiazol B show very strong activity towards the bc-1 subunit, and the thiazol-thiazolin melithiazol B analogue melithiazol A inhibits the respiratory chain at the same order of magnitude. In contrast, the single thiazol derivative melithiazol C inhibits the respiratory chain only at a concentration two orders of magnitude higher and fulvuthiacene A and B are non-toxic to human cell lines as well as a test panel of fungi and yeast up to 64 µg/ml, indicating the fulvuthiacenes not to inhibit the bc-1 subunit at all (Figure 8). Therefore, to produce a strong bc-1 inhibitor with a C-3 extended β-methoxyacrylate warhead, a bisthiazol or thiazolin-thiazol type structural element is needed. Figure 8 A illustrates that only the C-3 connected compounds harboring this bicyclic structural element are able to efficiently inhibit electron transfer at the bc-1 subunit. Since the C-3 bisthiazol type compounds like myxothiazol and melithiazol are comparably stronger antifungal agents, outperforming the strobilurins and oudemansins as fungicides, it is worth revising the SAR to include the bisthiazol unit in order to develop efficient antifungal crop protection substances (Figure 8 B). More importantly, considering the structural differences between the oudemansins and strobilurins, there is a strong possibility that the 5-methoxy moiety is not needed for activity at the cytochrome bc-1 subunit indicating that a structure as depicted in Figure 8 B might already be sufficient to exhibit potent antifungal activity. From a synthetic point of view this structure would be more readily accessible as it has no stereocenters, significantly facilitating its chemical synthesis. As development of novel β-methoxymethacrylate antifungal agents almost solely focused on the strobilurine core structure so far, these results might pave the way for the development of a novel fungicide series.[32]

3.4   Conclusion

In this work, we pursued a comprehensive MS-based strategy for the discovery of natural products from myxobacteria. We combined statistics-based filtering of untargeted high-resolution MS data for improved coverage in subsequent targeted tandem-MS measurements with spectral networking for the early evaluation of structural similarity, leading to full structure elucidation of a new secondary metabolite scaffold from *Myxococcus fulvus*. Albeit the biological activity of newly identified fulvuthiacenes remains currently elusive, investigation of the underlying biosynthetic gene cluster provided insights into enzymatic particularities and shed light on the evolutionary relationship to myxothiazol and melithiazol biosynthesis. Notably, our finding of fulvuthiacenes extended a diverse family of natural products bearing the β-methoxymethacrylate motif, which has been considered crucial for inhibition of the bc-1 subunit of the respiratory chain. However, fulvuthiacenes did not show cytotoxicity or fungicidal activity and thus

appear unable to effectively inhibit the respiratory chain complexes bc-1 subunit in the way myxothiazol does.[29] This study therefore adds a new detail to the overall structure activity relationship scheme of the β-methoxymethacrylates that are commercially used as fungicides. Our results may promote the development of innovative β-methoxymethacrylate fungicides with higher cytotoxicity compared to commercially used derivatives of strobilurins, allowing the desirable reduction of material deployed for crop protection.

3.5   Acknowledgement

3.6   References

[1]   a) D. Krug, R. Müller, *Nat. Prod. Rep.* **2014**, *31*, 768; b) D. Krug, G. Zurek, B. Schneider, R. Garcia, R. Müller, *Anal. Chim. Acta* **2008**, *624*, 97.

[2]   T. Hoffmann, D. Krug, S. Hüttel, R. Müller, *Anal. Chem.* **2014**, *86*, 10780.

[3]   M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan et al., *Nat. Biotechnol.* **2016**, *34*, 828.

[4]   C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, S. Neumann, *J. cheminformatics* **2016**, *8*, 3.

[5]   M. Maansson, N. G. Vynne, A. Klitgaard, J. L. Nybo, J. Melchiorsen, D. D. Nguyen, L. M. Sanchez, N. Ziemert, P. C. Dorrestein, M. R. Andersen et al., *mSystems* **2016**, *1*, e00028-15.

[6]   a) R. D. Kersten, Y. L. Yang, Y. Xu, P. Cimermancic, S. J. Nam, W. Fenical, M. A. Fischbach, B. S. Moore, P. C. Dorrestein, *Nat. Chem. Biol.* **2011**, *7*, 794; b) M. H. Medema, Y. Paalvast, D. D. Nguyen, A. Melnik, P. C. Dorrestein, E. Takano, R. Breitling, *PLoS Comput. Biol.* **2014**, *10*, e1003822; c) D. D. Nguyen, A. V. Melnik, N. Koyama, X. Lu, M. Schorn, J. Fang, K. Aguinaldo, T. L. Lincecum, M. G. K. Ghequire, V. J. Carrion et al., *Nat. Microbiol.* **2016**, *2*, 16197.

[7]   T. Weber, K. Blin, S. Duddela, D. Krug, H. U. Kim, R. Bruccoleri, S. Y. Lee, M. A. Fischbach, R. Müller, W. Wohlleben et al., *Nucleic Acids Res.* **2015**, *43*, W237-W243.

[8]   S. C. Wenzel, R. Müller, *Mol. Biosyst.* **2009**, *5*, 567.

[9]   K. Blin, H. U. Kim, M. H. Medema, T. Weber, *Brief. Bioinform.* **2017**, 1.

[10] a) R. Tautenhahn, G. J. Patti, D. Rinehart, G. Siuzdak, *Anal. Chem.* **2012**, *84*, 5035; b) S. Chanana, C. S. Thomas, D. R. Braun, Y. Hou, T. P. Wyche, T. S. Bugni, *Metabolites* **2017**, *7*, 34.

[11] a) Y. J. Kim, K. Furihata, S. Yamanaka, R. Fudo, H. Seto, *J. Antibiot.* **1991**, *44*, 553; b) R. Jansen, H. Steinmetz, F. Sasse, W.-D. Schubert, G. Hagelueken, S. C. Albrecht, R. Müller, *Tetrahedron Lett.* **2008**, *49*, 5796; c) W. Trowitzsch-Kienast, E. Forche, V. Wray, H. Reichenbach, E. Jurkiewicz, G. Hunsmann, G. Höfle, *Liebigs Ann. Chem.* **1992**, *16*, 659.

[12] F. Sasse, B. Böhlendorf, M. Herrmann, B. Kunze, E. Forche, H. Steinmetz, G. Höfle, H. Reichenbach, M. Hermann, *J. Antibiot.* **1999**, *52*, 721.

[13] S. Schieferdecker, T. E. Exner, H. Gross, M. Roth, M. Nett, *J. Antibiot.* **2014**, *67*, 519.

[14] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12580.

[15] N. Garg, C. Kapono, Y. W. Lim, N. Koyama, M. J. A. Vermeij, D. Conrad, F. Rohwer, P. C. Dorrestein, *Int. J. Mass Spectrom.* **2015**, *377*, 719-717.

[16] B. Silakowski, H. U. Schairer, H. Ehret, B. Kunze, S. Weinig, G. Nordsiek, P. Brandt, H. Blöcker, G. Höfle, S. Beyer et al., *J. Biol. Chem.* **1999**, *274*, 37391.

[17] S. Weinig, H.-J. Hecht, T. Mahmud, R. Müller, *Chem. Biol.* **2003**, *10*, 939.

[18] Y. Minowa, M. Araki, M. Kanehisa, *J. Mol. Biol.* **2007**, *368*, 1500.

[19] J. Kennedy, K. Auclair, S. G. Kendrew, C. Park, J. C. Vederas, C. R. Hutchinson, *Science* **1999**, *284*, 1368.

[20] I. Müller, S. Weinig, H. Steinmetz, B. Kunze, S. Veluthoor, T. Mahmud, R. Müller, *ChemBioChem* **2006**, *7*, 1197.

[21] I. Müller, R. Müller, *FEBS J.* **2006**, *273*, 3768.

[22] N. Parveen, K. A. Cornell, *Mol. Microbiol.* **2011**, *79*, 7.

[23] A. Kitsche, M. Kalesse, *ChemBioChem* **2013**, *14*, 851.

[24] W. Trowitzsch, G. Höfle, W. S. Sheldrick, *Tetrahedron Lett.* **1981**, *22*, 3829.

[25] R. Bhushan, H. Bruckner, *Amino Acids* **2004**, *27*, 231.

[26] L. A. Carpino, H. Imazumi, A. El-Faham, F. J. Ferrer, C. Zhang, Y. Lee, B. M. Foxman, P. Henklein, C. Hanay, C. Mügge et al., *Angew. Chem. Int. Ed.* **2002**, *41*, 441.

[27] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales et al., *Nucleic Acids Res.* **2017**, *45*, D200-D203.

[28] S. Weinig, T. Mahmud, R. Müller, *Chem. Biol.* **2003**, *10*, 953.

[29] W. F. Becker, G. von Jagow, T. Anke, W. Steglich, *FEBS Lett.* **1981**, *132*, 329.

[30] T. Leibold, F. Sasse, H. Reichenbach, G. Höfle, *Eur. J. Org. Chem.* **2004**, 431.

[31] F. Sasse, T. Leibold, B. Kunze, G. Höfle, H. Reichenbach, *J. Antibiot.* **2003**, *56*, 827.

[32] D. W. Bartlett, J. M. Clough, J. R. Godwin, A. A. Hall, M. Hamer, B. Parr-Dobrzanski, *Pest Manag. Sci.* **2002**, *58*, 649.

Supporting Information

Novel β-methoxymethacrylate natural products uncovered by statistics-based mining of the *Myxococcus fulvus* secondary metabolome

Fabian Panter[a,b], Daniel Krug[a,b] and Rolf Müller*[a,b]

Affiliation

[a]Department Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmaceutical Biotechnology, Saarland University, Campus E8.1, 66123 Saarbrücken, Germany

[b] German Centre for Infection Research (DZIF), Partner Site Hannover–Braunschweig, Hannover–Braunschweig, Germany

The printed version of the Supporting Information does not contain data that cannot be visualized satisfactorily on paper. To access this data please refer to the enclosed storage medium or the on-line version of this research paper.

S 3.1   Molecular biology protocols

S 3.1.1   Isolation of genomic DNA from myxobacteria

To isolate total DNA for sequencing purposes such as illumina sequencing, phenol-chloroform gDNA extraction is used.

S 3.1.2   Gene cluster identification and assembly

The MCy9280 genome was sequenced using illumina sequencing at GMAK at the Helmholtz institute for infectious diseases in Brunswick. Three putative fulvuthiacene gene clusters were identified among the BGCs predicted by antiSMASH.[1] The identified fulvuthiacene cluster was assembled via PCR amplification of the gap sequences followed by cloning into the pJet 1.3 vector utilizing clone Jet kit (Thermo scientific, U.S.A.).

S 3.1.3   List of all primers used in this study

Table S 1. *List of all primer sequences used in this study*

Primer name	Primer sequence
Fulvuthiacene cluster inactivation and inactivation mutant verification	
FtaD_KO_fwd	GAATCTGGTCCTGGTTCCCG
FtaD_KO_rev	CTCCCAACCCCTCCATCAAC
FP pCR2.1 in	CCTCTAGATGCATGCTCGAGC
FP pCR2.1 out	GGATCCACTAGTAACGGCCG
FtaD_KO_fwd_verif	TCTCGCTCACCCAATCTCTGTCAGGTGA
FtaD_KO_rev_verif	ACATCACCGCCACCATCCTCCCCACCT
FP_Kan_fwd	GATGGATTGCACGCAGGTTC
FP_Kan_rev	TGATATTCGGCAAGCAGGCA
Gap closing PCR primer used to complete Fulvuthiacenene gene cluster sequence	
FulvuthiacenGap2_rev	GGTGGACATGTTGGAGACCGTCTCC
FulvuthiacenGap2_fwd	AATCTCTGTCAGGTGAGGGCTCAAATG
FulvuthiacenGap1fwd	AGGAGCCTGGAGGAGCTGAAGGAAGAGTT
FulvuthiacenGap1rev	ACTCCCCAGCATCTCCTCCCAATCACTCCT
Primers used for creation of the heterologous gene transfer plasmid pFPVan FtaMN th5 GH	
Mx8_HindIII_fwd	AGTTCTAGACGGCCGCGCTTGTCGAGCG
Mx8_HindIII_rev	ACTCTAGATTCAGAAGAACTCGTCAAGAAGGCGATAGAAG
FtaMN_fwd	AATCATATGAGCGAGCAGAACACCATC
FtaMN_rev	GAAGAATTCAGGGCTTCTCCGCGTGGACG
FtaGH_fwd	TGGGAATTCTATGAGCAAGGAC
FtaGH_rev	GCAAGCTTGGACACTCACCGCCGCGTGAC
FtaG_tn5_rev	CCACATCGCCACACCCAGCGCCGTCCTATCCGCTTGCCGCTTGTCCTTGCTCATATGCTGTACCTCCTTATCCTGTCTC
FtaN_CmR_fwd	AGTGGCTGACGCTGTTCAACTGCGACGAGTACGCCATCGTCCACGCGGAGAAGCCCTGACTGAACAGGAGGGACAGCTG

tn5_CmR_fwd	GATGAGTGGCAGGGCGGGGCGTAATTGTAATACGACTCACTATAGGG
tn5_CmR_rev	CCCTATAGTGAGTCGTATTACAATTACGCCCCGCCCTGCCACTCATC

S 3.1.4   Polymerase chain reaction protocols

S 3.1.4.1   Thermo scientific Taq Polymerase

Taq Polymerase Protocol:

2µL NH_4SO_4 Buffer

(alternatively 2µL KCl Buffer)

2µL $MgCl_2$ solution (25mM)

4µL dNTP's (2mM)

4µL 50% Glycerol

1µL Primer fwd and rev (50mM)

.2 µL Taq DNA Polymerase

1µL gDNA Template (~50 ng/µL)

4.3µL ddH_2O

Taq Polymerase cycler program:

Table S 2. *Taq DNA Polymerase PCR program*

Step	Time [min:s]	Temperature [°C]
Initial denaturation	4:00	95
	0:30	95
Cycle, repeat 30x	0:15	63
	1:20	72
Final elongation	10	72
Store	forever	8

S 3.1.5   Thermo scientific Phusion Polymerase

Phusion Polymerase Protocol:

5µL GC Buffer

2.5µL dNTP's (2mM)

1.25µL DMSO

0.5µL Primer fwd and rev (100mM)

0.2 µL Phusion DNA Polymerase

0.5µL gDNA Template (~50 ng/µL)

14.55µL ddH₂O

Phusion Polymerase cycler program:

Table S 3. Phusion DNA Polymerase PCR program

Step	Time [min:s]	Temperature [°C]
Initial denaturation	4:00	95
	0:30	98
Cycle, repeat 30x	0:15	63
	1:20	72
Final elongation	10	72
Store	forever	8

S 3.1.6 Transformation protocols

S 3.1.6.1 Transformation protocol *E. coli DH10ß* TOPO TA kit cloning

1) Take competent *E. coli DH10ß* from -80°C on ice
2) Mix competent cells with pCR2.1 TOPO reaction mixture
   (0.5µL salt solution, 0.5µL pCR2.1 TOPO vector (linearized), 1-5 µL PCR product cleaned by gel electrophoresis incubated for 10 minutes)

3) Electroporate the cells at 1250 V, 25 µF, 200 Ohm and 1mm in a standard electroporation cuvette (100 µL).

4) Add 500 µL of fresh LB medium

5) Incubate on a shaker at 37°C for 1h

6) Plate on LB Kan50 Agar and incubate the plates overnight at 37°C

7) Pick around 6 clones and cultivate them for plasmid preparation via alkaline lysis

S 3.1.6.2 Transformation protocol *E. coli DH10ß* Thermo Scientific pJet 1.2 Kit

3) Take competent *E. coli DH10ß* from -80°C on ice
4) Mix competent cells with pJet 1.2 reaction mixture
   (0.5µL salt solution, 0.5µL pJet 1.2 vector (linearized), 1-5 µL PCR product cleaned by gel electrophoresis, 1 µL T4 DNA Ligase, ddH₂O to 10µL)

3) Electroporate the cells at 1250 V, 25 µF, 200 Ohm and 1mm in a standard electroporation cuvette (100 µL).

4) Add 500 µL of fresh LB medium

5) Incubate on a shaker at 37°C for 1h

6) Plate on LB Amp100 Agar and incubate the plates overnight at 37°C

7) Pick around 6 clones and cultivate them for plasmid peparation via alkaline lysis

S 3.1.7   Ligation protocol used in this study

Ligations in this study are performed by Thermo scientific T4 ligase.

T4 DNA Ligase mix:

2µL vector DNA (carefully dephosphorylated using FastAP (Thermo scientific)

2 – 6.5 µL insert DNA (triple the molar amount of the vector DNA)

1 µL T4 DNA ligase buffer

0.5 µL T4 DNA ligase

to 10µL ddH$_2$O

The ligase mix is incubated at 16°C for 15-24h and 1-5 µL ligation mix are subsequently electroporated into *E. coli* DH10ß using standard protocol.

S 3.1.8   Plasmid verification

S 3.1.8.1   Plasmid preparation via alkaline lysis

For verification of created constructs, plasmids are prepared via standard alkaline lysis.

S 3.1.8.2   Verification of the pCR2.1 and pJet 1.2 constructs

1) Find a restriction Enzyme that cuts one time on the vector backbone and one time on the insert.

2) Mix 8.5µL Plasmid from alkaline lysis with 1µL restriction enzyme buffer and 0.5 µL enzyme.

3) Incubate for 3-4 h

4) Verify the bands on agarose gel (0.8 %) by gel electrophoresis

S 3.1.9   Transformation of *Myxococcus fulvus* MCy9280

The transformation protocol for *M. fulvus* MCy9280 was elaborated from CTT medium cultures, since CTT medium leads to growth in perfect suspension. This is a prerequisite for efficient electro transformation of myxobacteria.

S 3.1.9.1   Transformation protocol

1) Centrifuge 2ml of an MCy9280 culture in CTT Medium at OD_{600} of approx. 0.3 at 8000 rpm for 2 minutes at a table centrifuge.

2) Wash the residual cell pellet 2 times with 500 µL autoclaved ddH_2O and discard the supernatant.

3) Resuspend cells in 50µL of ddH_2O, add 5 µL of plasmid solution (prepared from E. coli with Thermo Scientific Miniprep Kit) at a conc. of 0.3-0.4 ng/µL and transfer the solution into an electroporation cuvette.

4) Electroporation at 675 V, 400 Ohms, 25µF and 1mm Cuvette Settings for optimum electroporation efficiency.

5) Flush the cells out with 1 mL fresh CTT medium, transfer the Cell suspension into a 2ml Eppendorf tube.

6) Incubate the Cells for 5h on a Shaker thermostated to 37°C at 300 rpm.

7) Plating the cells on Kan50 CTT Agar after mixing the cell suspension with 3 ml of Kan50 CTT Softagar.

8) Clones appear after 5-7 Days in the 30°C incubator.

S 3.1.10   Transformation of *Stigmatella aurantiaca* DW4/3-1

The transformation protocol for *S. auramntiaca DW4/3-1* was elaborated from TS-6 medium cultures, since TS-6 medium leads to growth in perfect suspension necessary for efficient electrotransformation.

S 3.1.10.1   Transformation protocol

9) Centrifuge 2ml of a *S. aurantiaca* DW4/3-1 culture in TS-6 Medium at OD_{600} of approx. 0.3 at 8000 rpm for 2 minutes at a table centrifuge.

10) Wash the residual cell pellet 2 times with 500 µL autoclaved ddH_2O and discard the supernatant.

11) Resuspend cells in 50µL of ddH_2O, add 5 µL of plasmid solution (prepared from E. coli with Thermo Scientific Miniprep Kit) at a conc. of 0.3-0.4 ng/µL and transfer the solution into an electroporation cuvette.

12) Electroporation at 650 V, 400 Ohms, 25µF and 1mm cuvette Settings for optimum electroporation efficiency.

13) Flush the cells out with 1 mL fresh TS-6 medium, transfer the cell suspension into a 2ml Eppendorf tube.

14) Incubate the Cells for 5h on a Shaker thermostated to 37°C at 300 rpm.

15) Plating the cells on Kan50 TS-6 Agar after mixing the cell suspension with 3 ml of Kan50 TS-6 Softagar.

16) Clones appear after 10-14 Days in the 30°C incubator.

S 3.1.11   Vector creation for Fulvuthiacene cluster inactivation

S 3.1.11.1   pCR2.1 FtaD homology

For single crossover inactivation of the suspected fulvuthiacene secondary metabolite cluster a pCR 2.1 vector from the TOPO-TA cloning kit (Thermo scientific TOPO-TA cloning Kit) was equipped with a 1 kb homology sequence amplified from the FtaD gene in order to obtain single crossover inactivation mutants of the Fulvuthiacene biosynthesis gene cluster (main text Figure 3). The homology sequence is amplified with the primers FtaD_KO_fwd and FtaD_KO_rev (Table S 1).
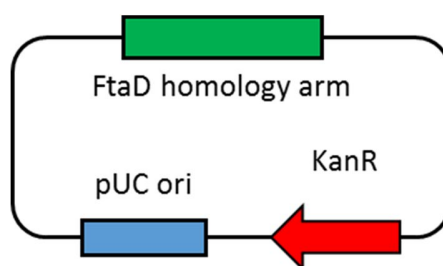
Figure S 1. *Schematic view of pCR2.1 FtaD homology vector used for single crossover inactivation of the fulvuthiacene biosynthetic gene cluster*

S 3.1.12   Verification of the Fulvuthiacenes biosynthetic origin via single crossover inactivation of the corresponding gene cluster

In order to verify the predicted gene cluster to secondary metabolite relationship, the assembled putative fulvuthiacene gene cluster was inactivated via a single crossover insertion inactivation of the FtaD gene by a pCR2.1 vector containing 1kb homology to the said gene. An electroporation protocol for *Myxococcus fulvus* MCy9280 that yields ca. 30 clones per electroporation was elaborated (section S 1.9). Verification of the single crossover mutant was done by PCR with *Myxococcus fulvus* MCy9280 mutant gDNA prepared using the Quiagen gDNA isolation kit. The frequency of homologous recombination events at the correct position of the MCy9280 seems rather low, as less than 1 per 10 screened clones displaying kanamycin resistance contain the plasmid integrated into the correct location on their bacterial chromosome. Therefore, further work on the fulvuthiacene biosynthesis elucidation was shifted towards

S. aurantiaca DW4/3-1 as this strains displays a much better reliability in genetic manipulation experiments.

S 3.1.13 Protocols for λ-Red prophage mediated recombination (Red/ET)

The strain used for all λ-Red based recombination experiments was *E. coli HS996* carrying the thermosensitive pSC101 BAD red gba plasmid. As the origin of this plasmid stops replicating at 37°C the strain has to be grown at 30°C for cryo cultures.

S 3.1.13.1 Incorporation of plasmids into the λ-Red recombination (Red/ET) strain

1) Inoculate 25 ml LB containing 12 µg/ml Oxytetracycline and grow the strain overnight at 30°C 200 rpm on a rotary shaker

2) Inoculate 20 µL of the well grown *E. coli* overnight culture into a 2 ml Eppendorf tube with a punctured lid containing 1.5 ml LB with 12 µg/ml Oxytetracycline

3) Incubate the culture for 4h until OD reaches approx. 0.6

4) Mix target vector or Ligation mixture (ca.1-3µL) with competent *E. coli* and pipet the mixture into a Electroporation cuvette (100µL)

5) Electroporate the mixture at 1250V, 25µF, 200 Ohm, 1mm Cuvette

6) Add 1mL fresh sterile LB medium, incubate for 1h at 900 rpm 30°C in a 2ml Eppendorf tube with a punctured lid

7) Plate on LB Kan50 Otc12 Agar (provided the plasmid has kanamycin resistance, if not use appropriate antibiotic) and incubate the plates overnight at 30°C

8) Pick around 6 clones and cultivate them for plasmid preparation via alkaline lysis

9) Confirm the electroporated plasmid via restriction digestion

S 3.1.13.2 Modification of plasmids by Red/ET

1) Inoculate 25 ml LB containing 12 µg/ml Oxytetracycline and 50 µg/ml Kanamycin and grow the strain overnight at 30°C 200 rpm on a rotary shaker

2) Inoculate 20 µL of the well grown *E. coli* overnight culture into a 2 ml Eppendorf tube with a punctured lid containing 1.5 ml LB with 12 µg/ml Oxytetracycline and 50 µg/ml Kanamycin

3) Incubate the culture for 2h until OD reaches approx. 0.2 (sometimes longer incubation times are necessary)

4) Add 20 µL 10% arabinose solution, incubate at 37°C, 200 rpm for additional 45 minutes to ensure optimum expression of the recombinases

5) Centrifuge the culture at 8000 x g for 5 minutes to pellet the cells

6) Wash the residual cell pellet 2 times with 1000 µL autoclaved milliqH$_2$O and discard the supernatant.

7) Re suspend the cell pellet in 50 µL autoclaved milliq H$_2$O and add the PCR product for modification

8) Electroporate the mixture at 1300V, 25µF, 600 Ohm, 1mm Cuvette

9) Add 1mL fresh sterile LB medium, incubate for 1h at 900 rpm 30°C in a 2ml Eppendorf tube with a punctured lid

10) Transfer the mixture into 50 ml LB Kan50 containing appropriate concentration of the antibiotic on the modification PCR product (in my case Chloramphenicol 25µg/ml) and incubate over night at 30°C

11) Prepare plasmid from this mixture via alkaline lysis

12) Electroporate the plasmid mix into *E. coli DH10ß* according to the *E. coli DH10ß* electroporation protocol and plate it on LB Kan50 Cm25

13) Pick around 4 clones and cultivate them for plasmid preparation via alkaline lysis

14) Confirm the electroporated plasmid via restriction digestion

The retransformation is necessary in order to avoid having clones harboring both plasmids (the modified and the unmodified one) in the same *E. coli* clone as well as to get rid of residual pSC101 BAD Red gba plasmid.

S 3.1.14   Verification of myxobacterial mutant clones

Verification of the myxobacterial mutant clones is done by PCR with TAQ DNA Polymerase (Thermo Scientific). Therefore, genomic DNA from the myxobacterium is isolated with the Puregene Core Kit A (Quiagen) according to the protocol for gDNA isolation from gram negative bacteria. The Taq DNA Polymerase master mix is prepared according to section S 3.1.5.

S 3.1.14.1   Taq polymerase cycler program for mutant verification:

Table S 4. Taq DNA Polymerase PCR program

Step	Time [min:s]	Temperature [°C]
Initial denaturation	4:00	95
	0:30	95
Cycle, repeat 30x	0:15	63
	1:20	72
Final elongation	10	72
Store	forever	8

S 3.1.14.2   Verification of the MCy9280::pCR2.1 FtaD homology mutant

MCy9280::pCR2.1_FtaD_homology DNA lanes 1, 2 3 and 5
MCy9280 wild type DNA lane 4
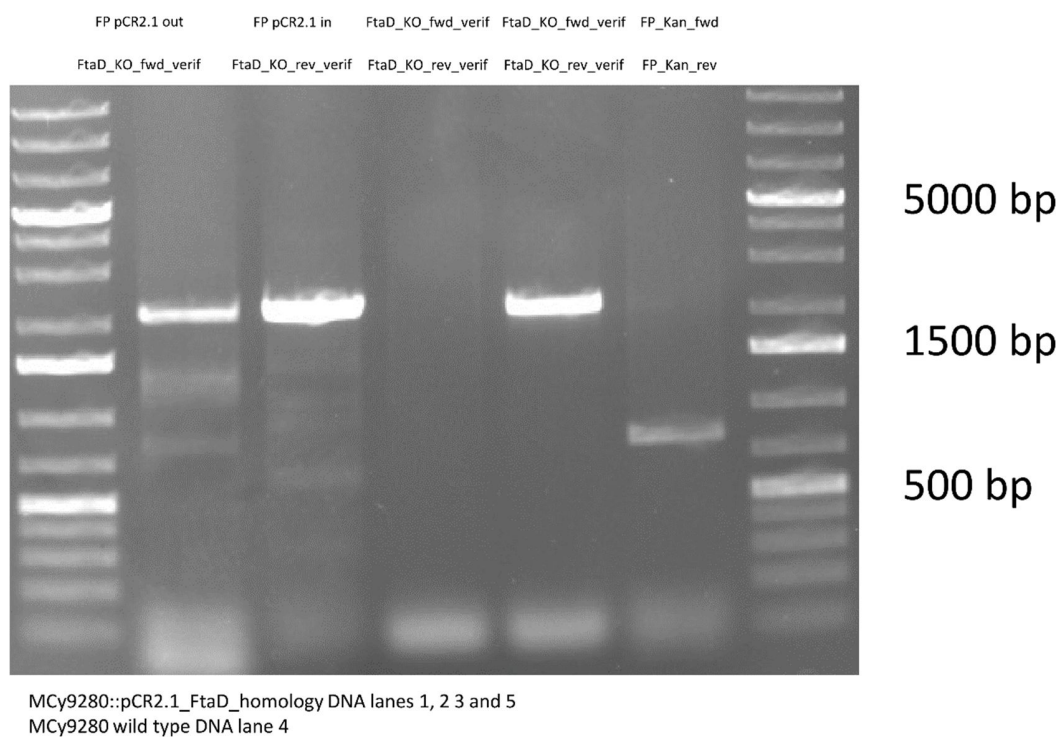
Figure S 2. Verification PCRs for the integration of pCR2.1_FtaD_homology into the FtaD gene of the Fulvuthiacene biosynthesis cluster

One can clearly see two 2 kb bands indicating the integration of the vector into the FtaD gene (lanes 1 and 2) as well as the absence of the band using both verification primers in the mutant gDNA (lane 3) that can be amplified from the wild type gDNA (lane 4). Also the control PCR for the presence of the kanamycin resistance gene on the pCR2.1 vector backbone is positive meaning that integration of the

pCR2.1 FtaD homology plasmid is complete and we do not observe residual intact FtaD that can be amplified from the mutant's gDNA. Presence of a band in lane 3 would mean the mutant is a mixed clone.

S 3.1.15   Creation of the vector for hybrid Myxothiazol production

S 3.1.15.1   Creation of the pFPVan FtaMN tn5 GH

In order to evaluate the functions of the enzymes FtaM, FtaN, FtaG and FtaH, these genes are to be transferred from the fulvuthiacene producer *Myxococcus fulvus* MCy9280 to the Myxothiazol A producer *Stigmatella aurantiaca* DW4/3-1.

Vector construction starts with the pCR2.1 vector (Thermo scientific TOPO TA cloning kit) featuring a vanillate promotor fused to the vanillate repressor called pFPvan as described in literature. [2] The genes *ftaM* and *ftaN* are amplified by PCR using the Phusion PCR protocol (section 1.4) and ligated into pFPvan directly downstream to the vanillate promotor using the restriction endonucleases EcoRI and NdeI.
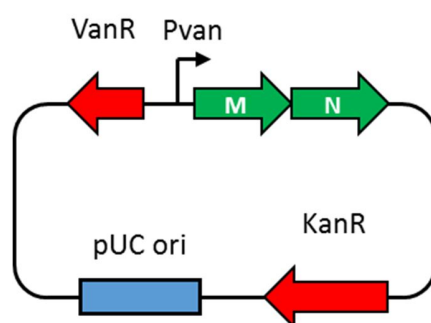
Figure S 3. *pFPVan vector featuring the genes* ftaM *and* ftaN

After construct verification by restriction digestion and Sanger sequencing (LGC ltd., U.K.), this construct is further modified by introduction of a PCR product containing the genes FtaG and FtaH on it using the EcoRI restriction endonuclease.
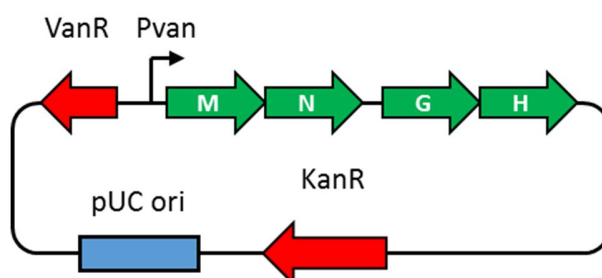
Figure S 4. *pFPVan vector featuring the genes* ftaM, ftaN, ftaG *and* ftaH

After selection of a clone having the PCR product ligated in the correct orientation using restriction digestion and subsequent sanger sequencing the construct is modified by Red E/T to integrate a tn5 promotor upstream of the *ftaGH* operon (section 1.11). Therefore, overlap extension PCR using the Phusion PCR protocol (section S 1.3) is used to fuse a tn5 promotor with a 50 base overhang of the FtaG gene start to a CmR resistance cassette having a 50 base overhang to the end of the FtaN gene. Furthermore, the construct pFPVan FtaMNGH is transformed into the RedET strain HS996 pSC101 BAD RedγβαA to be able to use λ-red prophage mediated recombination. Linear to circular Red E/T is used to integrate this linear DNA fragment harboring homology to the region downstream of *ftaN*, a chloramphenicol resistance cassette fused to a tn5 promotor equipped with a homology to the gene start of *ftaG*. The resulting plasmid is verified by restriction analysis and Sanger sequencing (LGC ltd., U.K.).
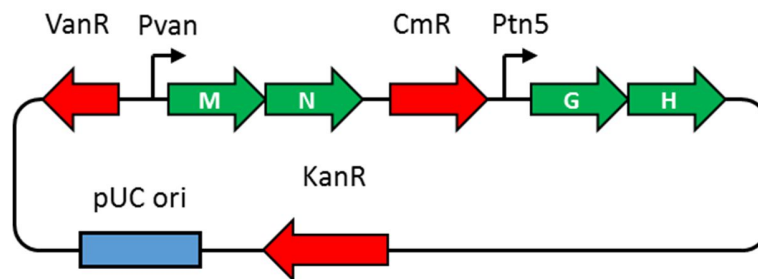
Figure S 5. *pFPVan vector featuring the genes* ftaM *and* ftaN *under the vanillate promotor and repressor control and* ftaG *and* ftaH *under tn5 promotor control*

In order to integrate this vector into the target organism that is the Myxothiazol A producer *S. aurantiaca* DW4/3-1 we integrated the Mx8 integrase gene into this construct by conventional restriction cloning with the endonuclease XbaI to obtain the pFPvan FtaMN tn5 GH construct.[3]
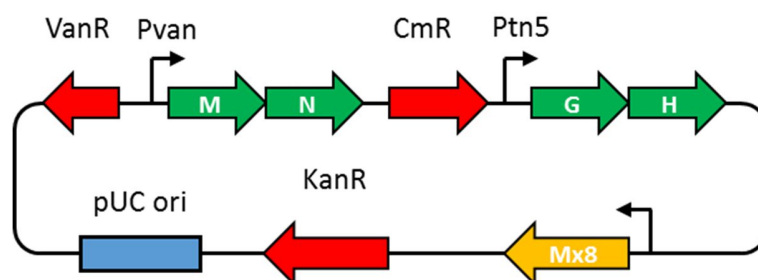
Figure S 6. *pFPvan FtaMN tn5 GH Mx8 vector to transfer the genes* ftaM,N,G *and* H *into* S. aurantiaca *DW4/3-1*

This construct is able to integrate into the *S. aurantiaca* DW 4/3-1 genome via the Mx8 integrase attB site.[3,4] The plasmid is electroporated into *Stigmatella aurantiaca* DW4/3-1 and clones appear after 2 weeks. The clones are genotypically verified by PCR and grown in TS-6 medium supplemented with

50μg/ml Kanamycin. Production of the hybrid Myxothiazol / Fulvuthiacene products is analyzed by extraction followed by LC-MS analysis according to standard protocol. We are now able to study the effect of *ftaG* and *ftaH* on the Myxothiazol production in *S. auratiaca* DW4/3-1 as well as the combined effects of *ftaM, ftaN, ftaG* and *ftaH* as the tn5 promotor is constitutively active while the vanillate repressor can be removed by adding 5 mM sodium vanillate into the medium during fermentation.

S 3.2   *In silico* analysis of the fulvuthiacene gene cluster

S 3.2.1   *In silico* analysis of the Fta biosynthetic gene cluster

Table S 5. genes of the fulvuthiacene biosynthesis gene cluster, predicted gene function and their closest homologue determined by blastp search against NCBI non-redundant protein database (nr)

Name	Length [bp]	Closest homologue in the nr database at NCBI	Identity [%] and length of alignment	Proposed function	Accession Nr.
ftaA	9384	MtaB (Stigmatella Aurantiaca)	63.3% (2846bp)	Polyketide synthesis	AAF19810
ftaB	9669	Hybrid NRPS/ type 1 PKS (Myxococcus stipitatus)	70.8% (3198bp)	Polyketide synthesis, Thiazole formation	WP_015350515
ftaC	5799	Polyketide synthase (Myxococcus Stipitatus)	79.4% (1931bp)	Polyketide Synthesis, O-Methylation	WP_015350514
ftaD	4071	Polyketide synthase (Stigmatella Aurantiaca)	84.2% (1363bp)	Polyketide synthesis, Enolether formation	WP_002618707
ftaE	5238	MelG (Melittangium Lichenola)	71.5% (1747bp)	Glycine attachment Glycine hydroxylation	CAD89778
ftaF	1026	Fumarylacetoacetate hydrolase (Myxococcus Stipitatus)	86.8% (341bp)	Alphahydroxyglycine hydrolysis	WP_015350511
ftaG	909	CtaK (Cystobacter Fuscus)	71.0% (297bp)	Methyl transfer	AAW03333
ftaH	1053	N-Carbamoylputrescine hydrolase (Myxococcus Stipitatus)	80.7% (332bp)	Hydolysis of terminal amide	AGC46253
ftaI	1431	Adenosylhomocysteinase (Myxococcus Stipitatus)	98.8% (476bp)	Adenosylhomocysteine detoxing	WP_015350518
ftaJ	897	bifunctional 5,10-methylene-tetrahydrofolate dehydrogenase/5,10-methylene-tetrahydrofolate cyclohydrolase (Myxococcus Stipitatus)	93.2% (293bp)	Methionine recycling from generated homocysteine	WP_015350519
ftaK	534	Hypothetical protein (Myxococcus Stipitatus)	80.9% (177bp)		WP_015350520
ftaL	1212	Acetyl-CoA Acetyltransferase / Thioesterase (Myxococcus Stipitatus)	95.5% (403bp)	might be proofreading	WP_015350521
ftaM	2178	EnoylCoA dehydratase (Myxococcus Stipitatus)	95% (725bp)	*in trans* enoyl reduction	WP_015350522
ftaN	969	Methyltransferase Typ11 (Myxococcus Stipitatus)	94.4% (322bp)		WP_015350523
orf1	198	Lysine Biosynthesis protein LysW (Myxococcus Stipitatus)	93% (43bp)		WP_063639811
orf2	864	Phage Major Capsid Protein (Micromonospora citrea)	52.5% (276bp)		SCL67435
orf3	1599	Lysyl Endopeptidase (Myxococcus Stipitatus)	90.2% (532bp)		WP_015350508
orf4	321	Hypothetical Protein (Myxococcus Stipitatus)	72.9% (107bp)		AGC46250

orf5	474	Hypothetical Protein (Stigmatella Aurantiaca)	74.5% (146bp)	WP_002609340
orf6	2004	Thioredoxin (Stigmatella Aurantiaca)	59.8% (636bp)	WP_002609295
orf7	639	Hypothetical Protein (Stigmatella Aurantiaca)	44.3% (206bp)	WP_013377569
orf8	690	Phage tail fiber Protein (Cystobacter fuscus)	43.0 (131bp)	EPX62440
orf9	198	No blast hit		
orf10	567	Hypothetical Protein (Myxococcus Stipitatus)	85.1% (188bp)	WP_015350505

The fulvuthiacene biosynthetic gene cluster from *M. fulvus* MCy9280 will be deposited in GenBank under accession number MH069655.

S 3.2.2 *In silico* comparison of the Fulvuthiacene, Melithiazol and Myxothiazol gene clusters

As it became evident that the three assembly lines of Fulvuthiacene, Melithiazol and Myxothiazol are strongly related, we did module by module protein alignments in order to determine possible ancestry and recombination events leading to the divergence of these clusters. Therefore, all modules having the same architecture are compared in a whole model alignment. An overview about the modules that are compared here can be found in the gene cluster comparison, main text Figure 7.
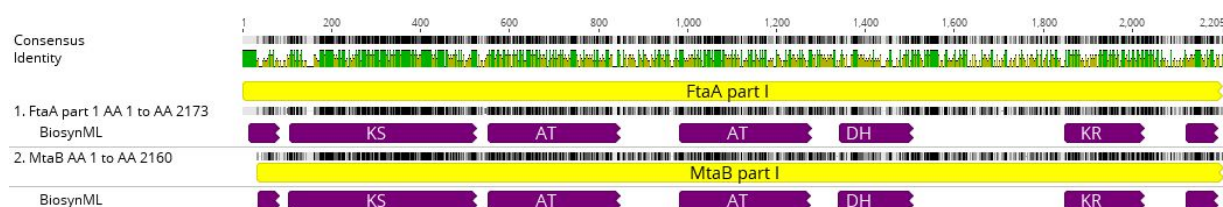
Figure S 7. *Protein alignment of FtaA module 1 and MtaB module 1 via muscle alignment* [5]

The modules FtaA module 1 and MtaB module 1 share a protein identity of 61.8 %.
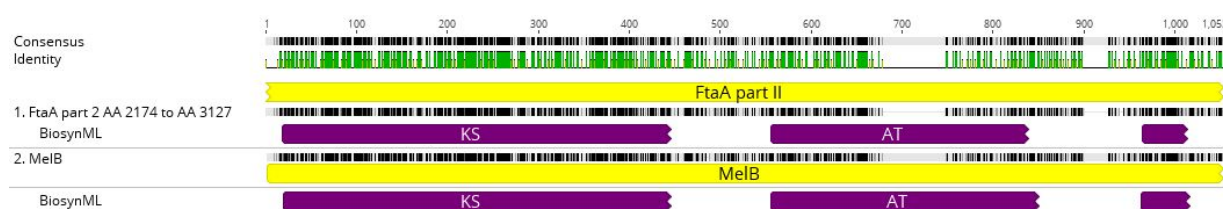
Figure S 8. *Protein alignment of FtaA module 2 and MelB via muscle alignment* [5]

The modules FtaA module 2and MelB share a protein identity of 53.5 % also it is worth noting that a 70 amino acid part of the Fta module 2 acyl transferase (AT) is deleted rendering Fta module 2 inactive.
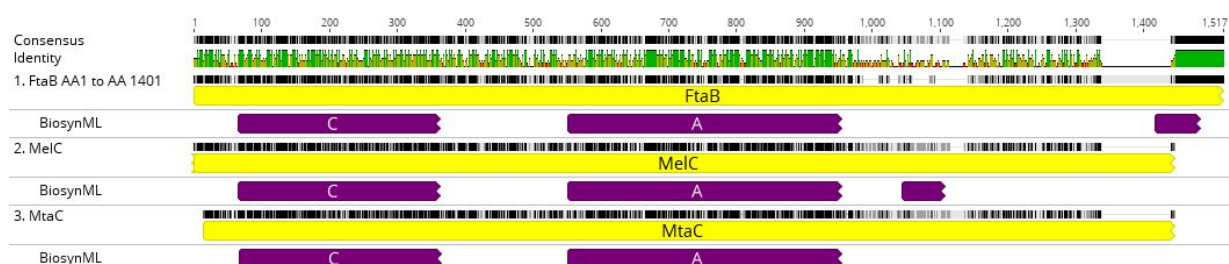
Figure S 9. Protein alignment of FtaB module 1, MelC and MtaC via muscle alignment [5]

pairwise identity	FtaB module 1	MelC	MtaC
FtaB module 1		50.42	51.68
MelC	50.42		65.31
MtaC	51.68	65.31	

Considering the genes compared in Figure 9 one can observe that homology of MelC and MtaC is generally well, while the homology of FtaB to MtaC and MelC is pretty good until the end of the cysteine specific A domain.
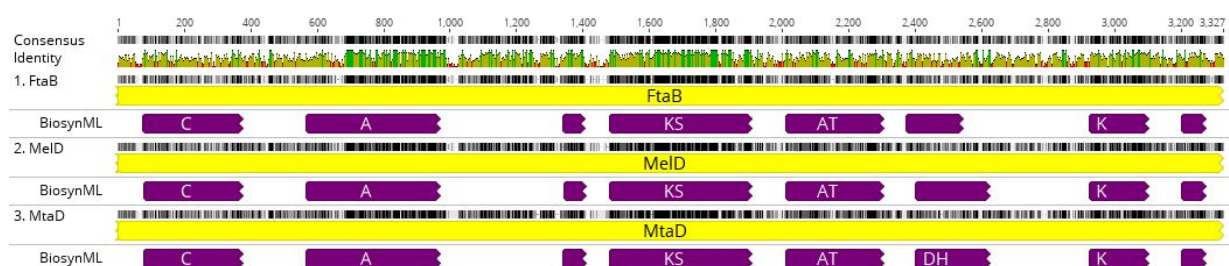
Figure S 10. Protein alignment of FtaB, MelD and MtaD via muscle alignment [5]

pairwise identity	FtaB	MelD	MtaD
FtaB		56.21	57.93
MelD	56.21		63.1
MtaD	57.93	63.1	

Once again we observe pretty good homology between MtaD and MelD while the protein FtaB is somewhat dissimilar from the gene start to the adenylation domain while the rest of the protein has good homology to its counterparts in the Melithiazol and Myxothiazol biosynthesis. It is therefore likely that FtaB was created by homologous recombination of a melC/MtaC like protein and a melD/mtaD protein and the crossover event occurred at the adenylation domain.
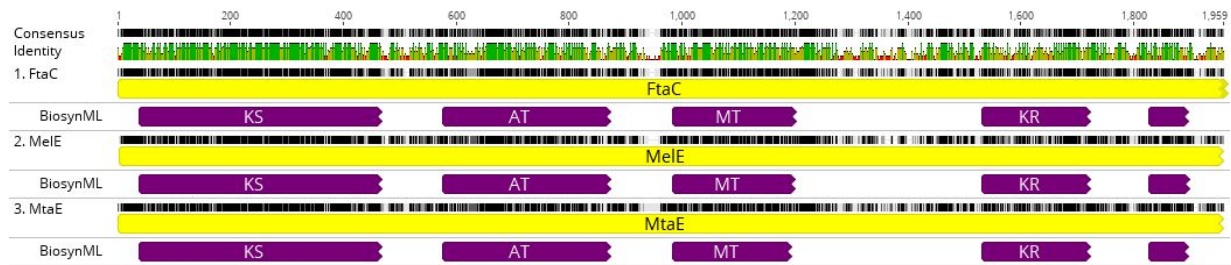
Figure S 11. *Protein alignment of FtaC, MelC and MtaC via muscle alignment* [5]

pairwise identity	FtaC	MelE	MtaE
FtaC		66.7	67.43
MelE	66.7		68.08
MtaE	67.43	68.08	
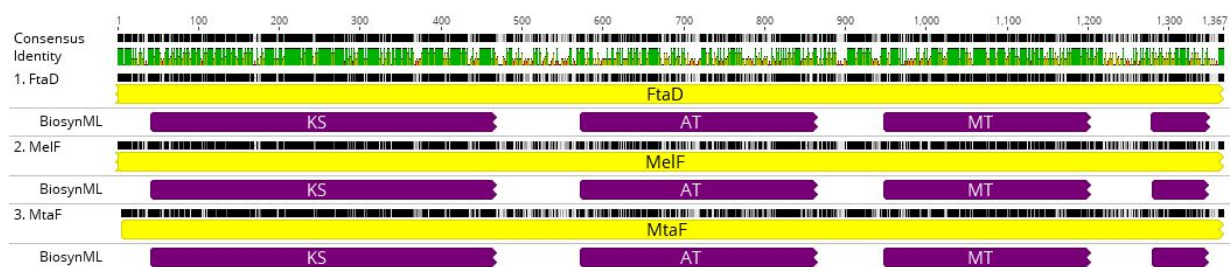
Figure S 12. *Protein alignment of FtaD, MelF and MtaF via muscle alignment* [5]

pairwise identity	FtaD	MelF	MtaF
FtaD		65.74	68.23
MelF	65.74		67.89
MtaF	68.23	67.89	
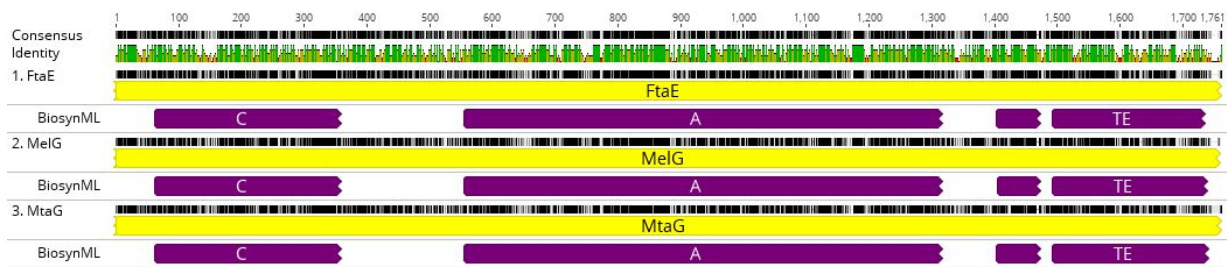
Figure S 13. *Protein alignment of FtaE, MelG and MtaG via muscle alignment* [5]

pairwise identity	FtaE	MelG	MtaG
FtaE		70.23	70.67
MelG	70.23		70.53
MtaG	70.67	70.53	
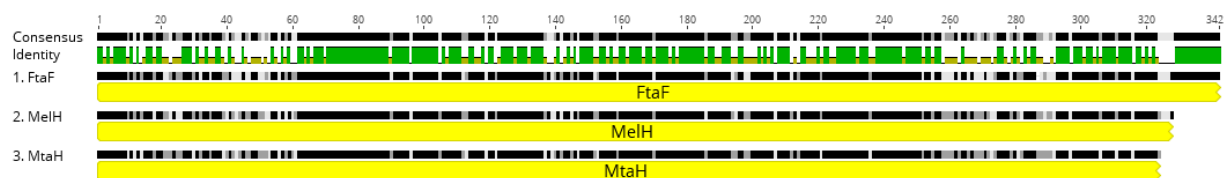
Figure S 14. *Protein alignment of FtaF, MelH and MtaH via muscle alignment* [5]

pairwise identity	FtaF	MelH	MtaH
FtaF		65.55	66.67
MelH	65.55		72.53
MtaH	66.67	72.53	
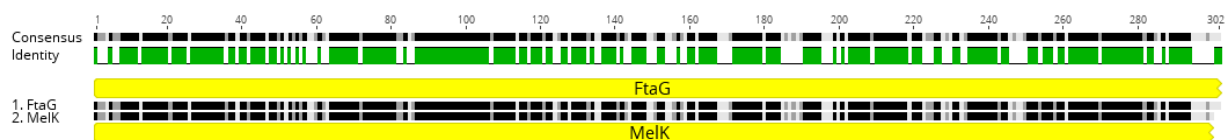
Figure S 15. *Protein alignment of FtaF and MelK via muscle alignment* [5]

FtaG and MelK share a protein identity of 68.7 %.
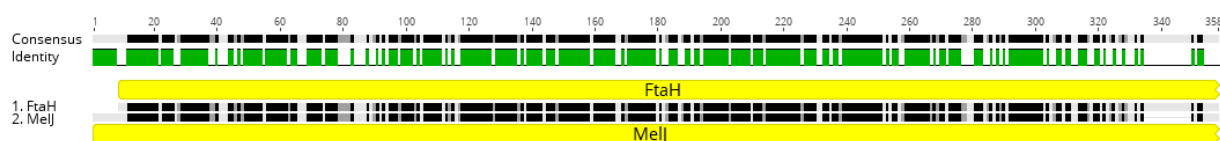
Figure S 16. *Protein alignment of FtaF and MelK via muscle alignment* [5]

FtaH and MelJ share a protein identity of 66.9 %.

S 3.2.3   *In silico* analysis of the FtaM protein

As described in this work, during the fulvuthiacene biosynthesis there has to be an *in trans* acting enoyl reductase as the FtaA protein does not contain an enoyl reductase (ER) domain while the fulvuthiacenes are fully reduced (main text, Figure 3). The most likely candidate for a *trans* acting enoyl reductase is the protein FtaM. This protein that is encoded in the biosynthetic gene cluster right next to the assembly line proteins is a homologue of the FabJ protein involved in the fatty acid β-oxidation pathway. These proteins consist of a crotonase fold, which can accommodate αβ unsaturated esters with their oxyanion hole as well as a NAD binding pocket with the corresponding dehydrogenase domain. The domains depicted in Figure S 17 are predicted by NCBI conserved domain (CD) search [6].
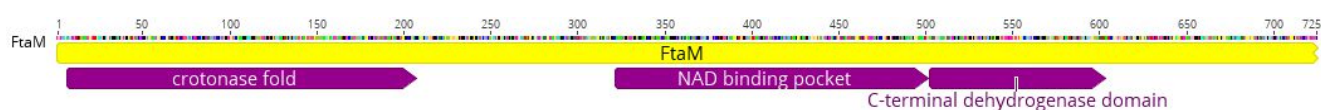
Figure S 17. *Position of the crotonase fold, the NAD binding pocket and the dehydrogenase domain on the FtaM primary sequence*

FabJ like proteins are mainly acting as oxidoreductases on carbon oxygen bonds and therefore slightly different from classical enoyl reductase proteins but since both of these proteins belong to the short chain dehydrogenase/reductase family and the crotonase fold has the right substrate specificity for α,β unsaturated esters, FtaM is likely to have evolved from FabJ to catalyze the carbon carbon bond reductions in module 1 of the Fulvuthiacene biosynthesis.

S 3.2.4 ***In silico*** **analysis of the Fulvuthiacene's stereochemistry**

In order to elucidate the stereochemistry of the Fulvuthiacenes at C-12 and C-13 in Fulvuthiacene A (Figure S 22) we chose to predict the stereochemistry at these two carbon atoms by in - silico analysis of the ketoreductase (KR) domain in module 5 on FtaC that forms both stereo centers (main text Figure 3). It has been shown by Kitsche et al. that analysis of the KR sequence is sufficient to predict the stereo chemistry outcome of a keto-reduction reaction and the stereochemistry of the corresponding α – methyl branch incorporated through the use of methyl – malonyl CoA in this module of the biosynthesis (main text Figure 3).[7]

The amino acid sequence of this KR domain used for the predictions is:

> KR module 5

GTYLITGGLGGLGVQTAKWLVARGVRHLVLAGRTEVPNAAVWDSLPEDSPHAAHVQMLRGLQAQGASVRYVRCDLG EPAQVRALVDTCGVGGVPLVGVLHAAGASAHHPWMETGPEVLASVFQAKALGAWLLHEFTRDLPLHCFVLFSSASAV WGAQGMAAYAAANHFLDALAHHRKARGLPATSVNWGRWSE

S 3.2.4.1   Prediction of the stereochemistry of the ketoreduction reaction:

KR of the Fulvuthiacene assembly line

```
 [1]  "G" "T" "Y" "L" "I" "T" "G" "G" "L" "G" "G" "L" "G" "V" "Q" "T" "A" "K"
[19]  "W" "L" "V" "A" "R" "G" "V" "R" "H" "L" "V" "L" "A" "G" "R" "T" "E" "V"
[37]  "P" "N" "A" "A" "V" "W" "D" "S" "L" "P" "E" "D" "S" "P" "H" "A" "A" "H"
[55]  "V" "Q" "M" "L" "R" "G" "L" "Q" "A" "Q" "G" "A" "S" "V" "R" "Y" "V" "R"
[73]  "C" "D" "L" "G" "E" "P" "A" "Q" "V" "R" "A" "L" "V" "D" "T" "C" "G" "V"
[91]  "G" "G" "V" "P" "L" "V" "G" "V" "L" "H" "A" "A" "G" "A" "S" "A" "H" "H"
[109] "P" "W" "M" "E" "T" "G" "P" "E" "V" "L" "A" "S" "V" "F" "Q" "A" "K" "A"
[127] "L" "G" "A" "W" "L" "L" "H" "E" "F" "T" "R" "D" "L" "P" "L" "H" "C" "F"
[145] "V" "L" "F" "S" "S" "A" "S" "A" "V" "W" "G" "A" "Q" "G" "M" "A" "A" "Y"
[163] "A" "A" "A" "N" "H" "F" "L" "D" "A" "L" "A" "H" "H" "R" "K" "A" "R" "G"
[181] "L" "P" "A" "T" "S" "V" "N" "W" "G" "R" "W" "S" "E"
```

Depending on a ScoreDiff value of -26.72 the L-form of secondary alcohols is predicted

S 3.2.4.2 Prediction of the stereochemistry of the α – methyl branch originating from methylmalonyl CoA incorporation:

KR of the Fulvuthiacene biosynthesis

```
  [1] "G" "T" "Y" "L" "I" "T" "G" "G" "L" "G" "G" "L" "G" "V" "Q" "T" "A" "K"
 [19] "W" "L" "V" "A" "R" "G" "V" "R" "H" "L" "V" "L" "A" "G" "R" "T" "E" "V"
 [37] "P" "N" "A" "A" "V" "W" "D" "S" "L" "P" "E" "D" "S" "P" "H" "A" "A" "H"
 [55] "V" "Q" "M" "L" "R" "G" "L" "Q" "A" "Q" "G" "A" "S" "V" "R" "Y" "V" "R"
 [73] "C" "D" "L" "G" "E" "P" "A" "Q" "V" "R" "A" "L" "V" "D" "T" "C" "G" "V"
 [91] "G" "G" "V" "P" "L" "V" "G" "V" "L" "H" "A" "A" "G" "A" "S" "A" "H" "H"
[109] "P" "W" "M" "E" "T" "G" "P" "E" "V" "L" "A" "S" "V" "F" "Q" "A" "K" "A"
[127] "L" "G" "A" "W" "L" "L" "H" "E" "F" "T" "R" "D" "L" "P" "L" "H" "C" "F"
[145] "V" "L" "F" "S" "S" "A" "S" "A" "V" "W" "G" "A" "Q" "G" "M" "A" "A" "Y"
[163] "A" "A" "A" "N" "H" "F" "L" "D" "A" "L" "A" "H" "H" "R" "K" "A" "R" "G"
[181] "L" "P" "A" "T" "S" "V" "N" "W" "G" "R" "W" "S" "E"
```

Depending on a ScoreDiff value of 36.03 and an a priori known L-form of the secondary alcohol the D-form of methyl branches is predicted

It is worth noting that D – and L – form as predicted by this analysis correspond to the absolute stereochemistry with respect to the thioester bound intermediate. Since Kitsche and Kalesse recommend a minimum scoreDiff value of at least ± 15 for a decision on the predicted stereochemistry we can assume our stereochemistry predictions to be valid (values close to 0 would mean an uncertain prediction, values above ± 15 were determined to be meaningful).[7] This stereochemistry prediction fits with the fulvuthiacene stereochemistry experimentally determined supporting these experimental results to be correct.

S 3.3 Extract generation, LC-MS data acquisition and clustering

S 3.3.1 Myxobacterial fermentation conditions

Cultures for UHPLC/HRMS analysis are grown in 300 ml shake flasks containing 50 ml of VYGS medium for *Myxococcus fulvus* MCy9280 or 50 ml of TS-6 medium for *Stigmatella aurantiaca* DW4/3-1 inoculated with 1 ml of pre culture. Media for mutant *S. aurantiaca* strains were supplemented with 50 mg/L kanamycin (Roth). After inoculation the medium is supplemented with 2% of sterile XAD-16 adsorber resin (Sigma Aldrich) suspension in water. *Myxococcus fulvus* strain MCy9280 is grown in VYGS medium at 30°C for 12 Days on an orbitron shaker at 160 rpm. *Stigmatella aurantiaca* DW4/3-1 is grown in TS-6medium at 30°C for 10 days. Fermentation cultures of *Myxococcus fulvus* strain MCy9280 used for isolation of the

fulvuthiacenes are grown in 5 5L shake flasks containing 2L of VYGS medium each. Fermentation duration is 14 days post inoculation; inoculation is done with 50 ml of pre-culture and the medium is supplemented with 2% of sterile XAD-16 adsorber resin suspension in water. Fermentation cultures of *Stigmatella aurantiaca* DW4/3-1 used for isolation of the myxothiazol Z and desmethyl-myxothiazol Z are grown in 4 5L shake flasks containing 2 L of TS-6 medium each. Fermentation duration is 14 days post inoculation; inoculation is done with 50 ml of pre-culture and the medium is supplemented with 2 % of sterile XAD-16 adsorber resin suspension in water.

S 3.3.2   Metabolite extraction procedure for analytical scale extractions

The frozen cell pellet is transferred into a 100 ml Erlenmeyer flask and a magnetic stirrer is added. 50 ml of acetone (fluka analytical grade, redistilled in house) are added onto the pellet and the mixture is stirred for 60 min. on a magnetic stirrer. The Acetone has to be distilled beforehand since fluka analytical reagent grade acetone still contains trace amount of polymer impurities that will otherwise concentrate in the extract. The acetonic extract is left to settle in order to sediment cell debris and XAD resin for a second extraction step. The supernatant is filtered with a 125 micron folded filter keeping cell pellet and XAD-16 resin in the Erlenmeyer flask for a second extraction step. The residual pellet and XAD-16 resin is extracted again with 30 ml of distilled acetone for 60 min on a magnetic stirrer and filtered through the same folded filter. The combined extracts are transferred into a 100 ml round bottom flask. The acetone is evaporated on a rotary evaporator at 260 mbar and 40 °C water bath temperature. The residual water is evaporated at 20 mbar until the residue in the flask is completely dry. The residue is taken up in 550 µl of methanol (Chromasolv HPLC grade, Sigma Aldrich) and transferred into a 1.5 ml Eppendorf tube. This tube is centrifuged with Heareus biofuge pico at 13000 rpm for 3 minutes to remove residual insolubilities such as salts, cell debris and XAD fragments.

S 3.3.3   Standardized UHPLC MS conditions

All measurements were performed on a Dionex (Germering, Germany) Ultimate 3000 RSLC system using a Waters (Eschborn, Germany) BEH C18 column (50 x 2.1 mm, 1.7 µm) equipped with a Waters VanGuard BEH C18 1.7 µm guard column. Separation of 1 µl sample was achieved by a linear gradient from (A) H_2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 600 µL/min and a column temperature of 45 °C. Gradient conditions were as follows: 0 – 0.5 min, 5% B; 0.5 – 18.5 min, 5 – 95% B; 18.5 – 20.5 min, 95% B; 20.5 – 21 min, 95 – 5% B; 21-22.5 min, 5% B. UV spectra were recorded by a DAD in the range from 200 to 600 nm. The LC flow was split to 75 µL/min before entering the Bruker Daltonics maXis 4G hrToF mass spectrometer (Bremen, Germany) using the Apollo II ESI source. Mass

spectra were acquired in centroid mode ranging from 150 – 2500 m/z at a 2 Hz full scan rate. Mass spectrometry source parameters are set to 500V as end plate offset; 4000V as capillary voltage; nebulizer gas pressure 1 bar; dry gas flow of 5 l/min and a dry temperature of 200°C. Ion transfer and quadrupole settings are set to Funnel RF 350 Vpp; Multipole RF 400 Vpp as transfer settings and Ion energy of 5eV as well as a low mass cut of 300 m/z as Quadrupole settings. Collision cell is set to 5.0 eV and pre pulse storage time is set to 5 µs. Spectra acquisition rate is set to 2Hz. Calibration is done automatically before every LC-MS run by injection of sodium formate and calibration on the sodium formate clusters forming in the ESI source. All MS analyses are acquired in the presence of the lock masses $C_{12}H_{19}F_{12}N_3O_6P_3$; $C_{18}H_{19}O_6N_3P_3F_2$ and $C_{24}H_{19}F_{36}N_3O_6P_3$ which generate the $[M+H]^+$ ions of 622.028960; 922.009798 and 1221.990638.

S 3.3.4 Methodology for statistics based metabolome filtering

To obtain all molecular features in the 6 LC/HRMS chromatograms of the bacterial extracts provided from three independent cultivations and the 6 LC/HRMS chromatograms of the corresponding three independent medium blanks, feature finding is applied that can subsequently be used for statistical data treatment. Therefore, the in-build molecular feature finder implemented in Bruker Compass Data Analysis 4.2 is used with the compound detection parameters SN threshold 1; Correlation coefficient 0.85; minimum compound length 10 spectra and smoothing width of 3 spectra. Bucketing is done with Bruker Profile Analysis 2.1 with advanced bucketing and advanced bucketing window parameters of 30s and 15 ppm. Bucket value is log transformed to avoid under evaluating low intensity signals in the presence of high intensity signals.[8] The PCA t-test function is used in order to separate medium derived MS features from the metabolome derived MS features since the t-test table that can be exported from Bruker Profile Analysis 2.1 contains information upon how many blanks and how many bacterial extracts contain said feature. This t-Test table is subsequently further processed with the KNIME software framework to obtain an SPL list that specifically and exclusively targets the bacterial metabolome for the acquisition of MS/MS spectra on our Bruker Daltonics maXis 4G hrToF spectrometer (Bremen, Germany).[9] The t-test table from the bucketing as well as the processed scheduled precursor list used to acquire SPL-MS/MS data will be supplied upon request.

S 3.3.5 Acquisition parameters for acquiring high-resolution tandem MS data

LC and MS conditions for SPL guided MS/MS data acquisitions are kept constant according to section standardized UHPLC-MS conditions. MS/MS data acquisition parameters are set to exclusively fragment scheduled precursor list entries. SPL tolerance parameters for precursor ion selection are set to 0.2 minutes and 0.05 m/z in the SPL MS/MS method. The method picks up to 2 precursors per cycle, applies

smart exclusion after 5 spectra and performs CID and MS/MS spectra acquisition time ramping. CID Energy is ramped from 35 eV for 500 m/z to 45 eV for 1000 m/z and 60 eV for 2000 m/z. MS full scan acquisition rate is set to 2Hz and MS/MS spectra acquisition rates are ramped from 1 to 4 Hz for precursor Ion intensities of 10 kcts. to 1000 kcts..

S 3.3.6   Spectral networking details for the acquired tandem MS data

 All supporting GNPS clustering data presented here is created based on exported .mzML files from the LC-MS² chromatograms using the parameters specified in the experimental section of the main text.[10] The entire MS/MS chromatograms are exported containing all MS/MS data as an .mzML data file and uploaded to the GNPS server at University of California San Diego via FileZilla FTP upload to ftp:// ccms-ftp01.ucsd.edu and all acquired SPL MS/MS spectra are used for spectral network creation. A molecular network was created using the online workflow at GNPS. The data was filtered by removing all MS/MS peaks within +/- 17 Da of the precursor m/z. MS/MS spectra were window filtered by choosing only the top 6 peaks in the +/- 50 Da window throughout the spectrum. The data was then clustered with MS-Cluster with a parent mass tolerance of 0.05 Da and a MS/MS fragment ion tolerance of 0.1 Da to create consensus spectra. Further, consensus spectra that contained less than 1 spectra were discarded. A network was then created where edges were filtered to have a cosine score above 0.65 and more than 4 matched peaks. Further edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes.[10] The complete clustered dataset is downloaded from the server and subsequently visualized using Cytoscape 3.6.1.

 S 3.3.6.1   Performance comparison of auto-MS² experiments versus SPL based MS² experiments

 As described in this work, the devised SPL based MS² method, as it is specifically designed to target the bacterial metabolome of interest, should have a much higher coverage of MS² spectra across our range of bacterial metabolome precursor peaks determined by PCA. We evaluated this statement by determination of the SPL coverage – a metric that determines the number of PCA precursors actually trigger acquisition of a MS² spectrum. We used sql queries in the KNIME software framework to match csv file entries and visualized the results as Venn diagrams.
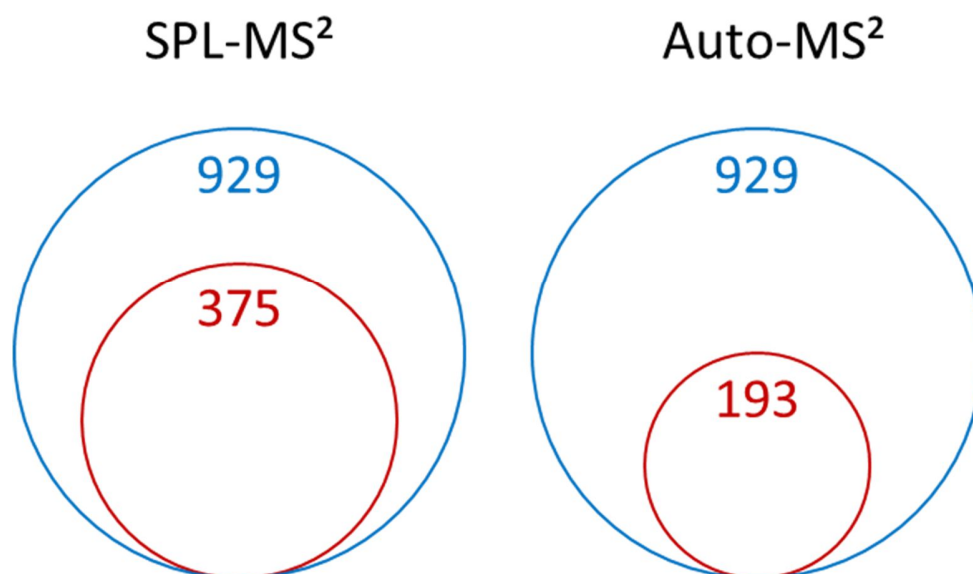
Figure S 18. Comparison of the SPL coverage using the SPL-MS² and the Auto-MS² based LC-MS² experiments for measurements on the MCy9280 Vy/2 medium extract

As expected and in line with previous studies, the SPL coverage reached with a SPL based MS² is significantly better than with the auto-MS² method. [9] Compared to the results presented by Hoffmann et al. our SPL coverages are overall lower, which is readily explained as we increased the cycle time to obtain fewer MS² spectra of better product ion intensity. Although SPL coverage drops, this leads to better predictivity of the resulting spectral networks in the spectral networking step. If one considers now that in AutoMS² mode there is a total of 997 MS² spectra acquired while only 193 of those are dedicated to cover the metabolome of interest one can estimate that around 4/5 of the MS² acquisition time in Auto MS² mode is dedicated to acquiring a priori non-interesting spectra.

S 3.3.7   Clustering of tandem MS² data acquired in auto-MS² mode

The complete set of MS² spectra clusters can be seen in Figure S 19. The cytoscape data can be provided as a .cys file upon request.
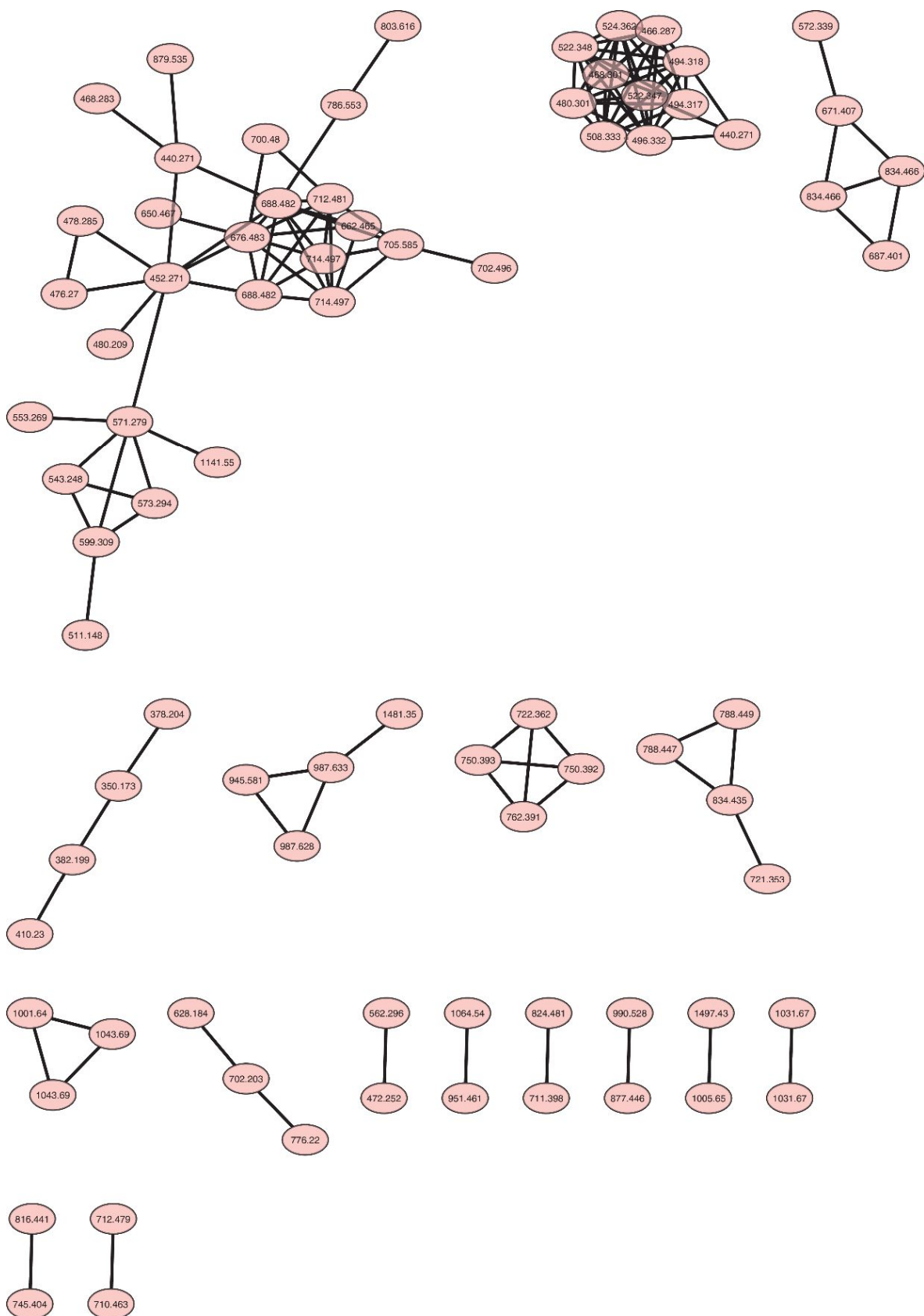
Figure S 19. *GNPS based MS² clustering of the automatic precursor selection MS² data visualized with Cytoscape*

As already mentioned in the main text, automatic precursor selection lacks the ability to selectively target the bacterial secondary metabolome. Therefore, the spectral network shown in Figure S 19 is not only comprised of features belonging to the bacterial metabolome but also of features that were fragmented that provide from the bacterial culture medium. Fragmentation of these features belonging to the culture medium represent wasted MS² acquisition time, as these spectra do not add any information about the microorganism in question. These features providing from the medium blank are best represented by the medium peptide block in the top left corner.
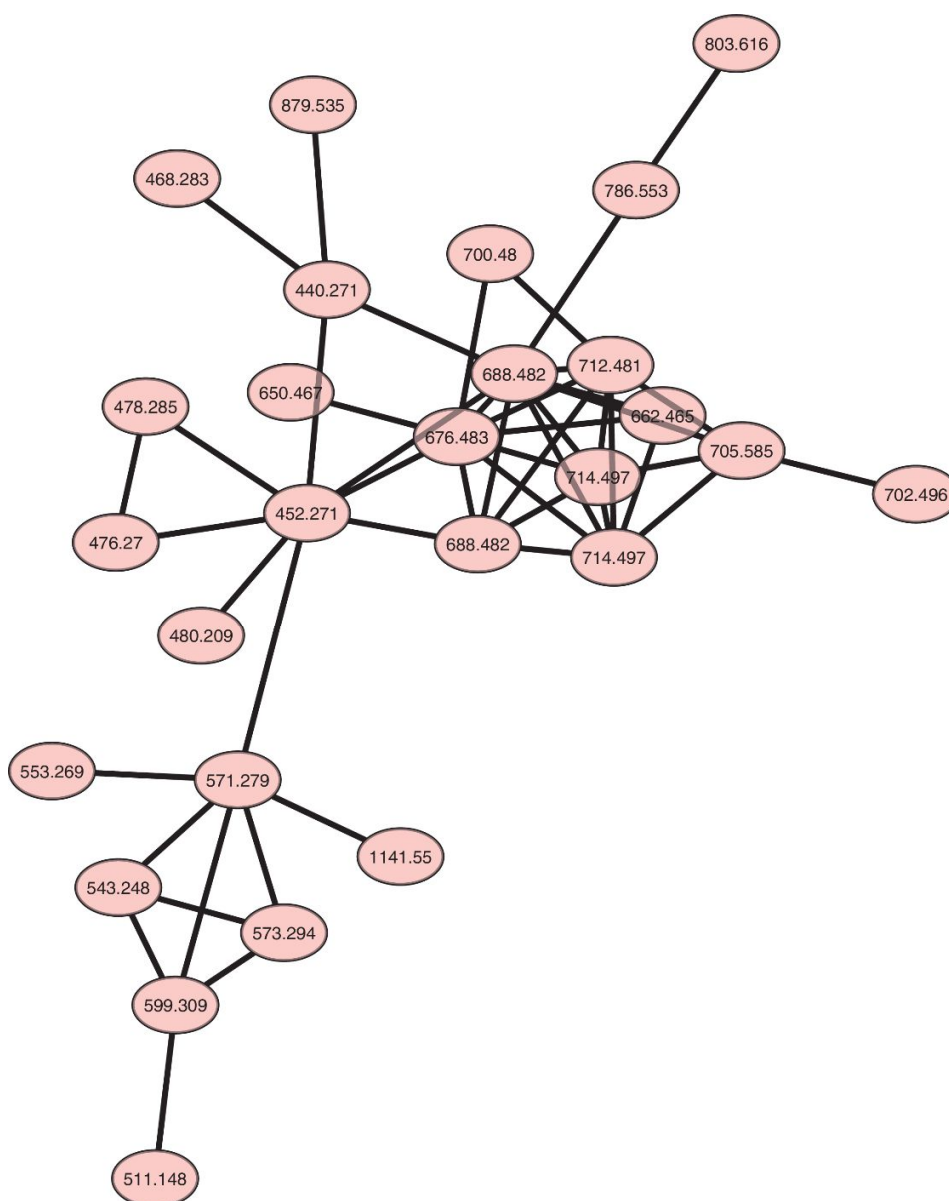
Figure S 20. Medium peptide cluster appearing in automatic precursor selection based GNPS MS² clustering

These features are characterized by a regular amino acid fragmentation as well as the fact that they are also present in the medium blank. The fact that the spectrometer is occupied with measuring MS² spectra of these peptide fragments while possibly features belonging to the secondary metabolome of lower intensity are eluting also explains the higher sensitivity of the SPL based precursor selection mode (Figure S 21). This finding is also verified by the incapacity of the automatic precursor selection MS² method to detect the fulvuthiacenes' biological precursors (the prefulvuthiacenes and desmethyl-fulvuthiacenes) as well as the minor derivative fulvuthiacene C.

S 3.3.8   Clustering of tandem MS² data acquired in SPL-MS² mode

The complete set of MS² spectra clusters can be seen in Figure S 21. The cytoscape data can be provided as a .cys file upon request.
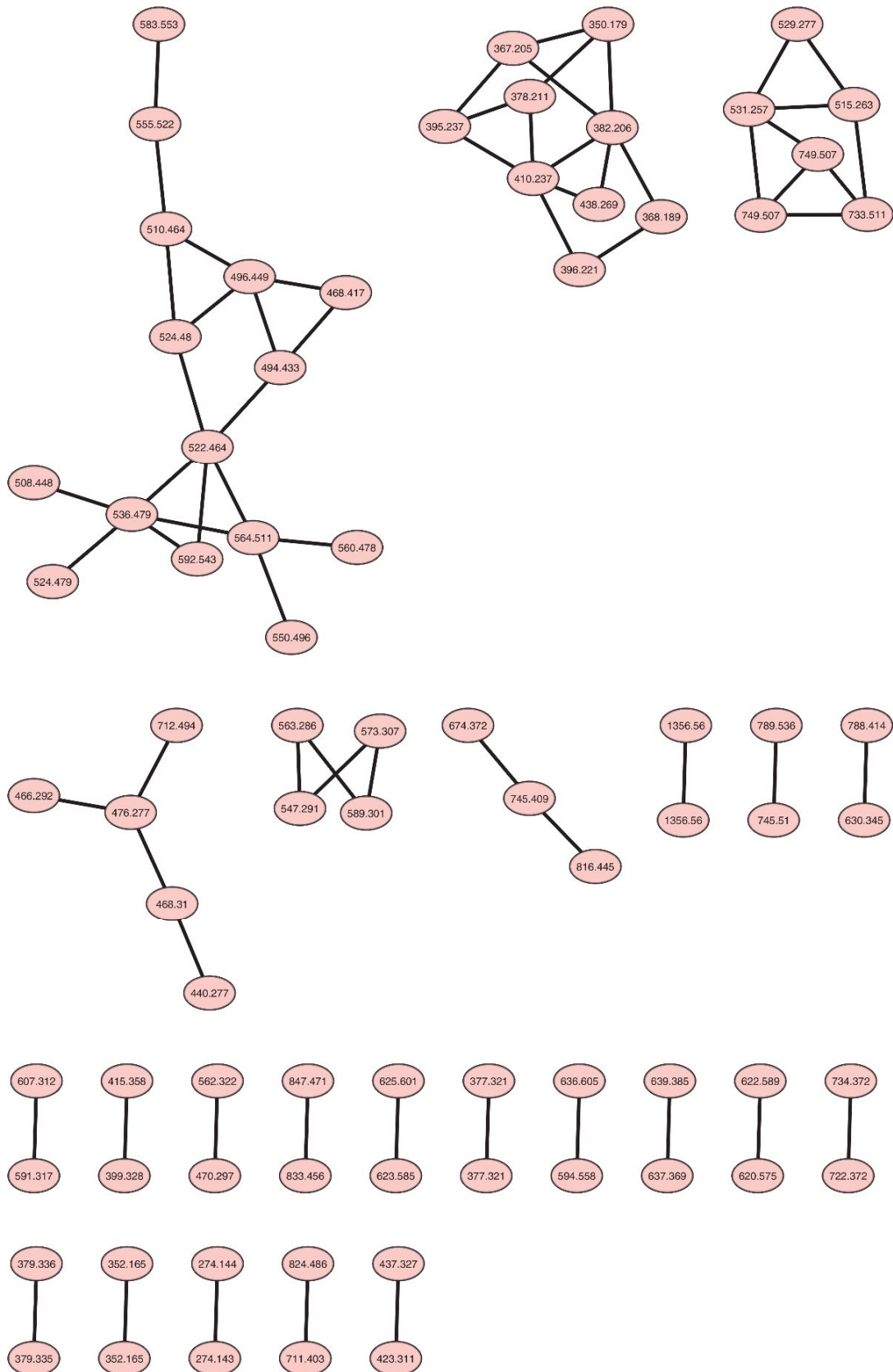
Figure S 21. *GNPS based MS² clustering of the SPL directed precursor selection MS² data visualized with Cytoscape*

Scheduled precursor list based MS² methods to specifically target the bacterium's secondary metabolome has multiple advantages. On the one hand, all nodes displayed in Figure S 21 are linked to bacterial metabolic activity of the bacterium significantly simplifying the distinction between bacterial metabolome and medium blank derived MS signals in the MS and MS² spectra. On the other hand, as less time is lost acquiring MS² spectra of medium derived molecular features we have a significantly greater coverage of low intensity bacterium derived molecular features as exemplified by the discovery of the fulvuthiacenes' biological precursors prefulvuthiacene A and B as well as desmethyl-fulvuthiacene A and B and the minor derivative fulvuthiacene C in the spectral network in the top middle of Figure S 21. It is worth noting that the cluster on the top left side includes spectra providing from the homospermidine lipid compound family as exemplified by their member Cmp-522 in the middle of the spectral network.[11] As this SPL based MS² fragmentation clustering method by design only acquires MS² spectra of compounds belonging to the bacterial metabolome and the corresponding MS² spectra show peptidic fragments including hydroxyvalins one could assume the cluster on the top right hand side around m/z of 749.507 and the cluster in the middle around m/z of 745.409 to be somewhat related to myxoprincomide biosynthesis but this remains to be investigated. [12] Still, as they show irregular peptidic fragmentation they seem to provide from non-ribosomal peptide biosynthesis. I would assume the two remaining spectral networks comprising more than two nodes to be also related to secondary metabolism although – judging from the tandem MS spectra – they don't seem to be peptides. The cluster comprising the m/z of 847.471 and 833.456 also seems to comprise non-ribosomal peptides as judged by its fragmentation pattern. A key issue in using the presented technique is that it efficiently reduces the number of molecular features to investigate manually from around 2000 to a handful of derivatives, which subsequently have to be mined for the most promising candidates based on like uncommon heteroatoms (as the sulfur atom in the fulvuthiacenes) interesting fragmentation patterns or biological activity.

S 3.4   Isolation and structure elucidation

S 3.4.1.1   Isolation and structure elucidation of the fulvuthiacenes

The *Myxococcus fulvus* strain MCy9280 is fermented in 50 ml VYGS medium as a seed culture flasks on an Orbiton shaker at 160 rpm and 30°C. The culture medium becomes translucid and red *M. fulvus* biofilm balls are visible after 5 to 7 days of fermentation. This preculture is used to inoculate 6 x 2L VYGS medium supplemented with 2 % XAD-16 resin suspension in sterilized water in 6 x 5L baffled shake flasks on an Orbiton shaker at 160 rpm and 30°C. Fermentation is complete after 14 days. Cells and XAD-16 resin are harvested by centrifugation on a Beckmann Avanti J-26 XP with the JLA 8.1 rotor at 6000 rcf. Combined

resin and cells are extracted with 2 x 500 ml of technical grade methanol (Fluka) and 2 x 500 ml of technical grade acetone (Fluka). The extracts are combined and all solvent is evaporated on a rotary evaporator. The crude extract from this 10 L fermentation in VYGS medium is separated on the Waters AutoPurification system preparative HPLC system. Purification was carried out on a Waters Autopurifier (Eschborn, Germany) high pressure gradient system, equipped with 2545 binary gradient module, SFO system fluidics organizer, 2767 sample manager and a 2998 photodiode array detector coupled to a 3100 single quadrupole mass spectrometer operated in positive ion mode. Source and voltage settings for the MS were as follows: mass range, m/z 300 - 1000; scan duration, 1 s; points per Dalton, 4; capillary voltage, 3.5 kV; cone voltage, 30 V; extractor voltage 3 V; RF lens, 0.1 V; source temperature 120 C, desolvation temperature, 250 C; desolvation gas flow, 400 L/hr; cone gas flow, 50 L/hr; ion counting threshold, 30. Autopurification is done by setting the fraction collection trigger masses 317.2 and 354.3 which correspond to the $[M-2MeOH]^+$ ions of the fulvuthiacenes. Separation was carried out on a Waters X-Bridge prep C-18 5 µm, 150 x 19 mm column using MeOH + 0.1% FA as B and H_2O + 0.1% FA as A and a flow rate of 25 ml/min. Separation is started with a plateau at 95% A for 2 minutes followed by a ramp to 72% A during 2 minutes and a ramp to 5% A during 29 minutes. The A content is kept at 5% A for 1.5 minutes. The A content is ramped back to starting conditions during 30 seconds and the column is re equilibrated for 3 minutes. This separation step provides two fractions, one enriched in fulvuthiacene A and one enriched in fulvuthiacene B. Further purification is done using a Dionex Ultimate 3000 SDLC low pressure gradient system on a Phenomenex Polar RP 250x10mm 4µm column with the eluents H_2O + 0.1% FA as A and ACN + 0.1% FA as B, a flow rate of 5 ml/min and a column thermostated at 30°C. fulvuthiacene A and B can be detected by UV absorption at 210 nm and 260 nm. Fulvuthiacene A is further purified with the gradient that starts with a plateau at 90 % A for 1 minute followed by a ramp to 70 % A during 1 minute. Then B content is ramped to 10 % A during 25 minutes and finally ramped to 95 % A during 30 seconds. A content is kept stable at 95 % for 1 minute and then ramped back to 90% A during 30 seconds. The column is reequilibrated at 90 % A for 3 minutes. Fulvuthiacene B is further purified with the gradient that starts with a plateau at 90 % A for 1 minute followed by a ramp to 58 % A during 1 minute. Then A content is ramped to 20 % during 25 minutes and finally ramped to 95 % A during 30 seconds. A content is kept stable at 95 % for 1 minute and then ramped back to 90 % A during 30 seconds. The column is reequilibrated at 90 % A for 3 minutes. The pure compounds are subsequently dried under nitrogen gas flow. Fulvuthiacene A and B are obtained as white solid.

S 3.4.1.2   Isolation and structure verification of Myxothiazol Z and Desmethyl-Myxothiazol Z

Myxothiazol Z as well as desmethyl-myxothiazol Z are isolated from the mutant strain *Stigmatella aurantiaca* DW4/3-1 attB :: attP pFPVan FtaLM tn5 FtaGH created in this study. The *S. aurantiaca* strain is fermented in 50 ml TS-6 medium as a seed culture flasks on an Orbiton shaker at 160 rpm and 30°C. After the culture medium reached OD of 0.8 after 4 to 5 days of fermentation it is used to inoculate 6 x 2L TS-6 medium supplemented with 2 % XAD-16 resin suspension in sterilized water in 6 x 5L baffled shake flasks on an Orbiton shaker at 160 rpm and 30°C. Fermentation is complete after 12 days. Cells and XAD-16 resin are harvested by centrifugation on a Beckmann Avanti J-26 XP with the JLA 8.1 rotor at 6000 rcf. Combined resin and cells are extracted with 2 x 500 ml of technical grade methanol (Fluka) and 2 x 500 ml of technical grade acetone (Fluka). The extracts are combined and all solvent is evaporated on a rotary evaporator. The crude extract from is partitioned in a liquid-liquid extraction step between chloroform and saturated brine water. Both myxothiazol derivatives end up quantitatively in the chloroform layer. Purification of desmethyl-myxothiazol Z was carried out using a Dionex Ultimate 3000 SDLC low pressure gradient system on a Phenomenex Luna C18 250x10mm 5µm column with the eluents H2O + 0.1% FA as A and ACN + 0.1% FA as B, a flow rate of 5 ml/min and a column thermostated at 30°C. Desmethyl-myxothiazol Z is detected by UV absorption at 315 nm and purification is done by time dependent fraction collection. Separation is started with a plateau at 70% A for 3 minutes followed by a ramp to 45% A during 2 minutes and a ramp to 5% A during 20 minutes. The A content is kept at 5% A for 2 minutes. The A content is ramped back to starting conditions during 30 seconds and the column is re equilibrated for 4.5 minutes. After evaporation, desmethyl-myxothiazol Z is obtained as an orange-red amorphous solid.

Purification of Myxothiazol Z was carried out on a Waters preparative SFC system, equipped with a Thar fluid delivery module, 2767 sample manager, an ABPR 20, a GLS system featuring a 515 GLS pump, a column oven thermostated to 45°C and a 2998 photodiode array detector coupled to a Waters Acquity QDA mass analyzer operated with a 515 Make up pump. Autopurification is done by setting the fraction collection trigger mass to 503.20 corresponding to myxothiazol Z [M+H]+. QDA probe temperature is set to 600°C, Cone voltage is set to 25 V while spray voltage was set to 1.75 kV. Separation was carried out on a Phenomenex Luna Silica 250x10mm 5µm column using MeOH as co solvent and a flow rate of 25 ml/min. GLS solvent was Methanol and make up solvent 80/20 Methanol/Water + 0.2 % formic acid. Separation started by keeping 5% Cosolvent for 1 minute followed by a co solvent ramp to 15% during 5 minutes. Cosolvent concentration is subsequently ramped to 50% during 1 minute and kept constant during 1 minute. Cosolvent concentration is ramped back to initial conditions during 30s and the column is reequilibrated for 1 minute prior to the next injection. After evaporation, myxothiazol Z is obtained as an orange-red amorphous solid.

S 3.4.2 Elucidation of the fulvuthiacenes' absolute stereochemistry

S 3.4.2.1 Synthesis of *L*-Phenylalanin-O-methylester

We dissolved 10g of *L*-phenlyalanin in 30 ml of methanol and add 5ml on concentrated hydrochloric acid. The reaction mixture is stirred under reflux conditions overnight in a 250 ml round bottom flask and subsequently dried on a rotary evaporator.

S 3.4.2.2 The derivatization and analysis method

In order to elucidate the absolute stereochemistry of fulvuthiacene A and B the compounds as well as myxothiazol A are ozonolysed and the ozonolysis products are coupled twice to *L*-phenylalanin-O-methylester. The created product possesses two stereo centers that are forcibly L-configured stemming from the phenylalanine moieties and therefore all different stereoisomers of the two stereo centers in the fulvuthiacene structure will forcibly generate diastereomers. It will thus be possible to separate them via LC/MS and confirm the equivalence of the stereochemistry of myxothiazol and the stereochemistry of fulvuthiacene.[13] There has already been proof that ozonolysis with oxidative workup works but the used method is not applicable here as the overall isolated yield of fulvuthiacenes is extremely low. To perform the ozonolysis reactions, 0.2 mg of fulvuthiacene A, 0.2 mg of fulvuthiacene B as well as 0.3 mg of myxothiazol A are dissolved in 20 ml of methanol and brought to -77°C. Ozone is passed through the solution until the methanol turns slightly ozone blue. The ozone is eliminated from the solution by flushing it with air and the mixture is mixed with 2 ml of 30% aqueous hydrogen peroxide solution(Sigma Aldrich, contains inhibitor). The mixture is gently stirred at 30°C for 3 hours and subsequently dried completely on a rotary evaporator. After drying, the mixture is taken up in 1 ml of DMF (Sigma Aldrich >99%). 10 equivalents of *L*-Phenylalanin-O-methylester and subsequently 10 equivalents of HATU coupling reagent (carbolution chemicals, Germany) are added (section S 2.4.1). After adding of two droplets of DIPEA (Sigma Aldrich >99%) the mixture is stirred in a 4 ml glass vial overnight. After 24 hours the mixture is extracted with 1ml ethyl acetate and the organic layer is taken for LC/MS analysis. The UHPLC gradient was elaborated to separate the possible stereoisomers as follows. All measurements were performed on a Dionex (Germering, Germany) Ultimate 3000 RSLC system using a Waters (Eschborn, Germany) BEH C_{18} column (50 x 2.1 mm, 1.7 µm) equipped with a Waters VanGuard BEH C_{18} 1.7 µm guard column. Separation of 1 µl sample was achieved by a linear gradient from (A) H_2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 600 µL/min and a column temperature of 45 °C. Gradient conditions were as follows: 0 – 1 min, 5% B; 1 – 2 min, 5 – 10% B; 2 – 16 min, 35% B; 16 - 21 min, 55% B; 21 - 22 min, 80% B; 22 – 26 min, 80% B; 26 – 26.5 min, 80-5% B; 26.5 – 32 min, 5% B. UV spectra were recorded by a DAD in the range from 200 to

600 nm. The LC flow was split to 75 µL/min before entering the Bruker Daltonics maXis 4G hrToF mass spectrometer (Bremen, Germany) using the Apollo II ESI source. MS data acquisition parameters are kept the same as in the acquisition of crude extract LC-MS chromatograms. LC/MSMS data to verify the presence of the correct species is acquired with the same gradient but MS/MS with automatic precursor selection. MS/MS parameters are kept as in the acquisition of automatic precursor selection MS² data.

S 3.4.2.3   Derivatization analysis results

All four theoretical stereoisomers of the fulvuthiacenes would lead to synthesis of an entity of a mass of 484.221 Daltons according to main text Scheme 1. We could clearly detect the reaction product with the correct exact mass for both derivatization reactions that occur at the same retention time. Therefore, the structural element excised from myxothiazol and fulvuthiacene by ozonolysis and derivatized as described in Scheme 1, has consequently the same configuration in both molecules. Spiking experiments of the two reaction mixtures and subsequent gradient variation did not lead to any peak splitting of the main product peak. Therefore, as the stereochemistry of the *L*-Phenylalanin-O-methylester is fixed, the reaction product stemming from **1** and **8** have the same configuration. As (since fulvuthiacenes are produced in very low amounts) the resulting entity could not be analyzed via NMR or in a polarimeter one had to prove the observed mass to actually represent the expected entity with a m/z value of 485.228 [M+H]⁺. Therefore the modified Marfey's analysis method was run to acquire tandem MS spectra that clearly showed the observed entity to be the derivatized ozonolysis product (main text Figure 5).

S 3.4.2.4   NMR based structure elucidation of the fulvuthiacenes

As the two methyl groups at the tips of the fatty acyl side chain of the structure solely couple to C-3 they form a clean doublet. The – for saturated carbons – unusually high carbon shift of C-4 further validates the iso-structure of the fatty acyl side chain in both cases for fulvuthiacene A and B. The rest of the fatty acyl chain is characterized by proton shifts around 2 ppm and carbon shifts of around 30 ppm as one would expect for these carbons. The thiazol heterocycle connected to the fatty acyl side chain is characterized by a very downfield singulett proton at 7 ppm on a downfield shifted carbon at around 120 ppm which shows HMBC coupling to the two quarternary aromatic carbons of the thiazole at 170 and 150 ppm. The right hand side of the molecule can be solved in good agreement to myxothiazol A and myxothiazol Z. [13,14]

S 3.4.2.5   NMR data of the fulvuthiacenes and recombinantly produced myxothiazols
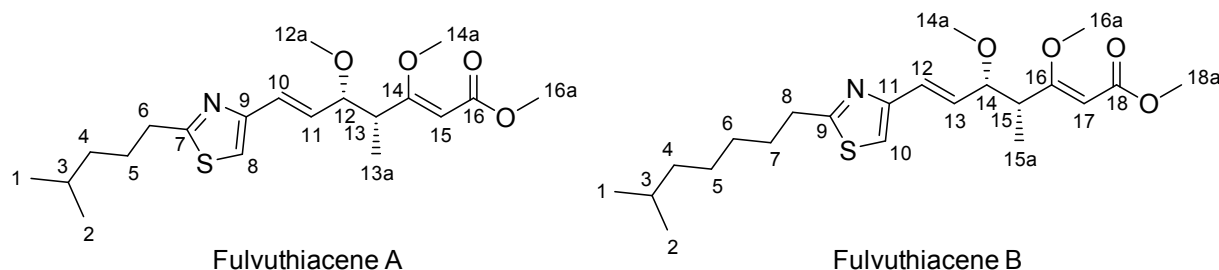
Figure S 22. Structures and carbon numbers of the isolated compounds named fulvuthiacene A and fulvuthiacene B

As both fulvuthiacene derivatives are rather hydrophobic and show low solubility in methanol, 2D NMR data used for structure elucidation is acquired in DMSO-d_6 and CDCl$_3$ on a Bruker Ascend 700 spectrometer with a 5 mm TXI cryoprobe (^{1}H at 700 MHz, ^{13}C at 175 MHz) NMR spectrometer. To elucidate the planar structure of the molecules ^{1}H, ^{13}C as well as H-H DQFCOSY, HSQC and HMBC data are acquired. DQFCOSY, HSQC, HMBC, and ROESY experiments were recorded using standard pulse programs. The NMR signals are grouped in the tables below and correspond to the numbering in the scheme above.

Table S 6. NMR Signals of fulvuthiacene A in DMSO-d_6

Carbon number	δ ^{1}H [ppm]	Multiplicity, J [Hz] and proton number	δ ^{13}C [ppm]	COSY correlations [ppm]	HMBC correlations [ppm]
1+2	0.895	d, 6.5, 6H	22.09	1.559	26.89, 37.7
3	1.559	d t, 6.8, 1H	26.89	0.895, 1.236	22.09, 37.04, 26.89
4	1.236	m, 2H	37.70	1.689, 0.895	26.86, 26.81, 32.7
5	1.689	t t, 7.5, 8.1, 2H	26.86	1.236, 2.910	26.89, 32.7, 37.4, 170.2
6	2.910	t, 7.5, 2H	32.70	1.689	26.86, 37.4, 170.2
7	-	-	170.20	-	-
8	7.320	s, 1H	115.40		151.5, 170.2
9	-	-	151.50	-	-
10	6.468	d, 15.5, 1H	125.80	6.140	83.2, 115.4, 129.6, 151.5
11	6.140	dd, 15.5 8.0, 1H	129.60	3.68, 6.442	38.8 , 83.3, 115.4, 126.2, 151.5
12	3.680	dd, 8.5 7.8, 1H	83.40	6.14, 4.07	38.8, 126.2, 129.6, 56, 175.6
12a	3.200	s, 3H	56.00	-	83.40
13	4.074	d q, 6.9 8.7, 1H	38.80	1.118, 3.680	83.3, 90.6, 129.5, 175.6

13a	1.118	d, 6.9, 2H	14.05	4.074	38.8, 83.4, 175.66
14	-	-	175.60	-	-
14a	3.500	s, 3H	55.96	-	166.50, 167.50
15	4.999	s, 3H	90.60		38.8, 166.3, 175.6
16	-	-	166.30	-	-
16a	3.500	s, 3H	50.15	-	166.50, 167.50

Table S 7. *NMR Signals of fulvuthiacene B in CDCl$_3$*

Carbon number	δ ^{1}H [ppm]	Multiplicity, J [Hz] and proton number	δ ^{13}C [ppm]	COSY correlations [ppm]	HMBC correlations [ppm]
1+2	0.845	d, 6.5, 6H	22.98	1.502, 1.155	28.2, 39.2
3	1.502	m, 1H	28.20	0.845, 1.155	0.845, 1.155
4	1.155	m, 2H	39.26	1.502, 1.315	22.7, 27.4, 28.8
5	1.315	m, 2H	27.36	1.155, 1.370	29.50
6	1.370	m, 2H	29.50	1.315, 1.761	27.4, 30.3, 34.0, 39.3
7	1.761	t t, 7.4 7.9, 2H	30.31	1.377, 2.959	27.36, 29.5, 34.0, 171.1
8	2.959	t, 7.9, 2H	34.00	1.761	171.1, 29.5
9	-	-	171.10	-	-
10	6.915	s, 1H	113.70		126.4, 153.2, 171.1
11	-	-	153.20	-	-
12	6.495	d, 15.6, 1H	126.40	6.274, 3.762	153.2, 131.0, 113.7, 84.3
13	6.274	d d, 15.6 8.0, 1H	131.80	3.726, 6.459	40.00, 84.30, 153.2, 113.7
14	3.762	m, 1H	84.90	6.274, 6.495, 6.274	126.4, 131.8, 176.4, 57.3, 40.0, 14.35
14a	3.280	s, 1H	57.30	-	84.90
15	4.116	d q, 6.9 7.8, 1H	40.00		131.8, 176.4, 57.3, 91.5
15a	1.182	d, 6.9, 2H	14.35	4.126	176.8, 84.9, 40.0
16	-	-	176.40	-	-
16a	3.569	s, 3H	55.90	-	176.40
17	4.931	s, 3H	91.50	-	40.0, 167.6, 176.4
18	-	-	167.60	-	-
18a	3.642	s, 3H	51.00	-	167.50
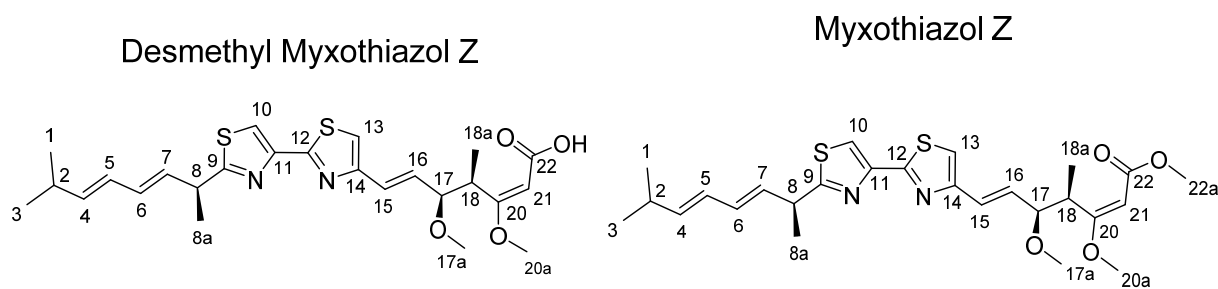
Figure S 23. *Structures and carbon numbers of the Myxothiazol derivatives produced in the* S. aurantiaca *attB :: attP pFPVan FtaMN tn5 GH Mx8 strain*

1D and 2D NMR data used for structure elucidation of the myxothiazol derivatives is acquired in Methanol-d_4 and CDCl$_3$ on a Bruker Ascend 700 spectrometer with a 5 mm TXI cryoprobe (^{1}H at 700 MHz, ^{13}C at 175 MHz) NMR spectrometer. To elucidate the planar structure of the molecules ^{1}H, ^{13}C as well as H-H DQFCOSY, HSQC and HMBC data are acquired. DQFCOSY, HSQC, HMBC, and ROESY experiments were recorded using standard pulse programs. The NMR signals are grouped in the tables below and correspond to the numbering in the scheme above. NMR data fits well to the data for myxothiazol A, desmethyl-myxothiazol Z and myxothiazol Z in the original publications. [13,14]

Table S 8. *NMR Signals of desmethyl-myxothiazol Z in CDCl$_3$*

Carbon number	δ ^{1}H [ppm]	Multiplicity, J [Hz] and proton number	δ ^{13}C [ppm]	COSY correlations [ppm]	HMBC correlations [ppm]
1+3	1.000	d, 7.10, 6H	22.20	2.330	30.9, 141.8
2	2.330	d sp, 7.10 ,1H	30.90	1.00, 5.68	22.20, 126.4, 142.3,
4	5.680	d d, 7.1 14.5, 1H	143.30	2.33, 6.01	22.20, 30.9, 131.7
5	6.010	d d, 10.5 15.3, 1H	126.40	5.68, 6.01	22.20, 30.9, 126.4, 132.2
6	6.170	d d, 10.5 15.0, 1H	131.00	5.78, 6.01	41.0, 126.7,132.2, 142.3, 176.1
7	5.780	d d, 7.5 15.0, 1H	132.20	6.17, 3.93	41.0, 20.5, 126.5, 131.7, 142.3, 176.1
8	3.930	d q, 7.4 7.5, 1H	41.00	1.53, 5.78	20.5, 126.4, 132.2, 143.3, 148.4, 176.1
8a	1.560	d, 7.4, 3H	20.50	3.930	41.0, 132.2, 176.1
9	-	-	176.10	-	-
10	7.810	s, 1H	115.72	-	148.4, 176.6
11	-	-	148.40	-	-
12	-	-	162.40	-	-
13	7.100	s, 1H	115.40	-	125.7, 148.4, 154.10, 162.4
14	-	-	154.10	-	-
15	6.580	d, 15.8, 1H	125.70	6.380	39.90, 83.30,115.40,131.40,154.10
16	6.380	d d, 8.1 15.8, 1H	131.10	3.81, 6.58	39.90, 83.30, 115.40,125.70, 154.10
17	3.810	dd, 8.0 8.9, 1H	84.30	4.15,6.38	14.3, 39.90, 56.70, 125.7, 131.1, 154.10, 178.2

17a	3.330	s, 3H	56.70	-	84.30
18	4.150	dq, 7.0 8.0, 1H	39.90	1.22, 3.81	14.3, 56.7, 84.3, 90.90, 178.2
18a	1.220	d, 7.0, 3H	14.03	4.150	39.9, 84.30, 178.20
20	-	-	178.200	-	-
20a	3.620	s, 3H	55.60	-	90.9, 171.2, 178.2
21	4.980	s, 1H	90.90	-	39.9, 171.2, 178.2
22	-	-	171.20	-	-

Table S 9. *NMR Signals of myxothiazol Z in Methanol-d$_4$*

Carbon number	δ ^{1}H [ppm]	Multiplicity, J [Hz] and proton number	δ ^{13}C [ppm]	COSY correlations [ppm]	HMBC correlations [ppm]
1+3	1.009	d, 6.74, 6H	22.61	2.340	32.2, 143.0
2	2.340	d sp,7 7.52, 1H	32.20	1.009, 5.68	22.6, 143.3, 127.7
4	5.680	d d,7 15.5, 1H	143.00	2.34, 6.03	22.6, 32.2, 133.1
5	6.030	d d,10.4 15.5, 1H	127.70	5.68, 6.21	32.2, 133.1, 133.3
6	6.210	d d, 10.4 15.3, 1H	133.10	5.78, 6.03	41.8, 127.7, 133.3, 143.3, 177.8
7	5.780	d d, 7.8, 15.3 1H	133.30	3.92, 6.21	21.0, 41.8, 127.7, 143.3, 170.2, 177.8
8	3.920	d q, 6.9 7.8, 1H	41.80	1.52, 5.78	21.0, 133.1, 133.3, 177.8
8a	1.520	d, 7.0,1H	21.00	3.920	41.8, 132.3, 177.8
9	-	-	177.800	-	-
10	8.000	s, 1H	116.90	-	41.8, 149.8, 164.2, 177.8
11	-	-	149.80	-	-
12	-	-	164.20	-	-
13	7.351	s, 1H	117.10	-	126.8, 164.2, 155.3
14	-	-	155.30	-	-
15	6.580	d, 15.4 1H	126.80	6.340	85.6, 117.1, 132.0, 155.3
16	6.340	d d, 8.6 15.4, 1H	132.00	6.58, 3.81	41.3, 85.6, 126.8, 155.3
17	3.805	d d, 8.6, 1H	85.60	4.18, 6.34	14.8, 41.3, 57.0, 126.8, 132, 178.4
17a	3.320	s, 3H	57.00	-	85.60
18	4.183	m, 1H	41.30	1.21, 3.81	14.8, 57.0, 85.6, 178.4
18a	1.210	d, 6.8, 3H	14.80	4.180	41.3, 85.6,178.4
20	-		178.400	-	-
20a	3.610	s, 3H	56.80	-	91.8, 178.4
21	5.040	s, 1H	91.80	-	41.3, 169.5,178.4
22	-	-	169.50	-	-
22a	3.610	s, 3H	51.30	-	91.8, 169.5

S 3.5   References

[1]   K. Blin, H. U. Kim, M. H. Medema, T. Weber, *Briefings in bioinformatics* **2017**.

[2]   F. Panter, D. Krug, S. Baumann, R. Müller, *Chem. Sci.* **2018**, *9*, 4898.

[3]   V. Magrini, C. Creighton, P. Youderian, *J. Bacteriol.* **1999**, *181*, 4050.

[4]   B. Silakowski, H. U. Schairer, H. Ehret, B. Kunze, S. Weinig, G. Nordsiek, P. Brandt, H. Blöcker, G. Höfle, S. Beyer et al., *J. Biol. Chem.* **1999**, *274*, 37391.

[5]   R. C. Edgar, *Bmc Bioinform.* **2004**, *5*, 113.

[6]   A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales et al., *Nucleic acids research* **2017**, *45*, D200-D203.

[7]   A. Kitsche, M. Kalesse, *ChemBioChem* **2013**, *14*, 851.

[8]   van den Berg,R. A., H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, van der Werf,M. J., *BMC Gen.* **2006**, *7*, 142.

[9]   T. Hoffmann, D. Krug, S. Hüttel, R. Müller, *Anal. Chem.* **2014**, *86*, 10780.

[10] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan et al., *Nat. Biotechnol.* **2016**, *34*, 828.

[11] M. Hoffmann, D. Auerbach, F. Panter, T. Hoffmann, P. C. Dorrestein, R. Müller, *ACS Chem. Biol.* **2018**, *13*, 273.

[12] N. S. Cortina, D. Krug, A. Plaza, O. Revermann, R. Müller, *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 811.

[13] W. Trowitzsch, G. Höfle, W. S. Sheldrick, *Tetrahedron Lett.* **1981**, *22*, 3829.

[14] J. W. Ahn, S. H. Woo, C. O. Lee, K. Y. Cho, B. S. Kim, *J. Nat. Prod.* **1999**, *62*, 495.

Chapter 4

Structure, Total Synthesis and Biosynthesis of Chloromyxamides - Myxobacterial Tetrapeptides Featuring an Uncommon 6-Chloromethyl-5-methoxypipecolic Acid Building Block

Jan Gorges, [a] † Fabian Panter,[b] † Louise Kjaerulff,[b] † Thomas Hoffmann,[b] Uli Kazmaier,*[a] and Rolf Müller*[b]

Affiliations

[a] Institute for Organic Chemistry, Saarland University, P.O. Box 151150, 66123 Saarbrücken, Germany

[b] Department Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmaceutical Biotechnology, Saarland University, Campus E8.1, 66123 Saarbrücken, Germany

† These Authors contributed equally to the manuscript

Contributions and Acknowledgements

Author's effort:

Contributions by others:

4.1 Abstract

Soil-living microbes are an important resource for the discovery of new natural products featuring great structural diversity, reflective of underlying biosynthetic pathways incorporating a wide range of intriguing small molecule building blocks. We report here the full structural elucidation, total synthesis and biosynthesis of chloromyxamides, new tetrapeptides from myxobacteria that display an unprecedented 6-chloromethyl-5-methoxypipecolic acid (CMPA) substructure. Chemical synthesis – including an approach to access the CMPA unit – was pursued to confirm the structure of chloromyxamides and enabled determination of the absolute configuration in the CMPA ring. A model for non-ribosomal assembly of chloromyxamides was devised based on the combined evaluation of biosynthetic gene cluster sequence information and feeding of stable isotope labelled precursors, providing an insight into formation of the various chloromyxamide derivatives and CMPA biogenesis.

4.2 Introduction

Natural products of microbial origin are a structurally diverse and ever increasing group of small molecules often exhibiting potent biological activity. Accordingly, bacterial secondary metabolites (SM) have a strong track record to become valuable lead structures in human disease areas and agriculture.[1] During evolution, bacteria developed complex biosynthetic machinery for SM biosynthesis that involves a wide range of enzymes catalyzing particular biotransformations. One prominent type of such machinery are the non-ribosomal peptide synthetases (NRPS). These multimodular megasynthetases accomplish the formation of peptide bonds independent from ribosomal translation, opening up the possibility to incorporate building blocks beyond the canonical amino acids.[2] Thus, non-ribosomally made peptides (NRP) cover an extended chemical space compared to ribosomally produced peptides. Structural diversity is further broadened by post-assembly modifications through the action of dedicated tailoring enzymes. Mirroring their wide structural variety, NRP fulfil diverse biological functions including defense against competing microorganisms or chemotaxis.[3] Several strategies are commonly employed for the discovery of new NRP from bacteria, including genome-mining, bioactivity-guided isolation and metabolome-mining approaches.[4] All three strategies have proven fruitful with myxobacteria, which are well-established producers of SM whilst being less extensively screened compared to e.g. the actinobacteria.[5] In this study, we applied a "compound first" screening method to uncover chloromyxamides as novel NRP from the rich secondary metabolomes of myxobacteria. Elucidation of their uncommon structures - including a building

block that is to date without precedent in known NRP - prompted confirmation through total synthesis as well as investigation of the underlying biosynthetic route.

4.3   Results and Discussion

The starting point of our present work is defined by the identification of a family of chlorinated peptides from the myxobacterium *Myxococcus sp.* MCy10608 based on high-resolution mass spectrometric data. Isotope pattern analysis indicated mono-chlorination and MS^2 analysis revealed non-standard amino acid fragments, suggesting that these signals belong to a novel chlorinated NRP scaffold (Figure 1, section S 5.2 to 5.3).[6] Production titers for the new molecules from MCy10608 were however unstable, therefore structural characterization was only possible following the identification of alternative producers using our myxobacterial metabolome repository, revealing that *Myxococcus sp.* MCy10608, MCy10615 and *Pyxidicoccus sp.* MCy10649 produce congeners **1** to **5** in varying patterns (Figure 1).
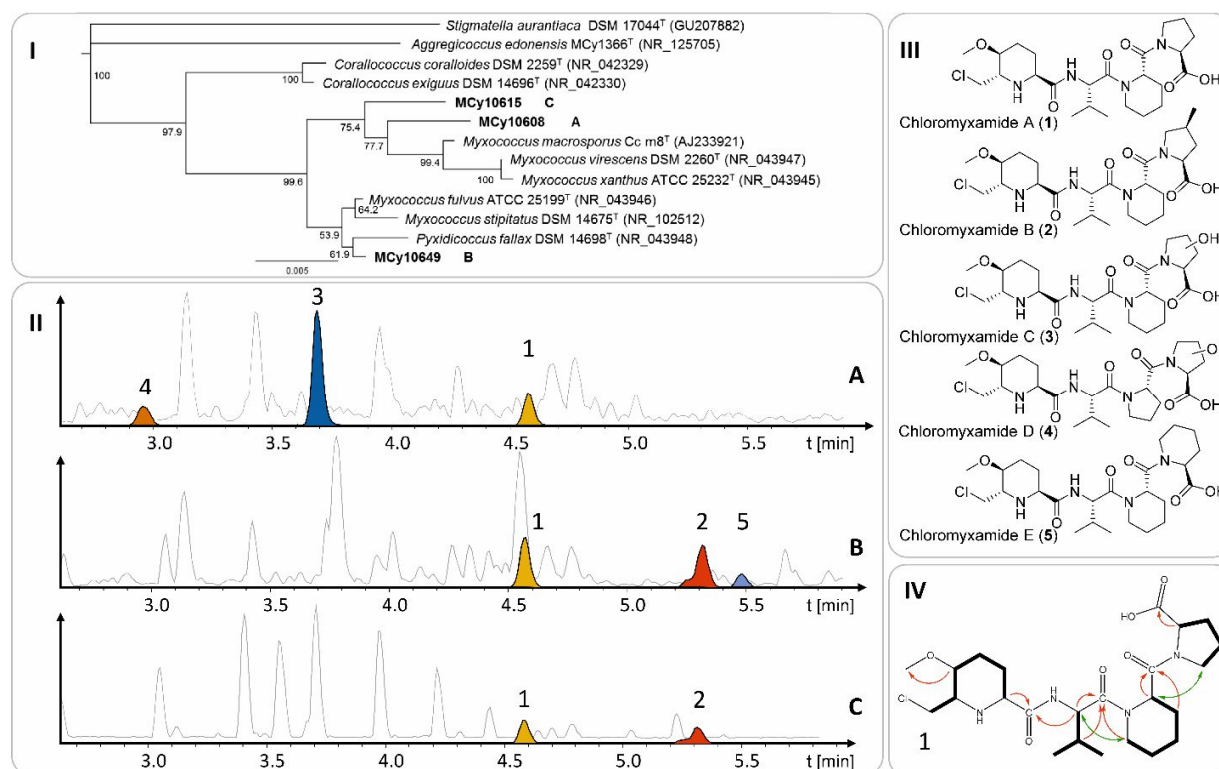
*Figure 1. I) Phylogenetic tree constructed from 16S RNA gene sequence similarity of producer strains A) MCy10608, B) MCy10649 and C) MCy10615 within the order Myxococcales. II) Comparative LC-MS analysis of the three chloromyxamide producers highlights different variants in their extracts (**1** chloromyxamide A, $C_{24}H_{39}N_4O_6Cl$; **2** chloromyxamide B, $C_{25}H_{41}N_4O_6Cl$; **3** chloromyxamide C, $C_{24}H_{39}N_4O_7Cl$; **4** chloromyxamide D, $C_{23}H_{37}N_4O_7Cl$; **5** chloromyxamide E, $C_{25}H_{41}N_4O_6Cl$). III) Structures of chloromyxamides elucidated by NMR and/or tandem MS spectrometry. IV) Key NMR correlations used for the structure elucidation of **1** (COSY correlation as a black dash, HMBC correlations as red arrow and ROESY correlation as green arrow).*

Only strain MCy10615 afforded reliable production at increased scale, permitting the purification of chlorinated peptides named chloromyxamide A (**1**) and B (**2**) with yields sufficient for full structure elucidation. NMR spectroscopy using 1D and 2D homo- and heteronuclear experiments in combination with mass spectrometric analysis showed that these peptides contain the unprecedented chlorinated amino acid 6-chloromethyl-5-methoxypipecolic acid (CMPA). In addition, valine, pipecolic acid, proline, and 4-methylproline moieties were identified in the 2D NMR analyses of **1** and **2**. Marfey's analysis showed that these compounds contained *L*-Val, *L*-Pip and *L*-Pro (in **1**), and *L*-Val, *L*-Pip, and (2*S*,4*S*)-4-MePro (in **2**) (section S 5.5).[7] Tandem-MS analysis confirmed amino acid sequences for all chloromyxamides and provided structural information for family members C (**3**), D (**4**) and E (**5**) not isolated due to low yields in larger scale cultures (Figure 2, section S 5.2). Subsequent feeding studies using stable isotope labelled amino acids proved that all chloromyxamides are synthesized via a common route and shed light on the biogenesis of incorporated non-canonical amino acids (Figure 2, section S 5.3).
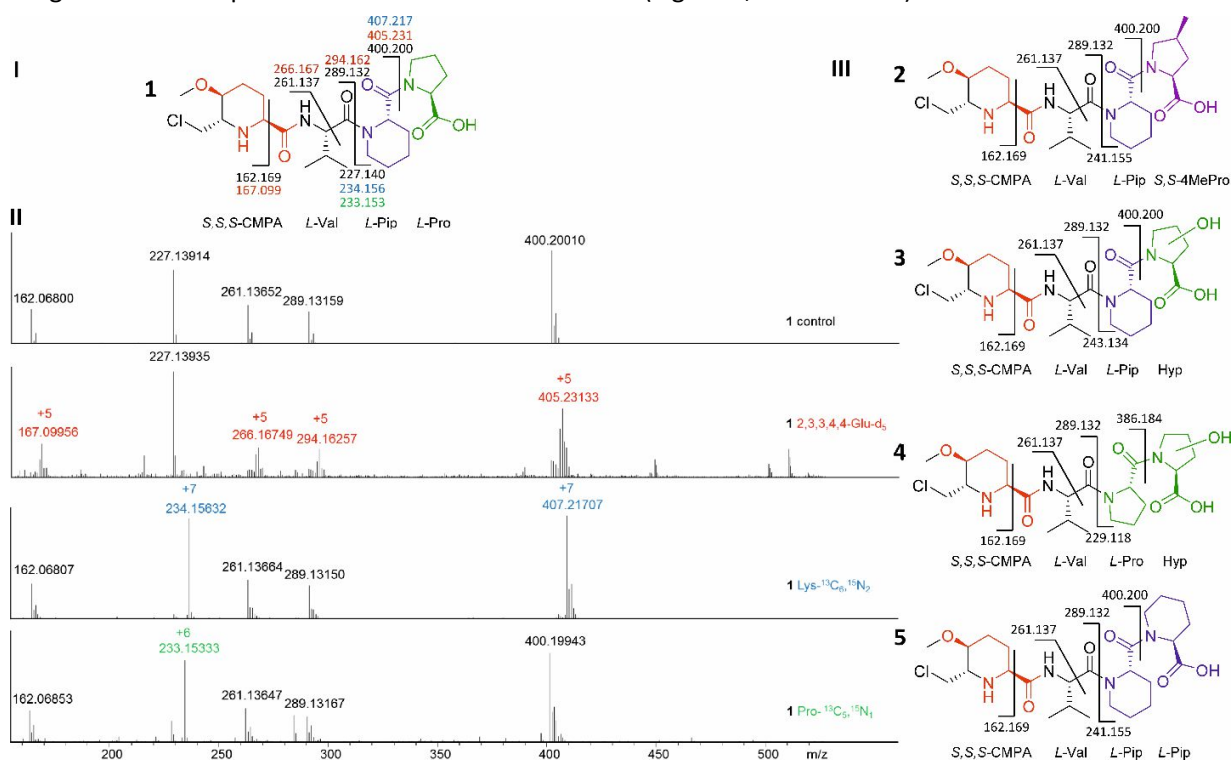
Figure 2. I) Fragmentation of *1* in tandem MS experiments when targeting the native form (black) or after feeding with labelled biosynthetic precursors (red, Glu-2,3,3,4,4-d₅; blue, Lys-¹³C₆,¹⁵N₂; green, Lys-¹³C₅,¹⁵N), II) corresponding tandem MS spectra of 1 after selectively targeting the MS peaks of the native structure (black) and after labelled precursor feeding (red, blue, green) and III) Fragmentation of *2* to *5* in tandem MS experiments when targeting the native form and color coding of the amino acid precursor incorporation (purple, L-Leu-5,5,5-d₃)

Taken together, we conclude that i) biosynthesis of the CMPA substructure found in all chloromyxamides starts with glutamic acid that is extended with a C_2 unit and subsequently cyclized, ii) pipecolic acid units in **1**, **2**, **3** and **5** originate from lysine, iii) hydroxyproline (Hyp) in **3** and **4** stems from

direct oxidation of proline, and iv) biosynthesis of the methyl proline unit in **2** is analogous to the methyl proline pathway described in griselimycin biosynthesis (Figures 2 and 3, section S 5.3 to 5.4).[8] Unfortunately, isolation of **3** and **4** in amounts sufficient for NMR analysis was unsuccessful; we were thus unable to determine whether these chloromyxamide congeners contain 3- or 4-hydroxyproline (section S 5.3 and 5.4). We next sought to elucidate the absolute stereochemistry of the CMPA moiety in chloromyxamides, but could not achieve this by Marfey's method since we lacked standard material. Therefore, we chose to determine its relative configuration via determination of the homonuclear coupling constants in an S^3 HMBC experiment. By analysis of J_{HH} and δ_H and comparison to reference values and coupling constants calculated by molecular modelling (section S 5.7), the relative configuration could be either (2*S*,5*S*,6*S*) or (2*R*,5*R*,6*R*).[9] In order to confirm the structure of the novel chloromyxamide scaffold and to clarify the absolute stereochemistry of the CMPA unit we set out for total synthesis of **1**. The synthesis of the C-terminal tripeptide **9** was straightforward, beginning with protected *L*-proline **6** (Scheme 1).
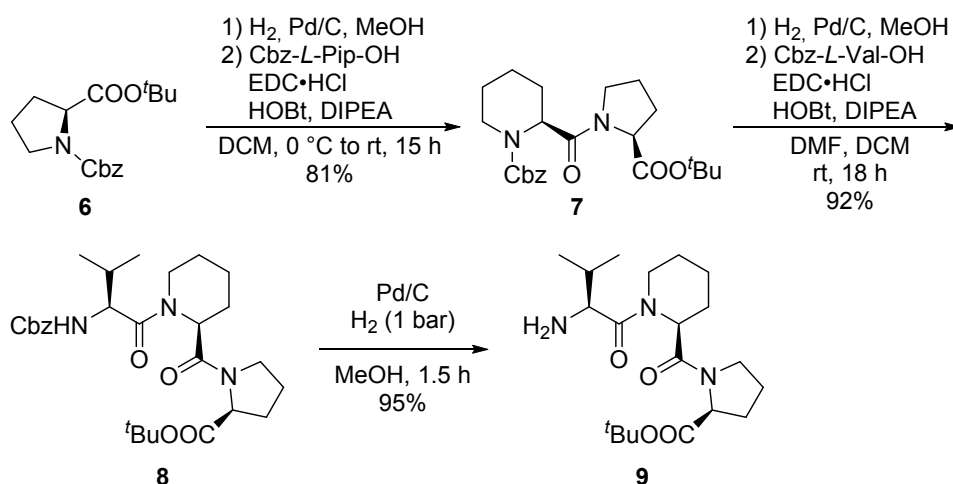
Scheme 1. *Synthesis of the tripeptide building block 9.*

The synthesis of the unusual *S*,*S*,*S*-configured *N*-terminal amino acid started from protected *D*-serine **10** which was first converted into the corresponding Weinreb amide **11** (Scheme 2). Treatment of **11** with Grignard reagent led to the ketone **12** which was stereoselectively reduced to alcohol **13**.[10,11] To determine the configuration of the new formed stereogenic center, **13** was converted into oxazolidinone **14** with sodium hydride. The coupling constant of 7.8 Hz is in agreement with literature values for the *syn*-oxazolidinone.[11] Selective *O*-methylation of **13** failed under several tested conditions (MeI/NaH, MeI/LHDMS, MeOTf/2,6-DTBMP or (MeO)$_3$BF$_4$/ proton sponge).[12] We either observed the formation of oxazolidinone **14**, *N*-methylation or low conversion rates.
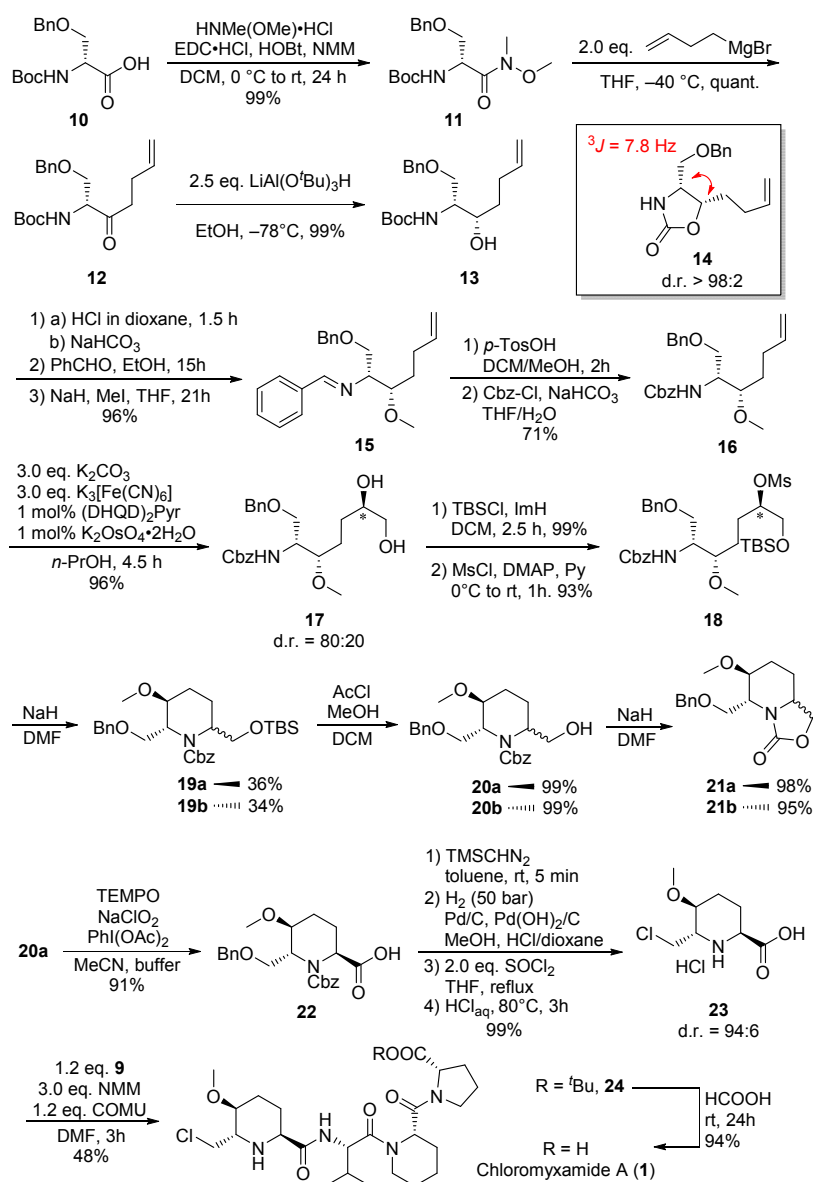
Scheme 2. *Synthesis of chloromyxamide A*

Finally, we achieved selective *O*-methylation by converting **13** into the corresponding benzylidene imine, which was then selectively methylated to **15**, which was converted into Cbz-protected **16**. To optimize the following Sharpless dihydroxylation several ligands and solvents have been tested, the best results delivered **17** in 96% yield and with an d.r. of 80:20. After selective TBS-protection of the primary alcohol, the secondary alcohol was transformed into mesylate **18**. Other leaving groups like tosylates or an epoxide led to unsatisfactory results in the ring closing reaction. To access the piperidine ring, **18** was subjected to sodium hydride, which afforded the two diastereomers **19a** and **19b**. The low diastereoselectivity might be caused as the reaction proceeds via an S_N1-mechanism. Both diastereomers **19a** and **19b** were then desilylated with acetyl chloride in methanol and converted into the corresponding

oxazolidinones **21** as described above. Their configuration was assigned by comparing NOE-DIFF-spectra and the crystal structure of **21a**, which possesses the natural product's configuration. **20a** was subsequently oxidized using a variant of the Epp-Widlanski oxidation to **22**.[13] The initial route towards **1** was designed to introduce the chlorine directly to the tetrapeptide at a late stage. To achieve this, we first coupled the two building blocks **22** and **9** to the linear tetrapeptide, but unfortunately, we were not able to cleave the benzyl ether from the *N*-terminus even under harsh hydrogenation conditions.[14] Therefore, we decided to synthesize CMPA before incorporation into the tetrapeptide. Esterification, catalytic hydrogenation and subsequent treatment with thionyl chloride provided the chlorinated methyl ester, which was cleaved to the free acid **23**. Coupling with **9** using COMU provided **24** which was purified by HPLC (section S 6.2.20). Cleavage of the *tert*-butyl ester with formic acid finally provided chloromyxamide A (**1**). The analytical data were in accordance with the isolated natural product proving the unprecedented CMPA amino acid to be (*S,S,S*)-configured (section S 6.3.2).

Bioinformatic analysis of the genome of *Myxococcus sp.* MCy10608 uncovered a non-ribosomal peptide synthetase (NRPS) biosynthesis gene cluster (BGC) matching the chloromyxamide scaffold (Figure 3, Table S 2).[15] Core functions of the *cmx* NRPS system are encoded by *cmxA-R*, including a three module NRPS assembly line encoded by *cmxB*, complemented by export-related (CmxC) and tailoring (CmxA, CmxD to CmxH) functions and enzymes involved in the biosynthesis of CMPA from glutamic acid (CmxI to CmxR). Since MCy10608 was inaccessible through genetic manipulation despite serious efforts, we devised a biosynthesis hypothesis based on feeding experiments and *in-silico* analysis (Figure 3). Chloromyxamide biosynthesis begins with formation of the CMPA unit: the integral adenylation domain of CmxM adenylates glutamic acid and transfers it to the amino acyl carrier protein CmxD, a homolog of LysW, which is bound to *L*-glutamate during lysine biosynthesis.[16] Next, the acetyl-CoA acetyltransferase-like protein CmxP catalyzes chain elongation of said glutamate through decarboxylative Claisen condensation utilizing malonyl-CoA. The CmxD bound reduced species is likely dehydrated by the enoyl-CoA hydratase protein CmxJ to produce the α,β- unsaturated precursor that is subsequently reduced by one of several dehydrogenase enzyme candidates (Table S 2). The crucial cyclisation step in CMPA biosynthesis is probably a chlorination-induced cyclisation giving rise to the pipecolic acid type backbone structure (Figure 3 panel II). CmxH encodes a matching diheme dioxygenase that might serve as a chloroperoxidase forming the Cl^+ species necessary for electrophilic attack on the remaining double bond. This halogenated cationic intermediate would be prone to intramolecular cyclisation, thus completing the observed six-ring structure of CMPA. Similar halogenation-induced cyclisation reactions are used for six membered ring formation in chemical synthesis.[17] Chlorination induced cyclisation to synthesize non-canonical amino acids has also been described for cyclopropyl ring formation during coronatin biosynthesis, although in

that case the chlorine does not remain on the CMA as it is lost in an S_N2 reaction.[18] The SAM-dependent O-methyl transferase CmxG finally methylates the amino acid's hydroxyl group to form the methoxy group of CMPA. Although we were not yet able to biochemically underpin all steps in the biosynthesis of the unprecedented CMPA unit, incorporation of five deuterium atoms into the chloromyxamides upon feeding with *L*-glutamic acid-d_5 proves the CMPA unit to be synthesized via glutamic acid. CMPA is transferred to the NRPS assembly line encoded on CmxB as CmxD-bound intermediate. The three modules with predicted substrate specificities for Val, Pip and Pro form the three peptide bonds present in all chloromyxamide derivatives. Pip required for the biosynthesis of **1, 2, 3** and **5** is synthesized by the TubZ-homolog CmxA (Figure S 2) as concluded from $^{13}C_6$ $^{15}N_2$ *L*-Lys feeding: loss of one ^{15}N atom is congruent with lysine cyclization.[19] Module 1 of the NRPS assembly line encoded on *cmxB* strictly incorporates valine into the final structure, whereas module 2 alternatively allows the incorporation of proline as seen in **4**.
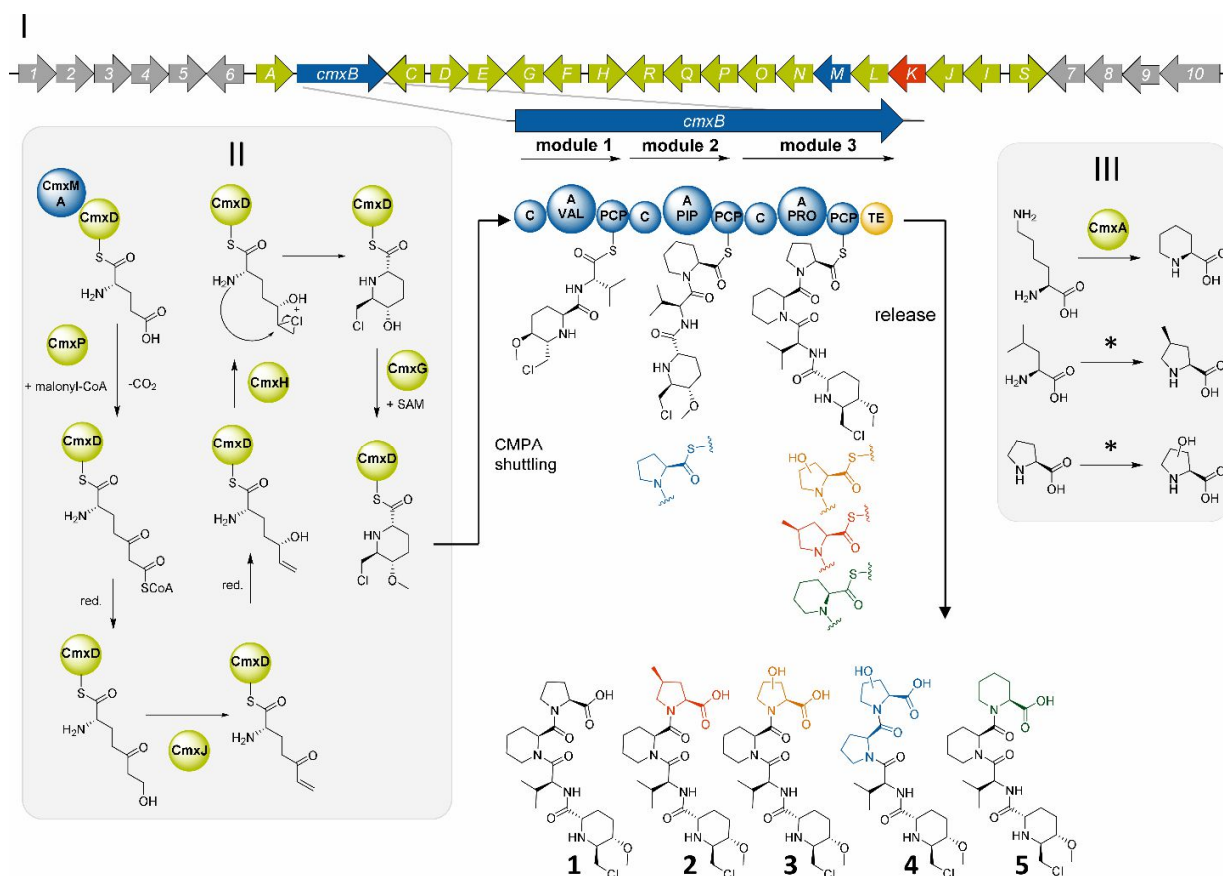
Figure 3. *I) Chloromyxamide BGC (~47000 bps, 28 ORFs) and proposed biosynthesis of **1** to **5** (blue: NRPS genes/domains, green: tailoring genes/enzymes, red: regulatory gene, grey: ORFs of unknown function). II) Proposed formation of the CMPA unit from L-glutamic acid. III) Reactions that lead to supply of non-canonical amino acids used in chloromyxamide biosynthesis; the asterisk marks reactions proven by stable isotope labelling that could not be attributed to a candidate biosynthetic enzyme encoded within the BGC.*

The flexibility of module 3 on CmxB is a noteworthy feature of chloromyxamide biosynthesis due to incorporation of cyclic amino acids proline, methyl proline, hydroxy proline or pipecolic acid, with ratios apparently depending on the specific producer. However, a methyl proline operon as known from griselimycin biosynthesis and a proline hydroxylation pathway for 3- or 4- hydroxyproline is not encoded within the *cmx* BGC.[8,20] These functions are thus likely encoded at distant genomic loci and the necessary building blocks recruited to CmxB as free amino acids. Despite their intriguing biosynthesis, identification of the biological role of chloromyxamides requires further studies as testing a panel of microorganisms and cell lines did not reveal significant biological activities (section S 5.9).

4.4   Conclusion and Outlook

The vast diversity of yet unidentified natural products contained in myxobacterial secondary metabolomes translates into outstanding chances for the discovery of new scaffolds through "compound first" approaches, exemplified in this study by applying a simple yet effective filter (i.e. halogenation and informative MS^2 fragmentation) to highlight NRP compounds likely to represent structural novelty. At the same time, our work underpins the importance of combining analytical, genomics-based and synthetic efforts for natural product discovery, since full structural elucidation of chloromyxamides including the absolute configuration of the unprecedented CMPA unit was achieved with the help of total synthesis. The CMPA building block present in chloromyxamides is peculiar both because of its biogenesis starting from glutamic acid and its intriguing substructure, as amino acids bearing a chlorine atom attached to an sp^3 carbon are rare in nature.

4.5   Acknowledgements

4.6   References

[1]   D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2016**, *79*, 629.

[2]   D. E. Cane, C. T. Walsh, *Chem. Biol.* **1999**, *6*, R319-R325.

[3]   A. L. Demain, A. Fang, *Adv. Biochem. Eng. Biotechnol.* **2000**, *69*, 1.

[4] J. J. Hug, C. D. Bader, M. Remškar, K. Cirnski, R. Müller, *Antibiotics* **2018**, *7*, 44.

[5] a) J. Herrmann, A. A. Fayad, R. Müller, *Nat. Prod. Rep.* **2017**, *34*, 135; b) T. Weber, K. Blin, S. Duddela, D. Krug, H. U. Kim, R. Bruccoleri, S. Y. Lee, M. A. Fischbach, R. Müller, W. Wohlleben et al., *Nucleic Acids Res.* **2015**, *43*, W237-W243; c) T. Hoffmann, D. Krug, S. Hüttel, R. Müller, *Anal. Chem.* **2014**, *86*, 10780; d) N. S. Cortina, D. Krug, A. Plaza, O. Revermann, R. Müller, *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 811.

[6] M. Meusel, F. Hufsky, F. Panter, D. Krug, R. Müller, S. Böcker, *Anal. Chem.* **2016**, *88*, 7556.

[7] C. B'Hymer, M. Montes-Bayon, J. A. Caruso, *J. Sep. Sci.* **2003**, *26*, 7.

[8] P. Lukat, Y. Katsuyama, S. C. Wenzel, T. Binz, C. König, W. Blankenfeldt, M. Brönstrup, R. Müller, *Chem. Sci.* **2017**, *8*, 7521.

[9] L. Kjaerulff, A. J. Benie, C. Hoeck, C. H. Gotfredsen, O. W. Sorensen, *J. Magn. Reson.* **2016**, *263*, 101.

[10] R. C. So, R. Ndonye, D. P. Izmirian, S. K. Richardson, R. L. Guerrera, A. R. Howell, *J. Org. Chem.* **2004**, *69*, 3233.

[11] R. V. Hoffman, N. Maslouh, F. Cervantes-Lee, *J. Org. Chem.* **2002**, *67*, 1045.

[12] a) P. Barbie, U. Kazmaier, *Org. Lett.* **2016**, *18*, 204; b) M. Kretschmer, M. Dieckmann, P. Li, S. Rudolph, D. Herkommer, J. Troendlin, D. Menche, *Chem. Eur. J.* **2013**, *19*, 15993; c) A. B. Smith, T. M. Razler, G. R. Pettit, J.-C. Chapuis, *Org. Lett.* **2005**, *7*, 4403; d) M. J. Zacuto, D. Tomita, Z. Pirzada, F. Xu, *Org. Lett.* **2010**, *12*, 684.

[13] G. Tojo, M. Fernández, *Oxidation of primary alcohols to carboxylic acids*. *A guide to current common practice*, Springer, New York, **2007**.

[14] a) M. T. Crimmins, K. A. Emmitte, *Adv. Ceram. Mater.* **2001**, *123*, 1533; b) M. Kanematsu, M. Shindo, M. Yoshida, K. Shishido, *Synthesis* **2009**, *2009*, 2893.

[15] K. Blin, T. Wolf, M. G. Chevrette, X. Lu, C. J. Schwalen, S. A. Kautsar, H. G. Suarez Duran, E. L. C. de Los Santos, H. U. Kim, M. Nave et al., *Nucleic Acids Res* **2017**, *45*, W36-W41.

[16] S. Fujita, S.-H. Cho, A. Yoshida, F. Hasebe, T. Tomita, T. Kuzuyama, M. Nishiyama, *Biochem. Biophys. Res. Commun.* **2017**, *491*, 409.

[17] S. A. Snyder, D. S. Treitler, *Angew. Chem. Int. Ed.* **2009**, *48*, 7899.

[18] F. H. Vaillancourt, E. Yeh, D. A. Vosburg, S. E. O'Connor, C. T. Walsh, *Nature* **2005**, *436*, 1191.

[19] A. Sandmann, F. Sasse, R. Müller, *Chem. Biol.* **2004**, *11*, 1071.

[20] a) M. A. Culpepper, E. E. Scott, J. Limburg, *Biochemistry* **2010**, *49*, 124; b) J. Mattay, S. Houwaart, W. Hüttel, *Appl. Environ. Microbiol.* **2018**.

Supporting Information

Structure, Total Synthesis and Biosynthesis of Chloromyxamides - Myxobacterial Tetrapeptides Featuring an Uncommon 6-Chloromethyl-5-methoxypipecolic Acid Building Block

Previously published in:

Jan Gorges, [a] † Fabian Panter,[b] † Louise Kjaerulff,[b] † Thomas Hoffmann,[b] Uli Kazmaier,*[a] and Rolf Müller*[b]

Affiliations

[a] Institute for Organic Chemistry, Saarland University, P.O. Box 151150, 66123 Saarbrücken, Germany

[b] Department Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmaceutical Biotechnology, Saarland University, Campus E8.1, 66123 Saarbrücken, Germany

† These Authors contributed equally to the manuscript

The printed version of the Supporting Information does not contain data that cannot be visualized satisfactorily on paper. To access this data please refer to the enclosed storage medium or the on-line version of this research paper.

S 4.1   Taxonomic evaluation of the chloromyxamide producer strains

The myxobacterial strains analyzed in this study belong to the *cystobacterinae* suborder among the myxobacteria. MCy10608 and MCy10615 were isolated by Ram Prasad Awal and MCy10649 was isolated by Ronald Garcia from top soil samples. As the strain MCy10608 has been sequenced by Pac-Bio sequencing (Pacific Biosciences, Menlo Park, CA, USA) while the other two strains (MCy10615 and MCy10649) were characterized by 16s RNA gene similarity (TAQ Polymerase, Thermo Scientific; sequencing at LGC Biosciences) we can investigate their phylogeny more closely.
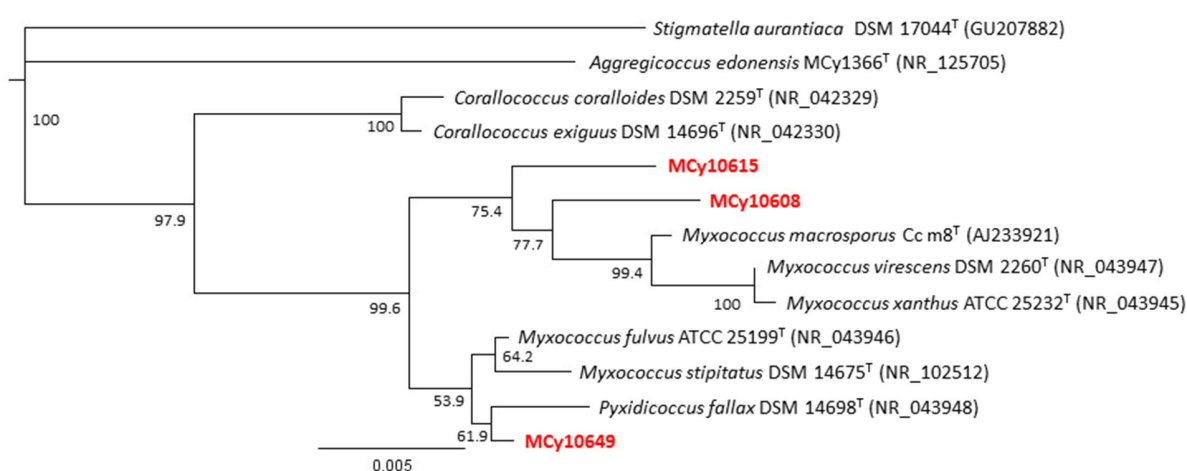
Figure S 1. *Phylogenetic tree of the published myxobacteria belonging to the* Myxococcaceae *family and highlighting of the different chloromyxamide producer strains MCy10608, MCy10615 and MCy1049*

The strains MCy10608 and MCy10615 belong to the *Myxococcus* clade and are related to the type strain for *Myxococcus macrosporus*. Still, sequence homology is too low to characterize MCy10608 and MCy10615 as a *Myxococcus macrosporus* meaning the strains will be referred to as *Myxococcus* species in this work. MCy10649 on the other hand is more closely related to the type strain of *Pyxidicoccus fallax* meaning that this bacterium most likely represents a new species of the *Pyxidicoccu*s genus and will thus be referred to as *Pyxidicoccus species*. Although the three strains MCy10608, MCy10615 and MCy10649 produce the same secondary metabolite (chloromyxamide A, **1**) they are phylogenetically rather far apart. More astoundingly, the two myxobacterial strains MCy10615 and MCy10649 producing **1** and the 4-methylproline derivative chloromyxamide B are actually more distantly related than MCy10615 and MCy10608 which only share the metabolite 1 (Table S 3). The distribution of the different types of chloromyxamide biosynthetic gene clusters is therefore rather widespread across the *Myxococcaceae* family and more closely related gene clusters are actually present in less closely related bacteria. It is thus

likely that this biosynthetic gene cluster has been passed on during evolution via horizontal gene transfer steps rather than divergent evolution.

S 4.2 Myxobacterial growth conditions and analytical setup

S 4.2.1 Myxobacterial culture media

Table S 1. *Medium recipe of VY/2-S medium used for cultivation of all myxobacteria in this study*

VYGS – Medium			
Amount	Ingredient	Concentration	Supplier
5 g/L	Baker's Yeast	-	Lynco
10 g/L	Starch (soluble)	-	Roth
1 g/L	CaCl$_2$	-	Sigma Aldrich
1 g/L	MgSO$_4$ x 7H$_2$O	-	Grüssing
10 mL/L	TRIS pH 8	1M	Sigma Aldrich
100 µL/L	Sterile Vit. B12 solution (add after autoclaving)	1 mg/ml	Roth
200 µL/L	Sterile FeEDTA solution (add after autoclaving)	8 mg/ml	Sigma Aldrich
50 µL/L	Sterile Riboflavin solution (add after autoclaving)	1 mg/ml	Roth
Dissolved in milliq. water pH adjusted to 7.2 with 1N KOH			

For the preparation of agar media, medium was prepared adding 14 g/L agar (BD) to the liquid medium.

S 4.2.2 Small scale myxobacterial fermentations for acquisition of UPLC-HRMS data

Cultures for UHPLC/HRMS analysis are grown in 300 ml Erlenmeyer flasks containing 50 ml of VY/2S medium for *Myxococcus sp.* MCy10608, *Pyxidicoccus sp.* MCy10649 or *Myxococcus sp*. MCy10615 inoculated with 1 ml of pre culture. After inoculation the medium is supplemented with 2% of sterile XAD-16 adsorber resin (Sigma Aldrich) suspension in water. The strains are grown in VYGS medium at 30°C for 12-15 days on a gyratory shaker at 160 rpm. Fermentation is complete once the white VY/2-S medium turned transparent and spherical myxobacterial bio-film clumps are visible in the shake flask.

S 4.2.3 Extraction procedures of myxobacterial fermentation cultures in analytical scale

The frozen cell pellet is transferred into a 100 ml Erlenmeyer flask and a magnetic stirrer is added. 50 ml of acetone (Fluka analytical grade, redistilled in house) are added onto the pellet and the mixture is stirred for 60 min. on a magnetic stirrer. The acetone has to be distilled beforehand since Fluka analytical reagent grade acetone still contains trace amount of polymer impurities that will otherwise concentrate

in the extract. The acetonic extract is left to settle in order to sediment cell debris and XAD resin for a second extraction step. The supernatant is filtered with a 125 micron folded filter keeping cell pellet and XAD-16 resin in the Erlenmeyer flask for a second extraction step. The residual pellet and XAD-16 resin is extracted again with 30 ml of distilled acetone for 60 min on a magnetic stirrer and filtered through the same folded filter. The combined extracts are transferred into a 100 ml round bottom flask. The acetone is evaporated on a rotary evaporator at 260 mbar and 40 °C water bath temperature. The residual water is evaporated at 20 mbar until the residue in the flask is completely dry. The residue is taken up in 550 μl of methanol (Chromasolv HPLC grade, Sigma Aldrich) and transferred into a 1.5 ml Eppendorf tube. This tube is centrifuged with Heareus biofuge pico at 13000 rpm for 3 minutes to remove residual insolubilities such as salts, cell debris and XAD fragments.

S 4.2.4   Standardized UHPLC MS method for screening of myxobacterial crude extracts

All measurements were performed on a Dionex (Germering, Germany) Ultimate 3000 RSLC system using a Waters (Eschborn, Germany) BEH C_{18} column (50 x 2.1 mm, 1.7 μm) equipped with a Waters VanGuard BEH C_{18} 1.7 μm guard column. Separation of 1 μl sample was achieved by a linear gradient from (A) H_2O + 0.1 % FA to (B) ACN + 0.1 % FA at a flow rate of 600 μL/min and a column temperature of 45 °C. Gradient conditions were as follows: 0 – 0.5 min, 5% B; 0.5 – 18.5 min, 5 – 95% B; 18.5 – 20.5 min, 95% B; 20.5 – 21 min, 95 – 5% B; 21-22.5 min, 5% B. UV spectra were recorded by a DAD in the range from 200 to 600 nm. The LC flow was split to 75 μL/min before entering the Bruker Daltonics maXis 4G hr-qToF mass spectrometer (Bremen, Germany) using the Apollo II ESI source. Mass spectra were acquired in centroid mode ranging from 150 – 2500 m/z at a 2 Hz full scan rate. Mass spectrometry source parameters are set to 500 V as end plate offset; 4000 V as capillary voltage; nebulizer gas pressure 1 bar; dry gas flow of 5 l/min and a dry temperature of 200°C. Ion transfer and quadrupole settings are set to Funnel RF 350 Vpp; Multipole RF 400 Vpp as transfer settings and Ion energy of 5 eV as well as a low mass cut of 300 m/z as Quadrupole settings. Collision cell is set to 5.0 eV and pre pulse storage time is set to 5 μs. Spectra acquisition rate is set to 2Hz. Calibration is done automatically before every LC-MS run by injection of sodium formate and calibration on the sodium formate clusters forming in the ESI source. All MS analyses are acquired in the presence of the lock masses $C_{12}H_{19}F_{12}N_3O_6P_3$; $C_{18}H_{19}O_6N_3P_3F_2$ and $C_{24}H_{19}F_{36}N_3O_6P_3$ which generate the [M+H]$^+$ ions of 622.028960; 922.009798 and 1221.990638.

S 4.2.5   UHPLC Coupled SPL-MS^2 method to determine stable isotope labelled amino acid incorporation

All measurements were performed on a Dionex (Germering, Germany) Ultimate 3000 RSLC system using a Waters (Eschborn, Germany) BEH C$_{18}$ column (50 x 2.1 mm, 1.7 µm) equipped with a Waters VanGuard BEH C$_{18}$ 1.7 µm guard column coupled to the Bruker Daltonics maXis 4G hr-qToF mass spectrometer (Bremen, Germany) as described in S 4.2.4. LC and MS conditions of the SPL guided MS/MS data acquisitions are kept constant according to section standardized UHPLC-MS conditions. MS/MS data acquisition parameters are set to exclusively fragment scheduled precursor list entries. Scheduled precursor list entries are edited manually to selectively target the precursor masses of the stable isotope labelled chloromyxamides (section S S 4.5.3). This allows following up the labelled part in the tandem MS spectrum. SPL tolerance parameters for precursor ion selection are set to 0.2 minutes and 0.05 m/z in the SPL MS/MS method. The method picks up to 2 precursors per cycle, applies smart exclusion after 5 spectra and performs CID and MS/MS spectra acquisition time ramping. CID Energy is ramped from 35 eV for 500 m/z to 45 eV for 1000 m/z and 60 eV for 2000 m/z. MS full scan acquisition rate is set to 2Hz and MS/MS spectra acquisition rates are ramped from 1 to 4 Hz for precursor Ion intensities of 10 kcts. to 1000 kcts..

S 4.3   Molecular biology protocols

S 4.3.1   Isolation of genomic DNA for PacBio sequencing

To isolate total DNA for sequencing purposes such as illumina sequencing, phenol-chloroform gDNA extraction is used. *In silico* analysis of the chloromyxamide gene cluster

S 4.4   *In silico* analysis of the *cmx* biosynthetic gene cluster

Identification of the biosynthetic gene cluster was done upon identification of a three modular NRPS gene called *cmxB* with a NRPS predictor 2 hidden markov model prediction of valine, proline and proline, which fits well to the structures observed for the chloromyxamides.[1] Prediction of the cluster borders was done by examination of the flanking regions to the right and to the left of the central *cmxB* gene. While *cmxA* to *cmxS* encode biosynthesis genes, open reading frame (orf) 1 to orf 10 encode genes most likely unrelated to natural products biosynthesis or natural product tailoring and are therefore excluded from the putative chloromyxamide biosynthesis gene cluster.

Table S 2. *Genes of the chloromyxamide biosynthesis gene cluster, predicted gene function and their closest homologue determined by blastp search against NCBI non-redundant protein database (nr)*

Gene	Length [bp]	Closest homologue [Organism of origin]	Identity [%] /length of alignment [AA]	Proposed function	Accession Nr.
cmxA	1101	Ornithine cyclodeaminase [Pseudomonas sp.]	71.2 / 341	Pipecolic acid synthesis	WP_057005626
cmxB	10347	hybrid non-ribosomal peptide synthetase/type I polyketide synthase [Archangium gephyra]	70.8 / 3200	Non-ribosomal peptide synthesis	WP_047856741
cmxC	1257	MFS transporter [Myxococcus xanthus]	75.3 / 391	Compound secretion	WP_014442384
cmxD	198	lysine biosynthesis protein LysW [Corallococcus coralloides]	73.0 / 50	Chloromethyl-Methoxy-Pipecolic acid carrier Protein	WP_014394897
cmxE	864	Phage major capsid Protein [Micromonospora peucetia]	73.8 / 271	-	SCL73402
cmxF	1125	hypothetical protein [Stigmatella erecta]	79.4 / 324	-	WP_093525581
cmxG	969	SAM-dependent methyltransferase [Myxococcus fulvus]	94.6 / 322	O-Methyltransfer	WP_046714637
cmxH	2022	rubber oxygenase A [Myxococcus fulvus]	83.5 / 674	Chloroperoxidase	AGT20509
cmxI	417	hypothetical protein [Myxococcus hansupus]	68.5 / 124	-	WP_082282800
cmxJ	843	enoyl-CoA hydratase [Myxococcus fulvus]	88.2 / 280	Dehydration reaction in Chloromethyl-Methoxy-Pipecolic acid synthesis	WP_074955256
cmxK	687	TetR/AcrR family transcriptional regulator [Myxococcus fulvus]	85.1 / 228	Chloromyxamide regulation	WP_046714629
cmxL	828	NAD(P)-dependent oxidoreductase [Myxococcus fulvus]	92.4 / 275	Reduction in Chloromethyl-Methoxy-Pipecolic acid synthesis	WP_074955260
cmxM	1839	AMP-dependent synthetase [Myxococcus fulvus]	85.9 / 612	Glutamic acid activation and loading protein	WP_046714631
cmxN	1140	acyl-CoA dehydrogenase [Myxococcus fulvus]	87.9 / 381	Reduction in Chloromethyl-Methoxy-Pipecolic acid synthesis	WP_074955266

cmxO	1149	acyl-CoA dehydrogenase [Myxococcus fulvus]	91.9 / 382	Reduction in Chloromethyl-Methoxy-Pipecolic acid synthesis	WP_046714633
cmxP	1212	acetyl-CoA acetyltransferase [Myxococcus stipitatus]	92.6 / 403	Glutamic acid acetyl transfer	WP_015350521
cmxQ	2178	3-hydroxyacyl-CoA dehydrogenase [Myxococcus fulvus]	92.0 / 725	Reduction in Chloromethyl-Methoxy-Pipecolic acid synthesis	WP_074955272
cmxR	1407	aldehyde dehydrogenase family protein [Myxococcus virescens]	84.0 / 468	Reduction in Chloromethyl-Methoxy-Pipecolic acid synthesis	WP_090487166
cmxS	1008	zinc-binding alcohol dehydrogenase family protein [Pseudomonas composti]	72.5 / 338	Reduction in Chloromethyl-Methoxy-Pipecolic acid synthesis	WP_061240899
orf1	1437	DUF3375 domain-containing protein [Myxococcus fulvus]	88.3 / 478	-	WP_074955289
orf2	615	DUF4194 domain-containing protein [Myxococcus stipitatus]	77.1 / 201	-	WP_015350529
orf3	3384	ATP-dependent exonuclease SbcD, C subunit-like protein [Myxococcus fulvus]	76.2 / 1127	-	WP_074955283
orf4	1188	hypothetical protein [Ectothiorhodospira sp. PHS-1]	53.6 / 390	-	WP_008932172
orf5	1284	phosphopyruvate hydratase [Myxococcus stipitatus]	94.3 / 425	-	WP_015350527
orf6	1032	hypothetical protein [Myxococcus fulvus]	81.5 / 343	-	WP_074955281
orf7	1545	VCBS repeat-containing protein [Cystobacter fuscus]	57.6 / 488	-	WP_002632442
orf8	516	hypothetical protein [Myxococcus stipitatus]	50.3 / 178	-	WP_015350520
orf9	903	bifunctional 5,10-methylene-tetrahydrofolate dehydrogenase/ 5,10-methylene-tetrahydrofolate cyclohydrolase [Myxococcus fulvus 124B02]	91.8 / 291	-	AKF82340
orf10	1719	cobalamin-binding protein [Myxococcus fulvus]	84.1 / 571	-	WP_046711853

The chloromyxamide biosynthetic gene cluster from *Myxococcus species* MCy10608 will be deposited in GenBank under accession number MH168382 upon publication.

S 4.4.1 *In silico* evaluation of the cmxA gene

In order to validate CmxA to be a pipecolic acid producing lysine cyclodeaminase the protein is aligned with the TubZ protein from An d48 that furnishes pipecolic acid in the tubulysin biosynthesis.[2]
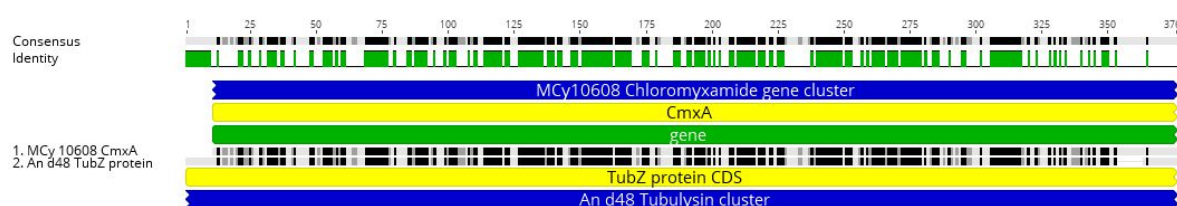
Figure S 2. *MUSCLE alignment of the CmxA gene to the TubZ protein, a validated lysine cyclodeaminase from* Pyxidicoccus fallax *An d48* [3]

S 4.5 Isolation and structure elucidation of chloromyxamide A and B

S 4.5.1   Isolation of chloromyxamide A and B from MCy10615

A 10-liter culture of MCy10615 in liquid VY/2 medium [4] added 2% amberlite resin XAD-16 (Sigma) was grown for 5 days at 160 rpm and 30 °C and after centrifugation the cell pellet and XAD-16 was extracted with 1:1 acetone/MeOH (8x250 mL) to yield 4.16 g of raw extract. This was dissolved in 10 ml MeOH and after centrifugation the extract was loaded onto a sephadex LH-20 column (Ø 2.5 × 80 cm) and eluted with MeOH at a reduced flow of about 4 sec/drop. Fractions of 450 droplets were collected and fractions were pooled based on LCMS analysis. Liquid extraction using water/CHCl$_3$ yielded chloromyxamide A (**1**) and chloromyxamide B (**2**) in the chloroform fraction containing a total of 425 mg solid material. This was further purified on a preparative Waters HPLC-MS system with H$_2$O and MeOH (+0.1 % FA) on an XBridge C$_{18}$ column (5 µm, 19x150 mm) at a flow rate of 17.06 ml/min going from 20-45% MeOH in 27 min with compound **1** eluting at 16 min (1.9 mg semipure fraction) and **2** eluting at 21 min (3.3 mg semipure fraction). Final purification steps on a semipreparative HCT HPLC-MS system using a Luna C$_{18}$ column (4.6 x 250 mm, 5µm) at 45 °C with H$_2$O and MeCN (+0.1% FA) going from 5% MeCN to 15% in 4 min, then to 25% (to 28% for **2**) in 18 min yielded pure **1** (0.31 mg, 14.6 min) and **2** (0.28 mg, 17.2 min).

S 4.5.2   Characterization of all detected chloromyxamides

Chloromyxamide A (**1**): white amorphous solid, [α] = −48.2° (*c* 0.19, MeOH); NMR (700 MHz, methanol-d_4 and DMSO-d_6) see NMR tables; HRESIMS *m/z* 515.2636 [M+H]$^+$ (calcd. for $C_{24}H_{40}ClN_4O_6$, 515.2631, Δ = 0.9 ppm).

Chloromyxamide B (**2**): white amorphous solid, [α] = −34.1° (*c* 0.15, MeOH); NMR (700 MHz, methanol-d_4) see NMR tables; HRESIMS *m/z* 529.2791 [M+H]$^+$ (calcd. for $C_{25}H_{42}ClN_4O_6$, 529.2787, Δ = 0.8 ppm).

Chloromyxamide C* (**3**): HRESIMS *m/z* 531.2583 [M+H]$^+$ (calcd. for $C_{24}H_{40}ClN_4O_7$, 531.2580, Δ = 0.6 ppm). This compound was not isolated due to low yields and unstable production.

Chloromyxamide D* (**4**): HRESIMS *m/z* 517.2424 [M+H]$^+$ (calcd. for $C_{23}H_{38}ClN_4O_7$, 517.2424, Δ = 0.1 ppm). This compound was not isolated due to low yields and unstable production.

Chloromyxamide E* (**5**): HRESIMS *m/z* 529.2793 [M+H]$^+$ (calcd. for $C_{25}H_{42}ClN_4O_6$, 529.2789, Δ = 0.7 ppm). This compound was not isolated due to low yields and unstable production.

Figure S 3. *Exact masses of all chloromyxamide derivatives as [M+H]⁺ ion acquired with our maXis 4G qTOF HRMS spectrometer in ESI positive mode, the retention time in the top right corner corresponds to the compounds retention time using our standard gradient, the number in the lower right corner corresponds to the respective chloromyxamide derivative*

S 4.5.2.1 Chloromyxamide production potential

As mentioned in the main text, the three chloromyxamide producer strains MCy10608, MCy10615 and MCy10649 produce different chloromyxamide derivatives. The chloromyxamide derivatives detected in each strain are grouped in Table S 3.

Table S 3. *Chloromyxamide production potential per strain*

Strain	MCy10608	MCy10615	MCy10649
Chloromyxamide A	+	+	+
Chloromyxamide B	-	+	+
Chloromyxamide C	+	-	-
Chloromyxamide D	+	-	-
Chloromyxamide E	-	-	+

Circular dichroism measurements were performed for natural chloromyxamide A and B as well as for synthetic chloromyxamide A at 1.0 mM in MeCN (Figure S 21 and Figure S 22).

S 4.5.3 Stable isotope labelling of the chloromyxamides

The non-canonical amino acids present in the chloromyxamide scaffold e.g. methyl proline, hydroxy proline and pipecolic acid are likely to be produced via the biosynthesis routes described in literature. These routes include the pipecolic acid synthesis by the TubZ analogon CmxA (Table S 2), methyl proline biosynthesis via oxidation and cyclisation of leucine as described in griselimycin biosynthesis and hydroxylation of proline by a prolyl-4-hydroxylase.[2,5,5,6] Secondly, according to our biosynthesis hypothesis as described by main text Figure 3, we hypothesized the novel 6-chloromethyl-5-methoxypipecolic acid (CMPA) amino acid to be synthesized from glutamic acid. In order to support these theories with experimental evidence we fed the stable isotopically labelled amino acid precursors *L*-leucine-5,5,5-d_3, *L*-glutamic acid-2,3,3,4,4-d_5, *L*-Lysine dihydrochloride-$^{13}C_6$,$^{15}N_2$ and *L*-proline-$^{13}C_5$,^{15}N (Cambridge Isotope Labs). Feeding was achieved by adding a total amount of 2ml of a 25mM amino acid stock solution of the respective amino acid into a 25ml MCy10608 VY/2-S medium fermentation culture (for **1,3** and **4**) or a MCy10649 VY/2-S medium fermentation culture (for **1,2** and **5**) in a 100ml Erlenmeyer flask. The amino acid stock solutions are prepared in dd-H_2O except for *L*-leucine-5,5,5-d_3 that is dissolved in 50:50 H_2O:MeOH due to solubility issues. Extraction is done according to standard protocol for LC-MS samples (section S S 4.2).

S 4.5.3.1   Feeding of *L*-glutamic acid-2,3,3,4,4-d$_5$

As expected, feeding of *L*-glutamic acid-2,3,3,4,4-d$_5$ led to a mass shift of approx. +5 Da in all the chloromyxamide derivatives as all of them possess the CMPA unit. The +4 Da peak is also elevated as amino acids featuring a deuterium atom in C-2 position suffer from H-D exchange that forms *L*-glutamic acid-3,3,4,4-d$_4$ which is also incorporated to a minor extent.
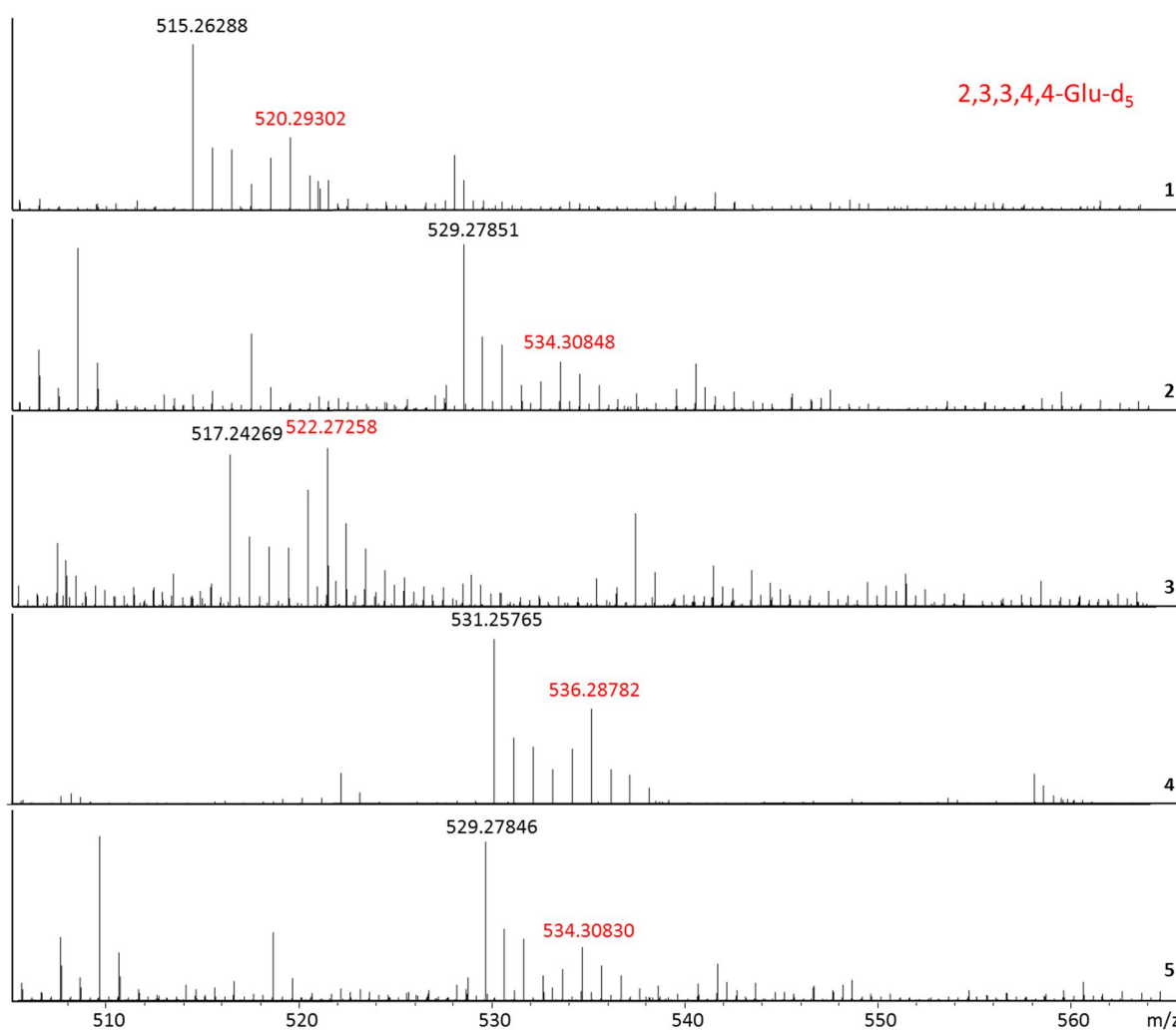
Figure S 4. *Exact masses of all chloromyxamide derivatives as [M+H]$^+$ ion acquired with our maXis 4G qTOF HRMS spectrometer in ESI positive mode after feeding of* L-*glutamic acid-2,3,3,4,4-d$_5$, the number in the lower right corner corresponds to the respective chloromyxamide derivative, exact monoisotopic mass of the unlabeled derivative is depicted in black, exact mass of the chloromyxamide-d$_5$ derivatives is depicted in red*

S 4.5.3.2   Feeding of *L*-lysine dihydrochloride-$^{13}C_6,^{15}N_2$

Since during pipecolic acid biosynthesis by the CmxA protein Lysine is cyclized we observe loss of one labelled nitrogen atom in *L*-lysine -$^{13}C_6,^{15}N_2$ during the formation of *L*-pipecolic acid-$^{13}C_6,^{15}N$. This explains the observed +7 Da shifts upon lysine incorporation as pipecolic acid. Since **4** does not contain pipecolic acid, no mass shift is observed. **5** displays incorporation of two labelled pipecolic acid building blocks as one would expect since the structure contains two pipecolic acid amino acids (main Text Figure 2).
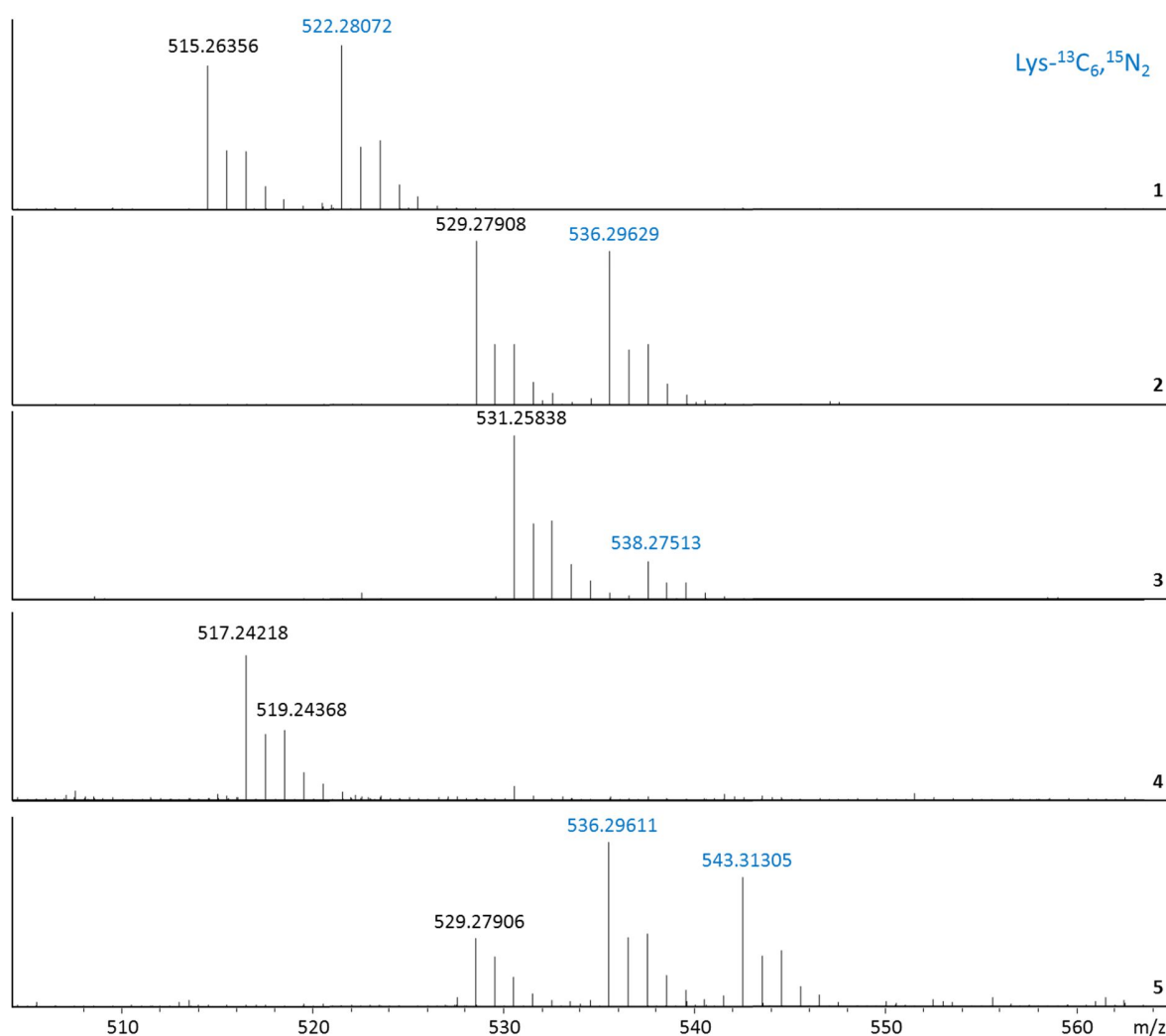
Figure S 5. *Exact masses of all chloromyxamide derivatives as [M+H]$^+$ ion acquired with our maXis 4G qTOF HRMS spectrometer in ESI positive mode after feeding of L-lysine dihydrochloride-$^{13}C_6,^{15}N_2$, the number in the lower right corner corresponds to the respective chloromyxamide derivative, exact monoisotopic mass of the unlabeled derivative is depicted in black, exact mass of the chloromyxamide-$^{13}C_6,^{15}N_2$ and chloromyxamide-$^{13}C_{12},^{15}N_4$ derivatives is depicted in blue*

S 4.5.3.3   Feeding of *L*-proline-$^{13}C_5,^{15}N$

As explained in S 4.5.3, the hydroxyproline in **3** and **4** as well as obviously the proline residues in **1** and **2** provide from proline either via direct incorporation or via incorporation after hydroxylation. Ass all of these proline containing chloromyxamide derivatives are produced by MCy10608, labelled proline was not fed to the strains producing **2** and **5**. Unfortunately, as explained in the main text, we were never able to isolate the hydroxyproline derivatives **3** and **4** in sufficient amounts to allow NMR based structure elucidation. Proline feeding as done in Figure S 6 proves that all hydroxyprolines provide from proline leading to labelling of **1** and **3** while double labelling **4** that contains proline and hydroxyproline. It is therefore likely that the hydroxyproline moiety is in fact 4-hydroxyproline as this amino acid is often observed in nature and also directly synthesized by hydroxylation of proline.[5] Unfortunately, as also 3-hydroxyproline is observed in natural products we are unable to determine the exact position of the hydroxyl group on the terminal proline moiety.[7]
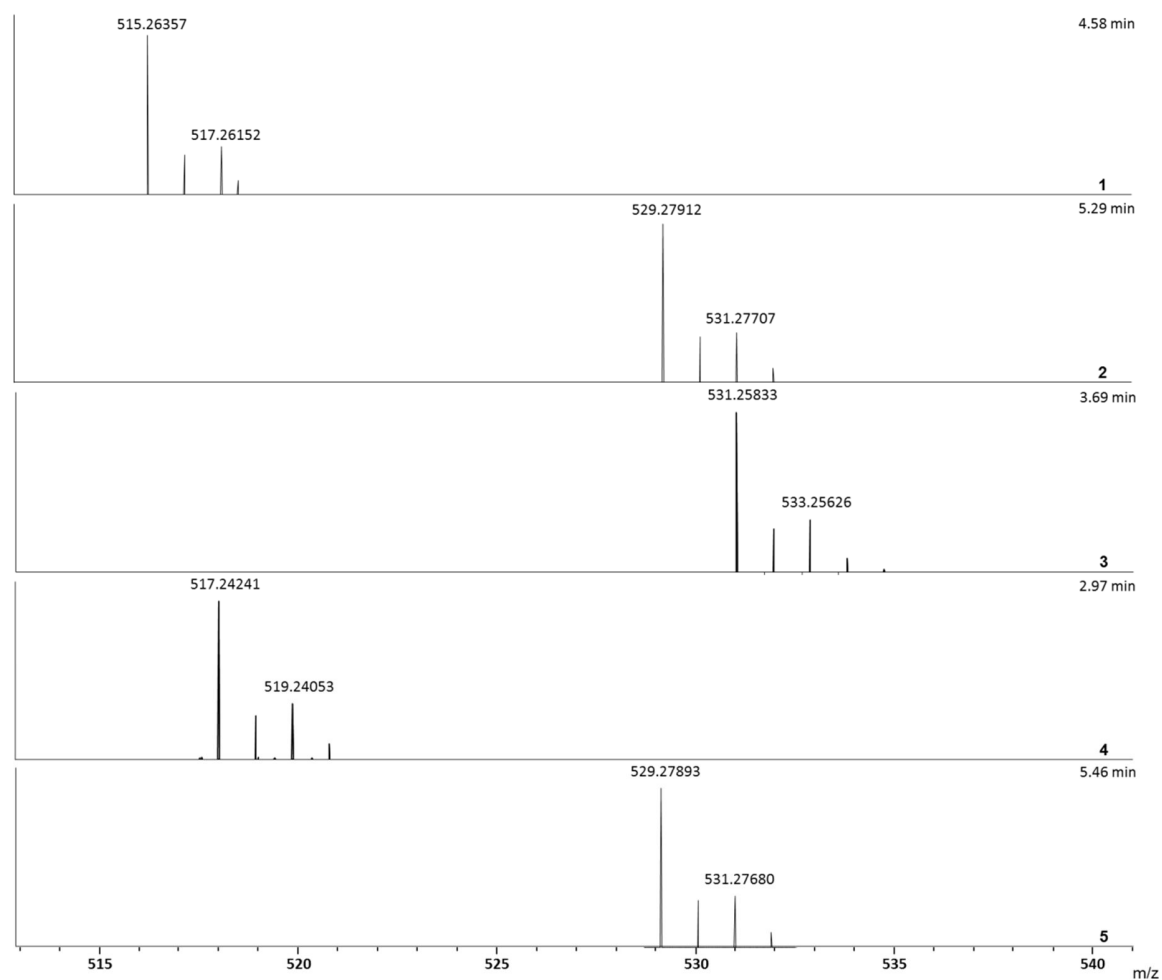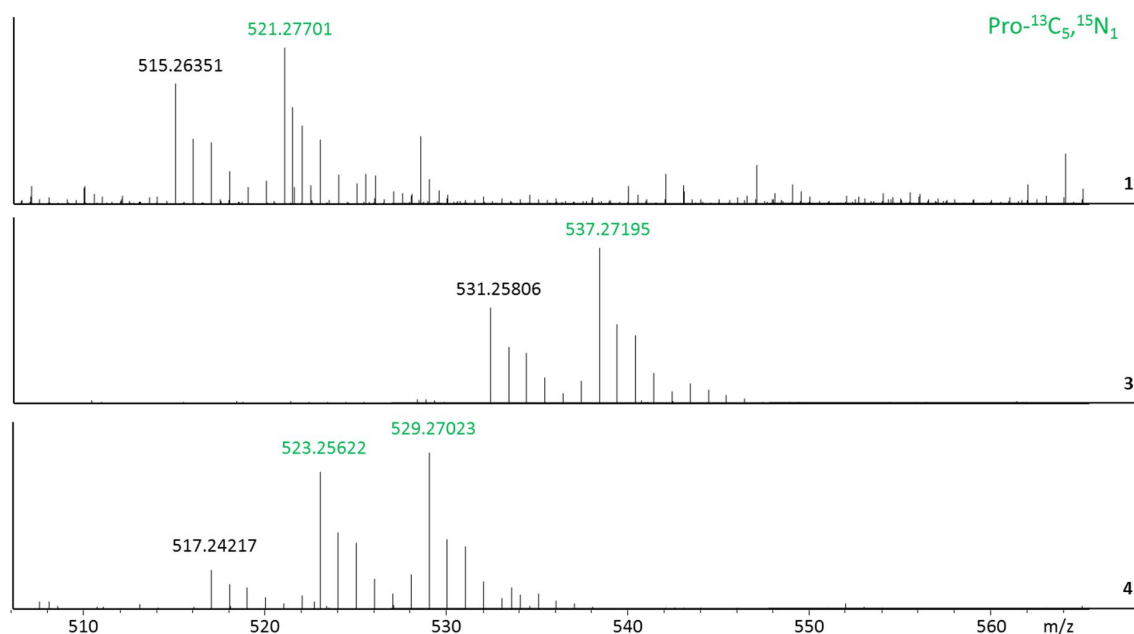

Figure S 6. *Exact masses of all chloromyxamide derivatives as [M+H]⁺ ion acquired with our maXis 4G qTOF HRMS spectrometer in ESI positive mode after feeding of L-proline-$^{13}C_5,^{15}N$, the number in the lower right corner corresponds to the respective chloromyxamide derivative, exact monoisotopic mass of the unlabeled derivative is depicted in black, exact mass of the chloromyxamide-$^{13}C_5,^{15}N$ and chloromyxamide-$^{13}C_{10},^{15}N_2$ derivatives is depicted in green*

S 4.5.3.4   Feeding of *L*-leucine-5,5,5-d$_3$

Since at this point, the origins of all non-canonical amino acids were determined except for the biosynthesis of the 4-methylproline we fed labelled lysine to MCy10649, the producer strain of the 4-methylproline containing **2**. According to the work on griselimycins, this amino acid should be synthesized from *L*-leucine although, as no genetic information is available on the producer strains of **2**, we were unable to search for the corresponding genes.[6] Nevertheless, feeding of *L*-leucine-5,5,5-d$_3$ to a MCy10649 VY/2-S medium culture led to observation of extensive labelling of **2** indicated by the +3 Da precursor mass shift. Thus, *Pyxidicoccus sp.* MCy10649 most likely uses analogous enzymology to synthesize 4-methylproline.
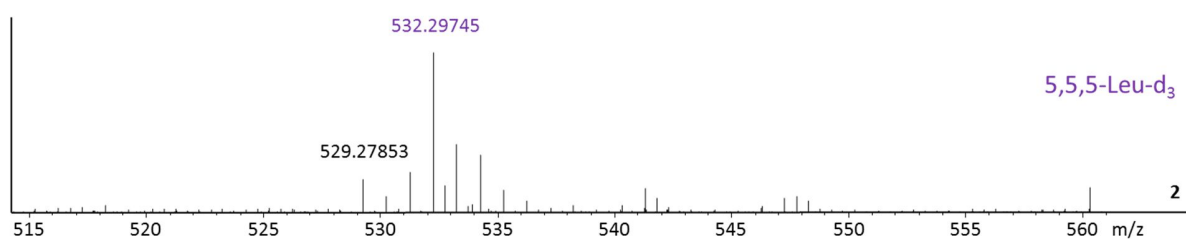
Figure S 7. *Exact masse of **2** as [M+H]$^+$ ion acquired with our maXis 4G qTOF HRMS spectrometer in ESI positive mode after feeding of L-leucine-5,5,5-d$_3$, exact monoisotopic mass of the unlabeled derivative is depicted in black, exact mass of the chloromyxamide B-d$_3$ (**2**) is depicted in violet*

S 4.5.4   Determination of the position of the labelled amino acids by tandem MS experiments

The chloromyxamides show very regular fragmentation in tandem MS experiments done according to our standard MS² protocol on our UHPLC hyphenated HRMS qTOF system (section S S 4.2). Therefore, upon stable isotope labelling of the chloromyxamides we can not only observe the general mass shifts of the precursor mass as described in S 4.5.3 but also follow up to determine the exact position of the labeled amino acid via MS² experiments.

S 4.5.4.1   Color-coded structure formulas indicating the incorporation of the different amino acids into the chloromyxamide scaffolds

In order to assign every labelled amino acid to their respective position in the chloromyxamide scaffold we applied scheduled precursor list based MS² experiments that specifically target the labelled chloromyxamide precursor. Therefore, we can validate the position the labelled amino acid ends up in by following up labelling to the respective product ion in the tandem MS experiment.
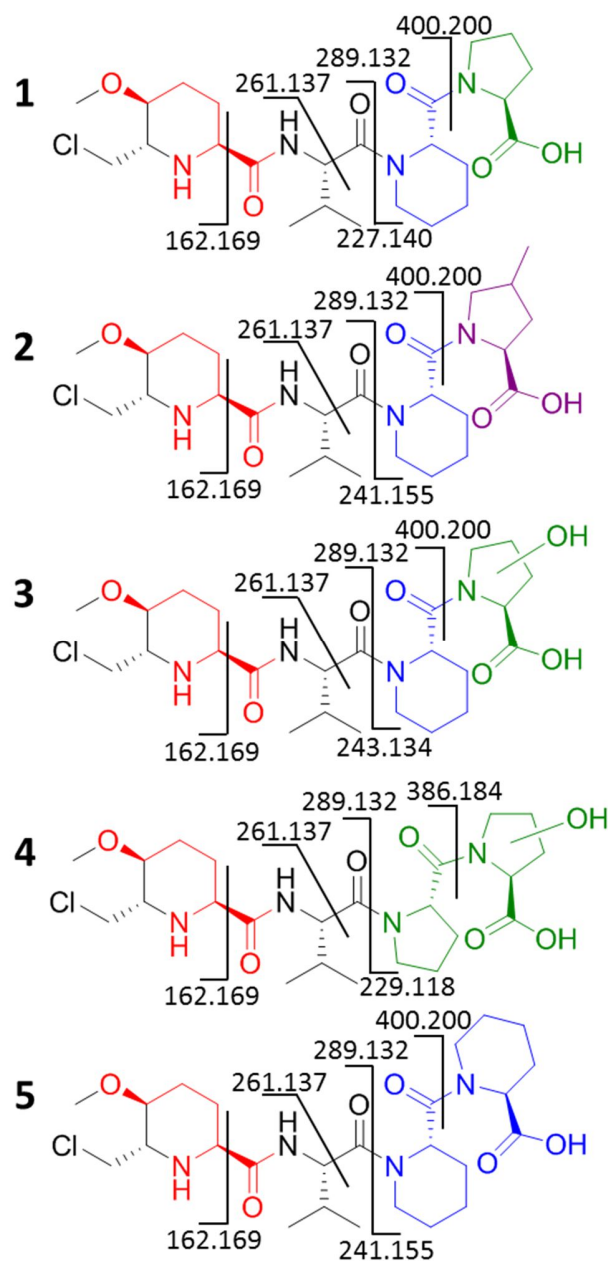
Figure S 8. *Color coded structure formulas indicating the incorporation of the different amino acids into the chloromyxamide scaffolds, the numbers represent the tandem MS product ions acquired with our maXis 4G qTOF HRMS spectrometer in ESI positive mode, the color coded parts represent carbon atoms providing from canonical amino acids as proven by tandem MS experiments on stable isotope labelled chloromyxamides as seen in S 4.5.4.*

S 4.5.4.2   Tandem MS analysis of stable isotope labelled chloromyxamide A (1)
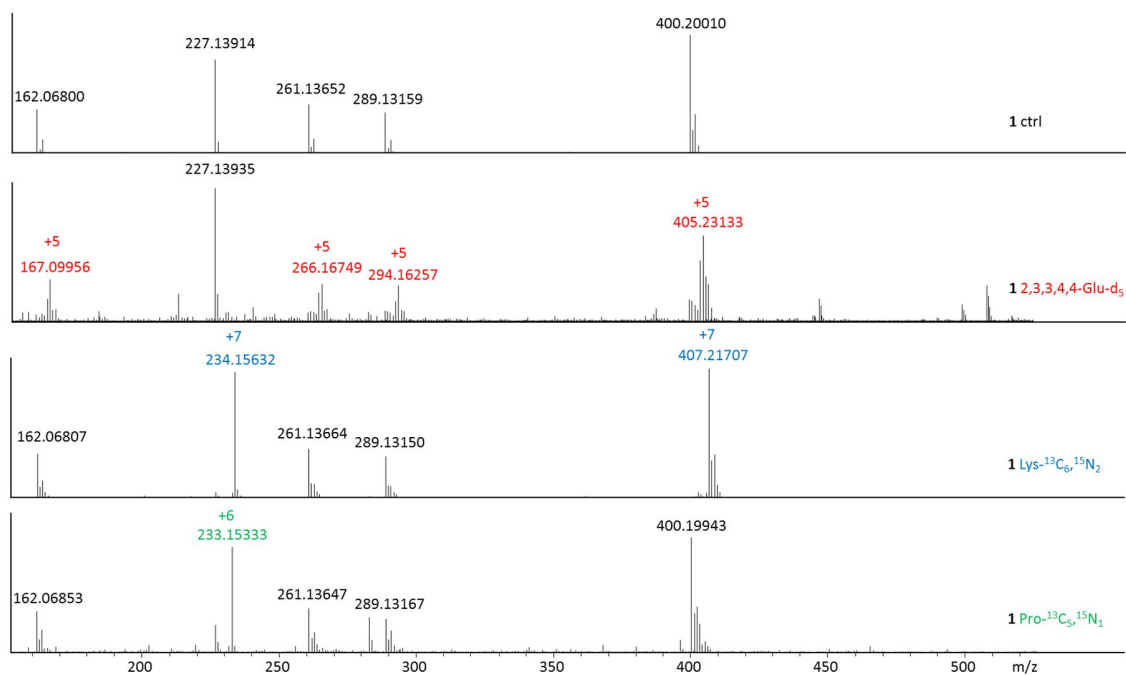
Figure S 9. *Tandem MS² spectra of chloromyxamide A (**1**) of the control and from feeding experiments with different stable isotope labelled precursors*

S 4.5.4.3   Tandem MS analysis of stable isotope labelled chloromyxamide B (2)
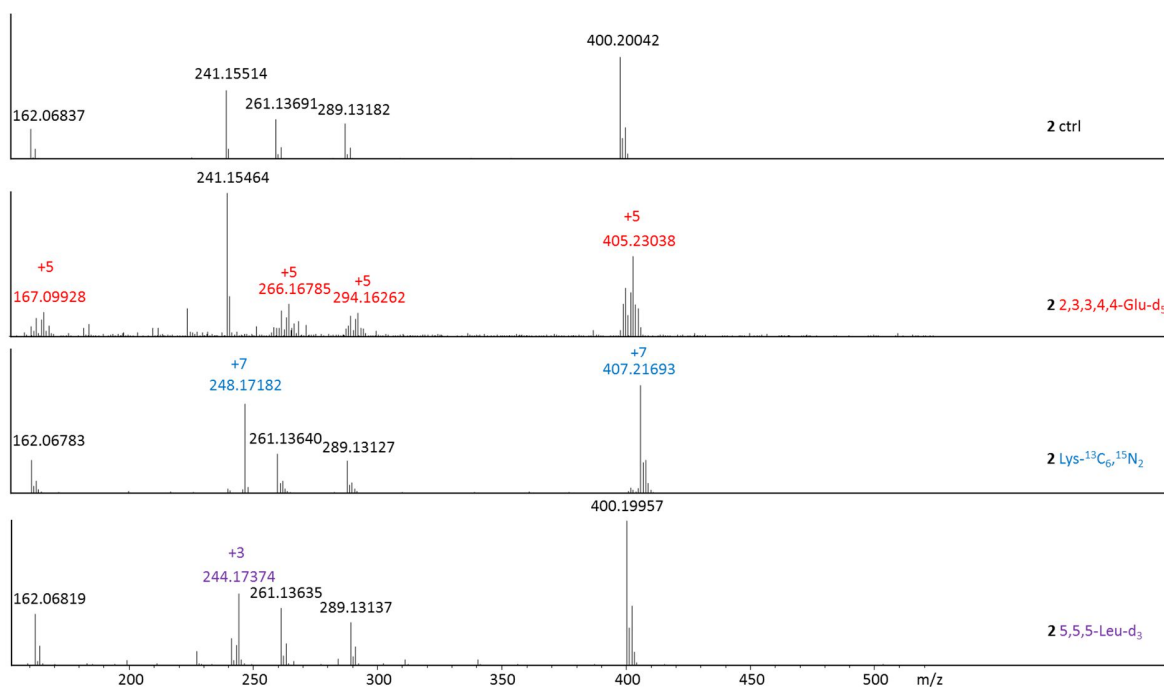
Figure S 10. *Tandem MS² spectra of chloromyxamide B (**2**) of the control and from feeding experiments with different stable isotope labelled precursors*

S 4.5.4.4   Tandem MS analysis of stable isotope labelled chloromyxamide C (3)
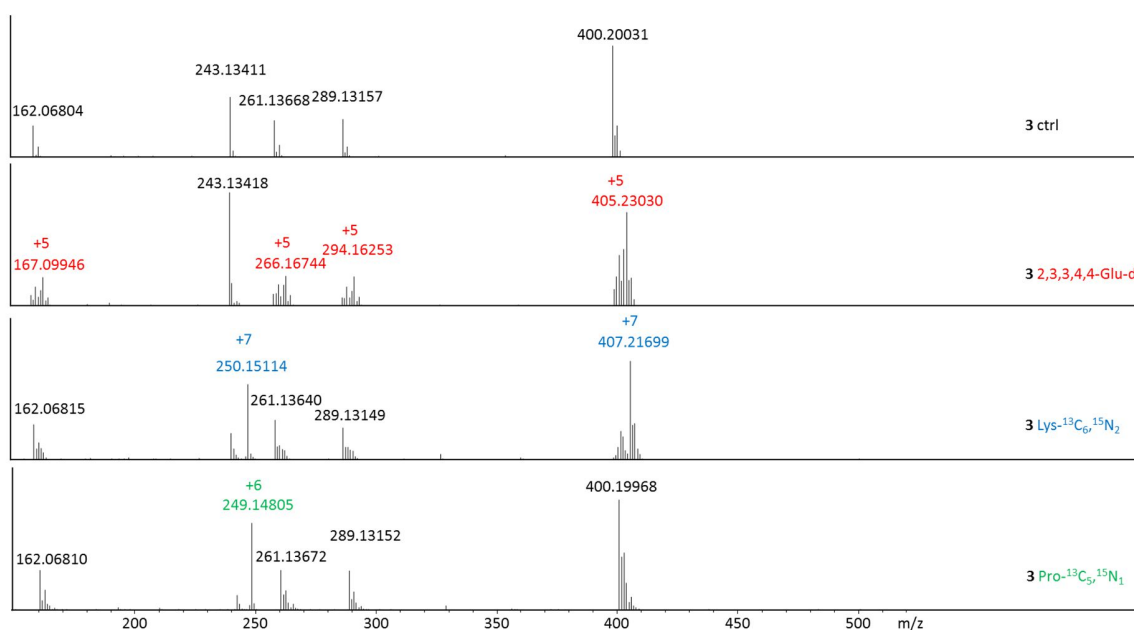
Figure S 11. *Tandem MS² spectra of chloromyxamide C (**3**) of the control and from feeding experiments with different stable isotope labelled precursors*

S 4.5.4.5   Tandem MS analysis of stable isotope labelled chloromyxamide D (4)
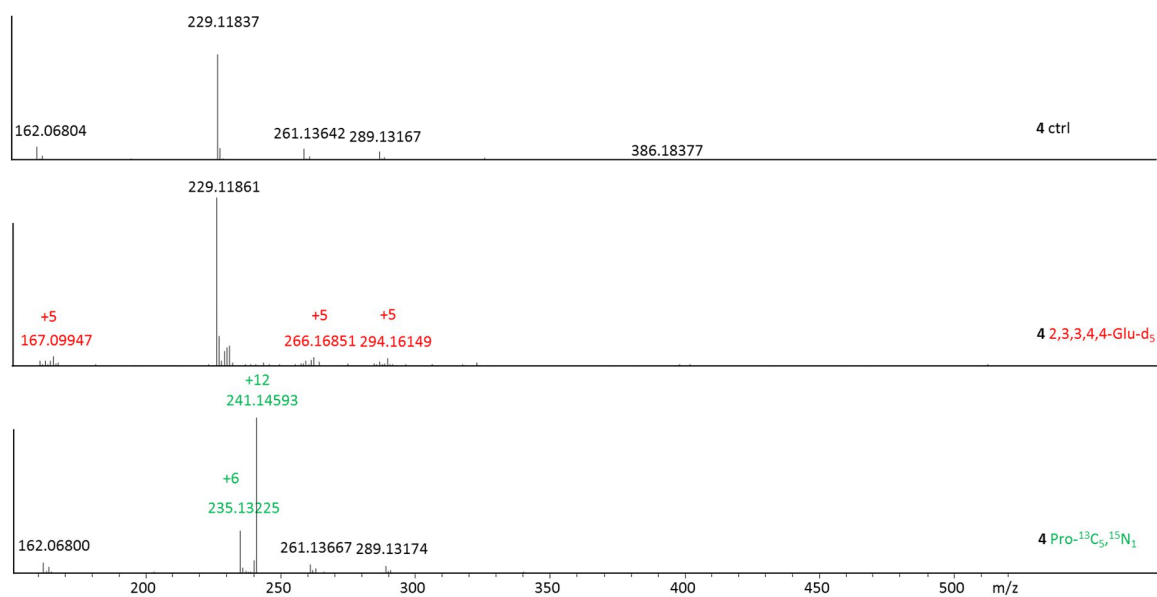
Figure S 12. *Tandem MS² spectra of chloromyxamide D (**4**) of the control and from feeding experiments with different stable isotope labelled precursors*

S 4.5.4.6   Tandem MS analysis of stable isotope labelled chloromyxamide E (5)
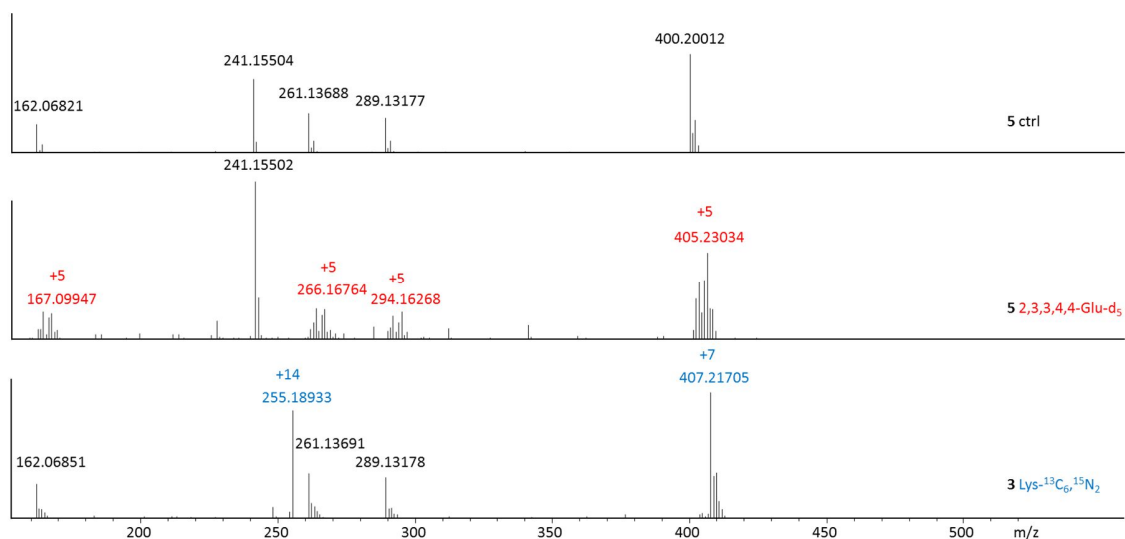
Figure S 13. *Tandem MS² spectra of chloromyxamide E (**5**) of the control and from feeding experiments with different stable isotope labelled precursors*

Tandem MS analysis of all chloromyxamide derivatives confirmed the amino acid sequences and furthermore made it possible to obtain structural information for additional chloromyxamide family members from MCy10608 and MCy10649 not isolated due to low yields in larger scale cultures. These compounds were named chloromyxamide C (**3**, $C_{24}H_{39}N_4O_7Cl$), D (**4**, $C_{23}H_{37}N_4O_7Cl$) and E (**5**, $C_{25}H_{41}N_4O_6Cl$). To obtain information-enriched tandem MS data for amino acid sequence analysis of chloromyxamides and to gain first insights into their biosynthesis, we supplemented MCy10615 cultures with stable isotope labelled amino acids. Feeding of L-glutamic acid-2,3,3,4,4-d_5, L-leucine-5,5,5-d_3, $^{13}C_5$,^{15}N-L-proline and $^{13}C_6$,$^{15}N_2$-L-lysine proved not only that the minor derivatives 3-5 are synthesized via the same route as 1 and 2 but also shed light on the biogenesis of all non-canonical amino acids in chloromyxamide (main text Figure 3). MS² experiments focusing on the isotopically labelled chloromyxamides allowed to follow up on the position of every labelled amino acid in the peptide, supporting the biosynthetic origin not only for **1** and **2** but also for **3**-**5**. The CMPA-Val substructure is conserved in all chloromyxamide congeners as judged on the basis of the corresponding product ion (main text Figure 2). Furthermore, *L*-glutamic acid-2,3,3,4,4-d_5 feeding shows incorporation of glutamic acid into the CMPA unit as proven by MS² analysis selectively applied to the labelled chloromyxamide. Therefore, CMPA biosynthesis has to start from glutamic acid that is extended with a C_2 unit and subsequently cyclized (main text Figure 3). All chloromyxamide congeners except **4** show a prominent product ion attributed to the CMPA-Val-Pip fragment, suggesting that **3** and **5** also contain Pip as the third amino acid. Furthermore, feeding of $^{13}C_6$, $^{15}N_2$ *L*-lysine led to a +7 Da shift in the mass spectra of **1, 2** and **3** while a +7 and a +14 shift is observed for **5** indicating the presence of 1 respective 2 pipecolic acid units in the molecule originating from lysine. Compound **4** shows instead a weaker fragment that is one CH_2-unit smaller than CMPA-Val-Pip and which is not labelled by $^{13}C_6$, $^{15}N_2$ *L*-lysine feeding. This fragment is assigned to CMPA-Val-Pro. Fragmentation is favored on the N-terminal side of Pro as well as on the C-terminal side of Pip, which in this case makes this position ideal for dissociation.[8] Compound **4** favors the formation of CMPA-Val and CMPA-Val-CO ions instead of the dissociation between the two proline residues. Contrary to all other chloromyxamides, **4** is not labelled by feeding of $^{13}C_6$, $^{15}N_2$ *L*-Lysin, which is the precursor of pipecolic acid, but instead displays a +12 Da shift upon feeding of $^{13}C_5$, ^{15}N *L*-proline; compound **4** thus seems to contain two proline moieties. As **1** and **2** show a +6 Da shift upon feeding of $^{13}C_5$, ^{15}N *L*-proline we can draw two conclusions: i) Tandem MS data and feeding studies suggest **4** to possess a proline residue instead of the pipecolic acid moiety, and ii) the hydroxyproline residue present in **3** and **4** is synthesized via direct oxidation of proline. Labelling of **2** by *L*-leucine-5,5,5-d_3 at amino acid 4 ultimately reveals biosynthesis of the methyl proline unit in **2** as analogous to the methyl proline pathway described in griselimycin biosynthesis.[6] Unfortunately, isolation of **3** and **4** in amounts sufficient for NMR analysis was unsuccessful; we were thus unable to determine whether these

chloromyxamide congeners contain 3- or 4- hydroxyproline. We can thus tell that the sequence of **1** to be CMPA-Val-Pip-Pro, the sequence of **2** to be CMPA-Val-Pip-4-MePro, the sequence of **3** to be CMPA-Val-Pip-4-HydroxyPro, the sequence of **4** to be CMPA-Val-Pro-4-HydroxyPro and the sequence of **5** to be CMPA-Val-Pip-Pip. Furthermore, we verified CMPA to contain glutamic acid, hydroxyproline to be synthesized from proline, while 4-methylproline is synthesized from leucine. All these findings are in good agreement with the biosynthesis hypothesis we drafted for the chloromyxamides in main text Figure 3.

S 4.5.5   Marfey's analysis of 1 and 2

Two portions each of 15 µg **1** and **2** were hydrolyzed in 100 µl 6 M HCl at 110 °C for 16 hours and redissolved in 50 µl H_2O followed by addition of 20 µl 1 M $NaHCO_3$ and 20 µl 1% *L*- or *D*-Marfey's reagent in acetone (*L*-FDLA and *D*-FDLA, respectively). After reaction at 40 °C for 90 min, 10 µl 2 M HCl was added to neutralize the mixtures and then diluted with 150 µl MeCN.

Derivatives of *L*- and *D*-amino acid standards (50 µl, 0.05 M of *D*- and *L*-Val, *L*-Pro, and *L*-Pip) were produced by the same procedure except the final dilution was done with 300 µl MeCN for the pure amino acids. Standard derivatives of (2S,4R)-4MePro and a mixture with all 4-MePro was generously provided by Kevin Sours. Analysis by LCMS on a Acquity BEH C18 (2.1 x 100 mm, 1.7 µm) using H_2O-MeCN (0.1 % FA) at 0.55 ml/min going from 5-10 % MeCN in 1 min, to 35% in 14 min, to 50 % in 7 min and to 80 % in 3 min. Detection by DAD and UHPLC maXis4G qTOF HRMS in positive mode.

L-FDLA derivatives eluted at 15.36 min (*L*-Pro), 17.02 min (*L*-Val), 17.07 min (*D*-Pro), 17.57 min (2S,4R-4MePro), 18.80 min (*D*-Pip), 19.65 min (*L*-Pip), and 21.05 min (*D*-Val). *D*-FDLA derivatives eluted at 15.45 min (*D*-Pro), 17.13 min (*D*-Val), 17.14 min (*L*-Pro), 18.68 min (*L*-Pip), 19.12 min (2S,4R-4MePro), 19.67 min (*D*-Pip), and 21.11 min (*L*-Val). Derivatives of the 4-MePro mixture eluted at 17.04, 17.54, 19.12, and 19.28 min for both *L*- and *D*-FDLA.[9]

Marfey's analysis of **1** and **2** showed that the compounds contained *L*-Val, *L*-Pip and *L*-Pro (for **1**), and *L*-Val, *L*-Pip, and (2*S*,4*S*)-4-MePro for **2**.[9] The structure of **1** thus has the amino acid sequence CMPA-*L*-Val-*L*-Pip-*L*-Pro (Figure 1), whereas in **2** the proline is replaced by a (2*S*,4*S*)-4-methylproline giving the sequence CMPA-*L*-Val-*L*-Pip-(2*S*,4*S*)-4MePro.

S 4.5.6 Structure elucidation of chloromyxamide A and B using multi-dimensional NMR analysis

As described in the main text, only MCy10615 produces the chloromyxamides in large scale cultures. Therefore, as seen in S 4.5.1, we were only able to obtain sufficient amounts of **1** and **2** to perform multi-dimensional NMR based structure elucidation. The other derivatives only occurring in MCy10608 and MCy10649 could only be assigned to their putative structures by tandem MS experiments combined with stable isotope labelling (section S 4.5.2 and S 4.5.3). Structure elucidation of **1** was achieved employing ^{1}H, DQF-COSY, HSQC, HMBC and ROESY experiments, while structure elucidation of **2** was achieved using ^{1}H, DQF-COSY and HSQC experiments. 2D NMR data used for structure elucidation is acquired in methanol-d_4 on a Bruker Ascend 700 spectrometer with a 5 mm TXI cryoprobe (^{1}H at 700 MHz, ^{13}C at 175 MHz) NMR spectrometer. DQFCOSY, HSQC, HMBC, and ROESY experiments were recorded using standard pulse programs. To elucidate the relative configuration of the chloromyxamides' CMPA unit, S^3 HMBC experiments are employed as described under S 4.5.7.
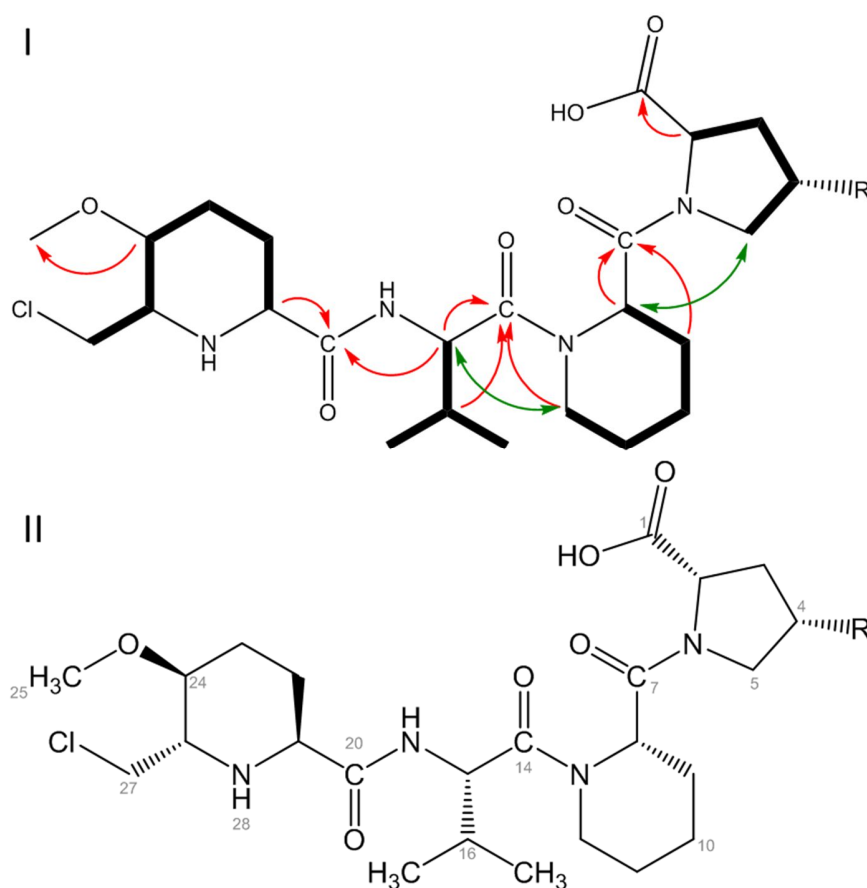
***Figure S 14. I)** Structure of **1** (R=H) and **2** (R=CH$_3$), showing key NMR correlations (black COSY, red HMBC, and green ROESY). **II)** Atom numbers for assignment of the NMR data in the natural products for **1** (R=H) and **2** (R=CH$_3$).*

Table S 4. *Tabulated 1D and 2D NMR data (700 MHz, methanol-d$_4$) for 1 and 2*

	Chloromyxamide A (1)				Chloromyxamide B (2)[b]	
#	δ$_H$, multiplicity, (*J* in Hz)	δ$_C$	HMBC	ROESY	δ$_H$, multiplicity, (*J* in Hz)	δ$_C$
Pro/4MePro						
1		176.0				
OH						
2	4.42, 1H, dd (8.5, 5.0)	60.6	1, 3, 4		4.33, dd, 1H, 10, 7	62.1
3a	2.25, 1H, m	29.9	1, 2, 4, 5		2.45, m, 1H	38.2
3b	1.97, 1H, m		1, 2, 4, 5		1.48, m, 1H	
4a	2.05, 1H, m	25.9	2, 3, 5		2.31, m, 1H	35.0
4b	2.00, 1H, m		2, 3, 5			
4-Me					1.10, d, 3H, 6.5	16.6
5a	3.74, 1H, m	48.0	(2), 3, 4	8	3.93, dd, 1H, 10.2, 7.5	55.4
5b	3.63, 1H, m		3, 4	8	3.14, dd, 1H, 11, 10	
6-N						
Pip						
7		172.1				
8	5.20, 1H, dd (6.3, 2.8)	52.6	7, 9, 10, 12, (14)	5a, 5b	5.25, dd, 1H, 6, 3	52.5
9a	2.12, 1H, m	26.2	(11)		2.17, m, 1H	26.3
9b	1.76, 1H, m		7, (8), 10		1.73, m, 1H	
10	1.66, 2H, m	20.1	(9)		1.65, m, 2H	20.2
11a	1.79, 1H, m	25.7			1.78, m, 1H	25.7
11b	1.58, 1H, m		(12)		1.56, m, 1H	
12a	3.97, 1H, br d (12.7)	44.6	8, 10, 11, 14	15	3.96, br d, 1H, 12	44.5
12b	3.74, 1H, m		14	15	3.76, m, 1H	
13-N						
Val						
14		173.4				
15	4.84[a], 1H	55.0	14, 16, 17, 18, 20	12a, 12b	4.85[a], 1H	54.8
16	2.12, 1H, m	31.4	14, 15, 17, 18		2.12, m, 1H	31.3
17	0.97, 3H, d (6.8)	20.0	15, 16, 18		0.97, 3H, d (6.8)	19.9
18	0.88, 3H, d (6.8)	17.7	15, 16, 17		0.87, 3H, d (6.8)	17.6
19-NH						
CMPA						
20		174.4				
21	3.56, 1H, dd (5.6, 2.8)	55.6	20, 22, 23, 25		3.55, 1H, dd (5.5, 2.7)	55.6
22a	2.32, 1H, m	25.9	21, 23		2.32, 1H, m	25.9
22b	1.63, m, 1H		20, 21, 23, 24		1.63, 1H, m	
23a	2.13, m, 1H	26.5	21, 22, 24, 25		2.12, 1H, m	26.4
23b	1.23, m, 1H		21, 22, 24, 25		1.22, 1H, m	
24	3.09, ddd, 1H, 10, 9, 4.0	77.8	(22), 25, 26, 27		3.09, 1H, ddd (10, 9, 4.1)	77.8
25-OMe	3.35, s, 3H	56.1	24		3.35, 3H, s	56.1
26	2.83, ddd, 1H, 9, 6.1, 3.1	58.8	(23), 24		2.82, 1H, ddd (8.6, 6.0, 3.1)	58.8
27a	3.83, dd, 1H, 11.0, 3.1	46.4	24, 26		3.83, 1H, dd (11.0, 3.1)	46.4
27b	3.75, dd, 1H, 11.0, 6.1				3.74, 1H, dd (11.0, 6.1)	
28-NH						

[a] Obscured by HDO resonance. [b] Only ^{1}H, DQF-COSY and gHSQC data available.

S 4.5.7   Molecular modelling and S^3 HMBC

Minimum energy calculations of the structures were performed in Maestro 10.1 using the Schrödinger software package. Conformational searching and minimum energy calculations were performed using Macromodel. Calculations were done in CHCl$_3$ with the OPLS2005 force field and minimum 10000 iterations and gradient converging for the minimizations. All calculations converged to

the structure in Figure S 17. $^3J_{HH}$ coupling constants were calculated with the built-in coupling constant analyzer, and therefore only standard range coupling constants were used in comparison to experimental coupling constants determined with the S^3 HMBC experiment.[10]

S 4.5.7.1   Molecular modelling of chloromyxamide A (1)

Minimum energy calculations of **1** with all *S* stereochemistry shows a strong hydrogen bond between the two peptide ends, which suggests that the peptide adopts a pseudo-cyclic structure in solution.

#	J_{HH} measured [Hz]	J_{HH} calculated [Hz]
H2$_{eq}$-H3$_{eq}$	3	2
H2$_{eq}$-H3$_{ax}$	6	5
H3$_{ax}$-H4$_{ax}$	>10*	14
H4$_{eq}$-H5$_{ax}$	4	5
H4$_{ax}$-H5$_{ax}$	10	11
H5$_{ax}$-H6$_{ax}$	9	9
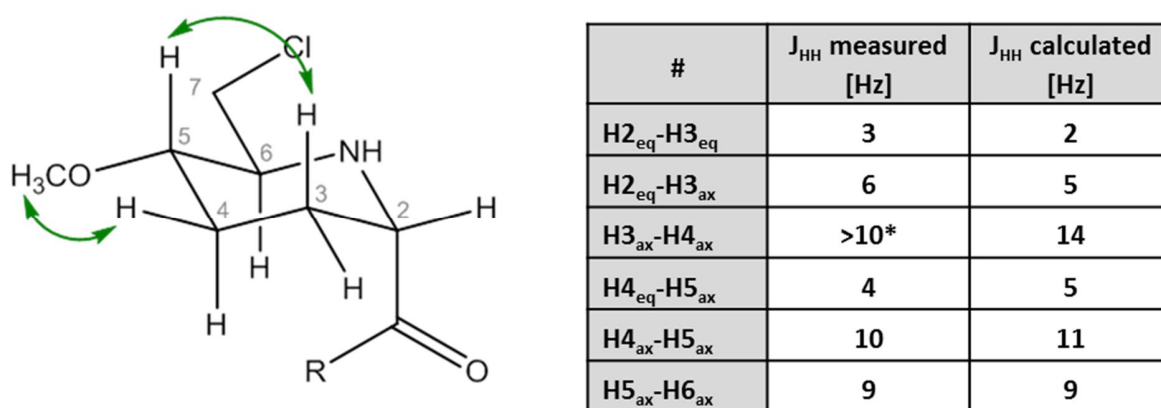
*Figure S 15. Proposed configuration of the 2S,5S,6S-CMPA as well as measured and calculated 1H-1H coupling constants in the ring system. Two key ROESY correlations are shown with green arrows. * seen as a large splitting bigger than 10 Hz.*
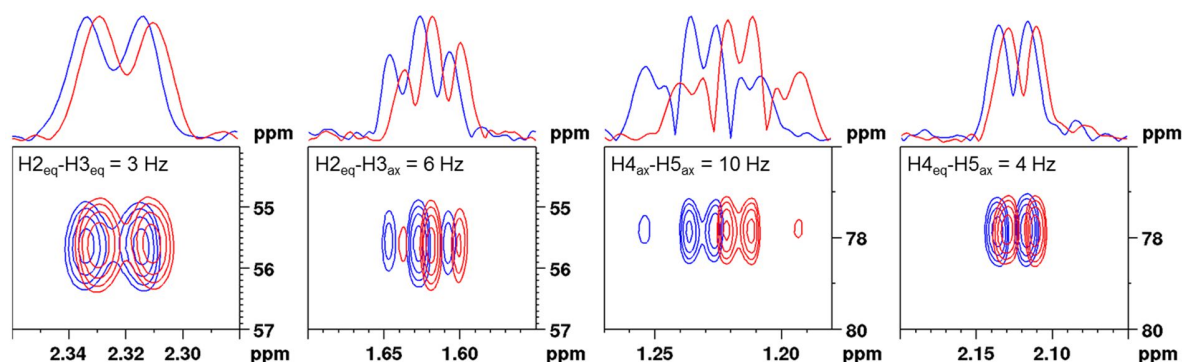
*Figure S 16. Example cross peaks from S^3 HMBC of 1, used for determination of homonuclear coupling constants. See full spectrum in section **Fehler! Verweisquelle konnte nicht gefunden werden.**.*
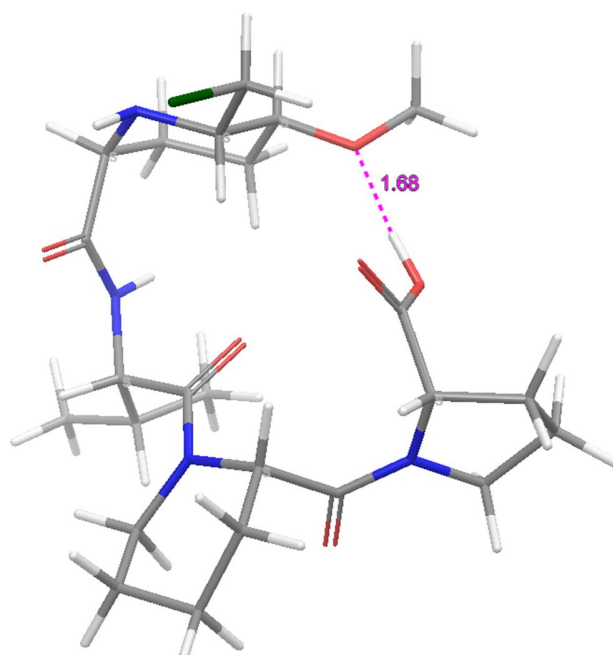
Figure S 17. *Minimum energy conformation for 1 in CHCl₃ showing a strong intramolecular hydrogen bond between the carboxyl group of Pro and the –OMe of CMPA. Using water as solvent yields a similar result.*

S 4.5.9   Biological activity evaluation of the chloromyxamides

Chloromyxamide A has been subjected to antimicrobial testing against our standard set of pathogens and cancer cell lines to evaluate biological activity of chloromyxamide A.

Table S 5. *MIC determinations for chloromyxamide A against common pathogens and cancer cell lines*

Pathogenic strain	MIC Chloromyxamide A [µg/ml]
S. aureus Newman	>64
E. coli ΔTolC	>64
E. coli DSM 1116	>64
M. smegmatis mc2-155	>64
P. aeruginosa PA14	>64
B. subtilis DSM10	>64
M. luteus DSM1790	>64
M. hiemalis DSM2656	>64
C. albicans DSM1665	>64
P. anomala DSM6766	>64
Chinese hamster ovary CHO-K1	>64

Since isolation yields of the other chloromyxamides B-E were very low, we were not able to subject these for MIC testing.

S 4.6.3.2   Comparison of synthetic and natural chloromyxamide A (1)
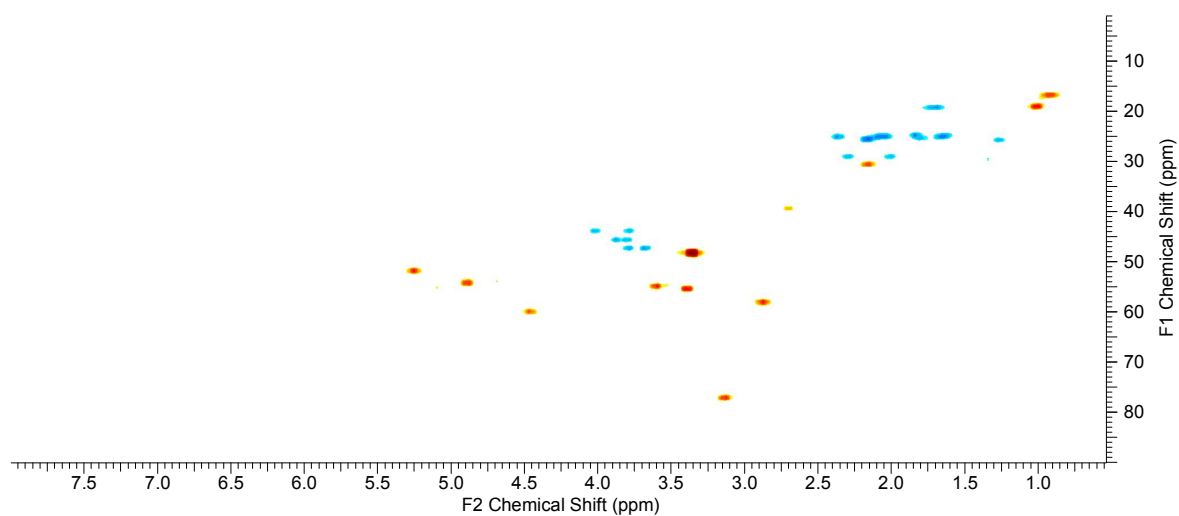
Figure S 18. *HSQC spectrum of natural **1**.*
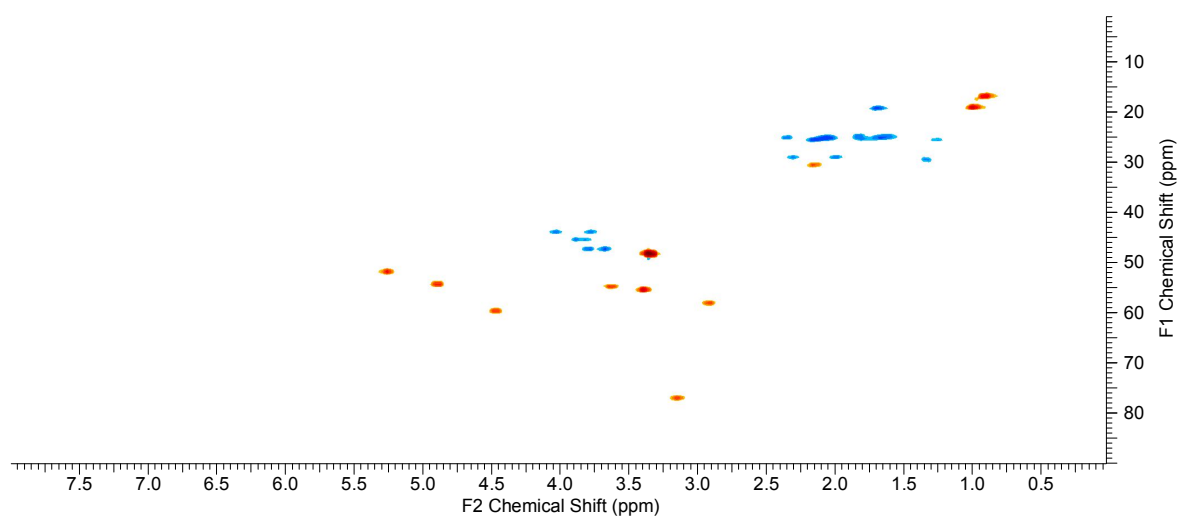
Figure S 19. *HSQC spectrum of synthetic **1**.*
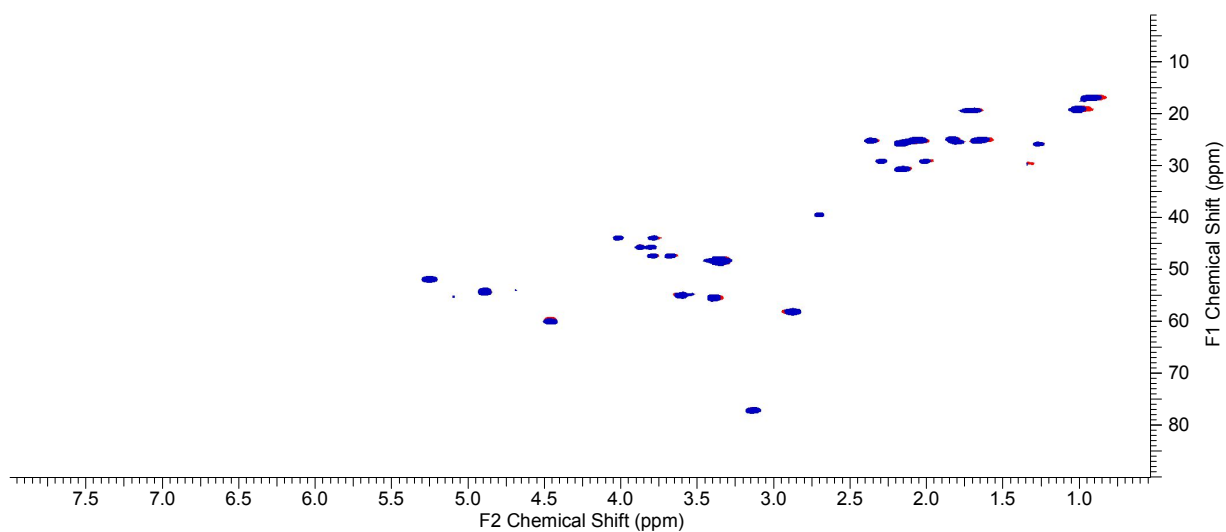
Figure S 20. *HSQC spectra overlay of synthetic **1** (red) and the isolated natural product (blue).*
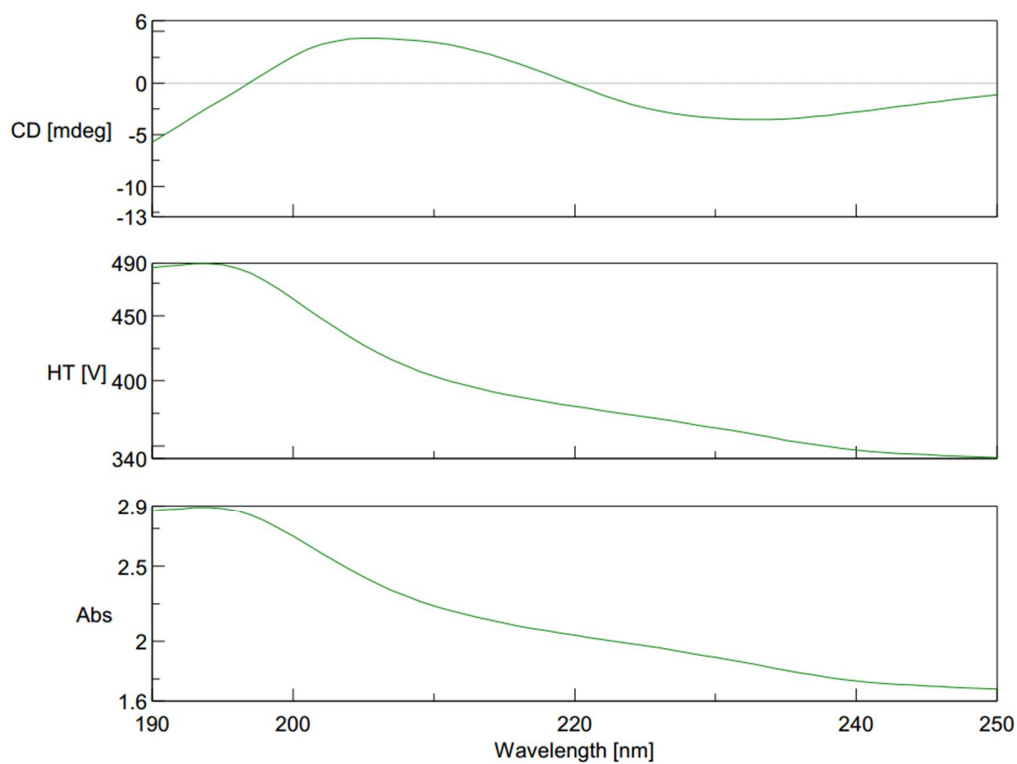
Figure S 21. *CD-Spectrum of **1 (synthetic chloromyxamide A)** at 0.1 mM in MeCN in the area 190-250 nm (with smoothing).*
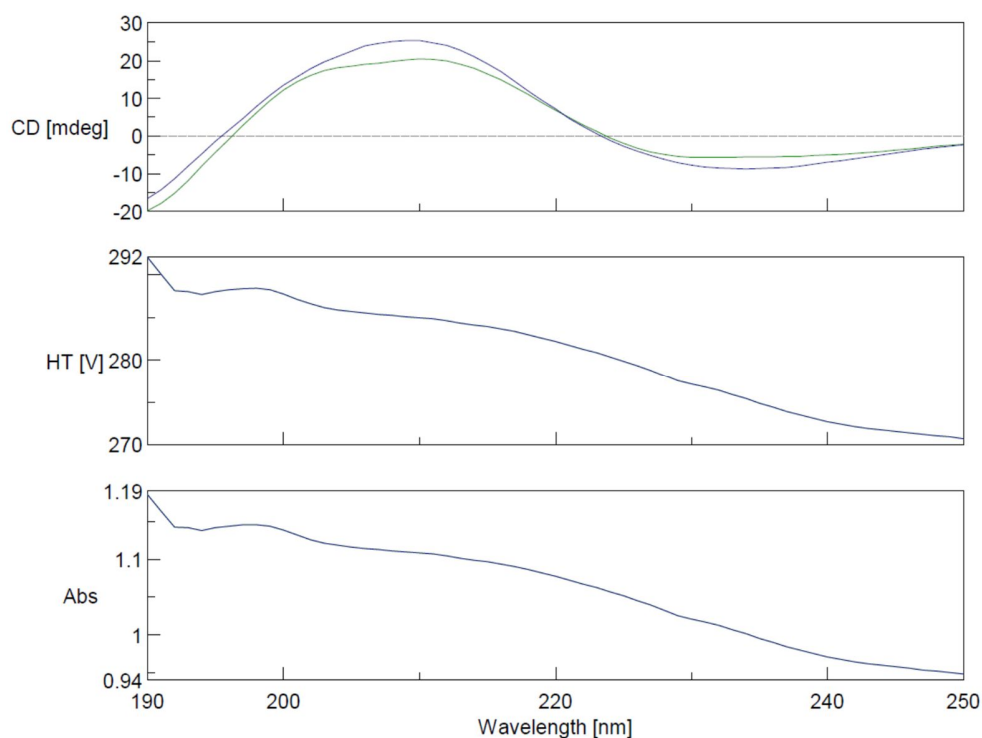
Figure S 22. *CD spectra of natural chloromyxamide A (blue) and B (green) at 0.1 mM in MeCN in the area 190-250 nm (with smoothing).*

S 4.7 References

[1] M. Röttig, M. H. Medema, K. Blin, T. Weber, C. Rausch, O. Kohlbacher, *Nucleic Acids Res.* **2011**, *39*, W362-W367.

[2] A. Sandmann, F. Sasse, R. Müller, *Chem. Biol.* **2004**, *11*, 1071.

[3] R. C. Edgar, *Nucleic Acids Res.* **2004**, *32*, 1792.

[4] R. O. Garcia, D. Krug, R. Müller in *Methods in Enzymology. Complex Enzymes in Microbial Natural Product Biosynthesis*. Part A: Overview Articles and Peptides (Ed.: David A. Hopwood), **2009**, pp. 59–91.

[5] M. A. Culpepper, E. E. Scott, J. Limburg, *Biochemistry* **2010**, *49*, 124.

[6] P. Lukat, Y. Katsuyama, S. C. Wenzel, T. Binz, C. König, W. Blankenfeldt, M. Brönstrup, R. Müller, *Chem. Sci.* **2017**, *8*, 7521.

[7] J. Mattay, S. Houwaart, W. Hüttel, *Appl. Environ. Microbiol.* **2018**.

[8] M. D. M. Raulfs, L. Breci, M. Bernier, O. M. Hamdy, A. Janiga, V. Wysocki, J. C. Poutsma, *Journal of the American Society for Mass Spectrometry* **2014**, *25*, 1705.

[9] C. B'Hymer, M. Montes-Bayon, J. A. Caruso, *J. Sep. Sci.* **2003**, *26*, 7.

[10] L. Kjaerulff, A. J. Benie, C. Hoeck, C. H. Gotfredsen, O. W. Sorensen, *J. Magn. Reson.* **2016**, *263*, 101.

Chapter 5

Pyxipyrrolones: Novel cytotoxic myxobacterial metabolites. Structure elucidation and biosynthesis proposal

Louise Kjaerulff[a] †, Ritesh Raju[a] †, Fabian Panter[a], Ullrich Scheid[a], Ronald Garcia[a,b], Jennifer Herrmann[a,b], and Rolf Müller*[a,b]

Affiliations

[a] Department Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmaceutical Biotechnology, Saarland University, Campus E8.1, 66123 Saarbrücken, Germany

[b] German Centre for Infection Research (DZIF), Partner Site Hannover-Braunschweig, D-38124, Braunschweig (Germany)

† These Authors contributed equally to the manuscript

Contributions and Acknowledgements

Author's effort:

The author designed and performed experiments, evaluated and interpreted resulting data. The author performed *in silico* analysis of the biosynthetic gene cluster and conceived a complete biosynthetic model for the Pyxipyrrolones. Furthermore, the author contributed significantly to conceiving and writing this manuscript.

Contributions by others:

Louise Kjaerulff significantly contributed to the conception of this study, designed and performed experiments, evaluated and interpreted resulting data. She re-isolated the compound and finalized the NMR based structure elucidation including assignment of all stereocenters except one and contributed significantly to conceiving and writing this manuscript. Ritesh Raju significantly contributed to the conception of this study, designed and performed experiments, evaluated and interpreted resulting data. He identified the pyxipyrrolones, performed the first compound isolation and contributed to NMR structure elucidation of the pyxipyrrolones. Ullrich Scheid identified the pyxipyrrolone biosynthetic gene cluster in the host strain's genome and performed *in silico* analysis of the gene clusters KS domains. Ronald Garcia isolated the *Pyxidicoccus fallax* strain MCy9557. Rolf Müller contributed by supervision of the project and conceiving, editing and proofreading of the manuscript.

5.1   Abstract

In our search for new secondary metabolites from myxobacteria, a strain from the genus *Pyxidicoccus* was investigated. This led to the identification of a new natural product class showing structural novelty and interesting biological activity. Isolation and structure elucidation of two analogs led to the identification of pyxipyrrolone A and B, harboring the novel 3-methylene-2,3,4,5,6,7,8,9-octahydro-1*H*-benzo[*e*]isoindol-1-one scaffold. Mosher's ester analysis combined with NMR studies allowed the determination of all stereocenters but one. Genome sequencing of the producer strain led to the identification of a putative biosynthetic gene cluster for the pyxipyrrolones. The compounds showed activity against several cancer cell lines (μM range) with pyxipyrrolone B having 2- to 11-fold higher activity than A, although they differ only by one methylene group.

5.2   Introduction

Microorganisms remain a highly prolific source of biologically active molecules for drug development. Novel chemical scaffolds exhibit the potential for unprecedented biological functions, and are therefore inherently interesting with respect to drug discovery.[1]

Myxobacteria are Gram-negative bacteria which are most noted for their large genomes and their ability to form fruiting bodies upon starvation. They have received considerable attention as a prominent source of novel secondary metabolites exhibiting new chemical scaffolds.[2] As part of our ongoing program involved in the screening and chemical profiling of myxobacterial strains, one such strain (MCy9557), identified as belonging to the under-explored genus *Pyxidicoccus* (section S2), was investigated.

5.3   Results and Discussion

HPLC-DAD-MS analysis of small scale cultivations (100 mL) revealed the presence of two analogs harboring a triene chromophore (λ_{max} 280 nm). Fractionation of repeated larger scale cultivations (10 L) permitted the isolation and structure elucidation of pyxipyrrolone A (1) and B (2) (Figure 1).

An account of the spectroscopic analysis leading to the assignment of 1 and 2 is presented below, followed by the description of a putative gene cluster for these novel molecules, and finally an initial characterization of their biological activities.

HRESIMS analysis of **1** revealed a protonated molecule at *m/z* 482.2897 [M+H]$^+$ (calcd for $C_{29}H_{40}NO_5$, 482.2901, $\Delta = 0.7$ ppm) indicative of a molecular formula $C_{29}H_{39}NO_5$ requiring eleven double bond equivalents.
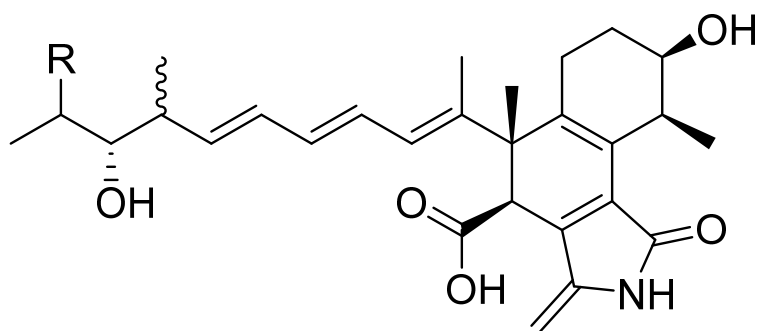
*Figure 1. Structure of pyxipyrrolone A (**1**, R = CH₃) and B (**2**, R = H).*

The NMR (methanol-d_4) data (Table S4) indicated resonances for two carbonyls (δ_C 173.8 and 171.3) and 12 sp^2 carbons (δ_C 94.7 – 143.3), requiring **1** to incorporate three rings. The ^{1}H NMR spectrum contains five olefinic protons (δ_H 5.57-6.31) and a singlet downfield methylene resonance (δ_H 4.89). Interpretation of the ^{1}H-^{1}H COSY correlations led to the identification of two isolated spin systems, with the first one delineating the sequence from the two secondary methyls H-29 and H-30 (δ_H 0.91 and 0.86), through the oxymethine H-27 (δ_H 3.07), and extending via the secondary methyl H-26 (δ_H 1.03) to five olefinic methines H-24 (δ_H 5.57) to H-20 (δ_H 5.90) (Figure 2).
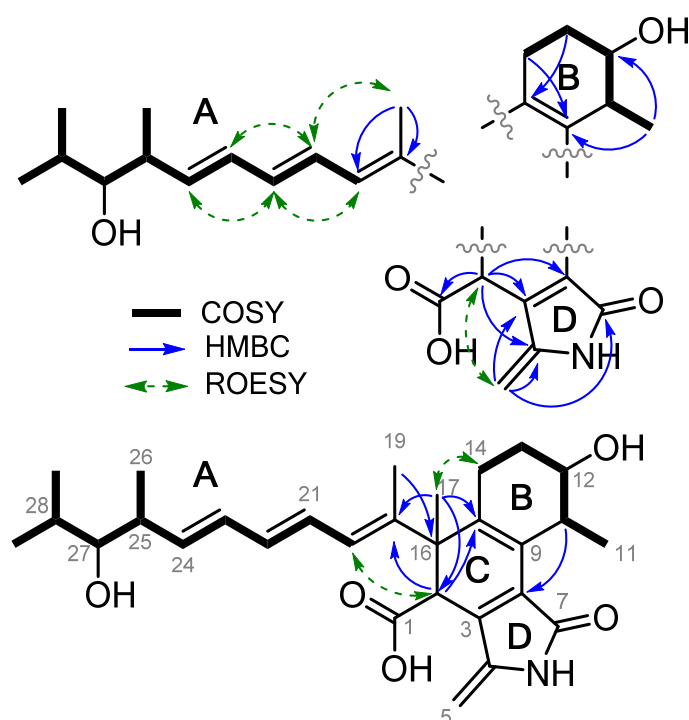
*Figure 2. Key 2D NMR correlations in methanol-d_4 used for the structure elucidation of pyxipyrrolone A (**1**).*

The second spin system includes two sets of heterotopic methylenes H-14 (δ_H 2.40, 2.02) and H-13 (δ_H 1.83, 1.73), continues through the oxymethine H-12 (δ_H 3.97), and terminates at the secondary methyl H-11 (δ_H 1.08). HMBC correlations from the methylene H-14 to the quaternary carbons C-9 (δ_C 131.6) and C-15 (δ_C 138.0), and from the secondary methyl H-11 to C-9, defined the structural fragment ring B, a 2-methylcyclohex-3-enol moiety (Figure 2). The triene fragment C-18 to C-24 is assigned *E* configuration based on two characteristically large coupling constants (J = 14.9 and 14.5 Hz, H-24/H-23 and H-22/H-21) along with a strong ROESY correlation from methyl H-19 (δ_H 1.76) to H-21 (δ_H 6.31). This structural fragment A is further extended by HMBC correlations from the methyl H-19 to C-16, C-18 and C-20. The methyl H-17 (δ_H 1.36) shows HMBC correlations to C-2, C-15, C-16 and C-18 which further extends structural fragment A. The downfield methine H-2 (δ_H 3.62) displays eight HMBC correlations to nearby non-protonated sp^2 carbons C-15, C-18, C-1, C-3, C-4, C-8, the quaternary C-16, as well as the tertiary methyl at C-17 which establishes the carbon connectivity between C-2 and C-16. The downfield methylene singlet resonance H-5 (δ_H 4.89, δ_C 94.8) correlates to several downfield non-protonated carbons C-3 (δ_C 143.3), C-4 (δ_C 140.1), and C-7 (δ_C 171.3). A long-range HMBC correlation from H-5 to C-9 (δ_C 131.6) is also observed through the conjugated π-system. The characteristic methylene group at C-5 (δ_H 4.89, δ_C 94.8) must be very electron-rich to be shifted this far upfield, and the chemical shifts resemble those observed for dehydroalanine in e.g. the myxobacterial compound argyrin.[3] The methylene protons, however, show a strong HMBC correlation to another olefinic carbon atom C-3 (δ_C 143.3), which would not be consistent with dehydroalanine in a linear peptide chain, and a combination of HMBC and ^{15}N-HMBC correlations in DMSO-d_6 (^{15}N to C-3, C-4, C-7, and C-8) warrant the identification of a 5-methylene-1,5-dihydropyrrol-2-one ring (fragment D, Figure 2). The downfield sp^2 hybridized C-3 and C-8 are connected to the last ring (fragment C), also shown by correlations from H-2. Furthermore, a single ROESY correlation from H-5 to H-2 suggests spatial proximity between these protons as well as H-5 pointing away from other protons. C-8 is correlated to only H-2 and H-10 which suggests linking fragments B and C at C-9, resulting in a highly substituted cyclohexadiene ring (fragment C) connected to the B ring at C-9 and C-15 and to the D ring at C-3 and C-8. The key HMBC correlations which establish the connection of the structural fragments A and B through C-16 are (i) H-17 to C-15, C-16, and C-18 and (ii) H-2 to C-15, C-16, C-17, and C-18. The final missing piece of the puzzle is C-1 (δ_C 173.8) associated only with H-2 by a strong HMBC correlation, which suggests a carboxylic acid functionality here. The numerous HMBC correlations to H-2 support the multiple branching points around C-2 and neighboring carbon atoms. The acid functionality is further verified by the observation of strong [M-H-CO$_2$]$^-$ fragment ions in negative mode ESIMS as well as the analytical scale preparation of a methyl ester with decreased polarity, allowing extraction into EtOAc from aqueous

solution, whereas **1** remains predominantly in the aqueous phase (sat. $NaHCO_3$ solution). The chemical shifts in the D fragment are in good agreement with those of a similar synthetic building block (4-ethyl-3-methyl-5-methylene-1,5-dihydropyrrol-2-one) which further supports the assignment.[4]

By sequencing the genome of the producing myxobacterium and subsequent retrobiosynthetic analysis, a putative biosynthetic cluster was found to be responsible for pyxipyrrolone production which was characterized by the presence of a single adenylation domain that had serine specificity. Feeding studies with $[2,3,3-d_3]$-L-serine indeed showed incorporation with a mass shift of +2 Da. This is in accordance with the structure and suggests that a dehydration reaction takes place to form the exocyclic double bond, although it is unclear when this takes place. An analog named pyxipyrrolone B (**2**) at *m/z* 468.2743 $[M+H]^+$ (calcd for $C_{28}H_{38}NO_5$, 482.2745, $\Delta = 0.4$ ppm) indicative of a molecular formula $C_{28}H_{37}NO_5$, was identified and by 2D NMR spectroscopy it was shown to be missing one methyl group (C-30), and thus harbored an ethyl group in place of the isopropyl group in the chain end of **1**. This is the result of the incorporation of different starter units in the biosynthetic machinery (Figure 3). The absolute configuration of the two secondary alcohols of **1** (C-12 and C-27) were determined by Mosher's method (Figure S4.3a).[5] Through interpretation of 2D NMR data for the (12,27)-*bis*-Mosher's esters of **1** the ^1H $\Delta\delta_{SR}$ values were calculated (Figure S4.3b and section S4.3). This revealed the absolute configuration of both C-12 and C-27 as *R*. Taking advantage of this, the configurations of three other stereogenic centers were elucidated. Through inspection of ROESY correlations and *J* coupling constants, the relative configurations in the B-C rings in combination with the knowledge from the Mosher analysis yielded the configuration of the stereogenic centers at C-2, C-10, C-16 (see section S4). The configuration at C-25 was not determined due to the free rotation at this position. H-2 and methyl H-17 are also correlated, but since this has weaker intensity than to H-19, we assume that H-2 and methyl C-17 have *anti*-orientation. A long-range ROE between H-12 and H-20 suggests that the triene adopts a conformation where H-20 is below the B fragment, and this observation also supports the C-16 stereochemistry. Altogether, these analyses resulted in assignment of the overall novel structures of pyxipyrrolone A and B (Figure 1). The 3-methylene-2,3,4,5,6,7,8,9-octahydro-1*H*-benzo[*e*]isoindol-1-one scaffold is novel, and only structural fragments are comparable to known compounds. The A fragment (Figure 2) shows similarity to delaminomycins,[6] which has a side chain like pyxipyrrolone B, but with only two conjugated double bonds, and similar pyrrole moieties (ring D, Figure 2) are known from e.g. congoensin A [7] and pyranonigrin E,[8] for which a protease homologue was responsible for the *exo*-methylene formation through dehydration of serine.[9] AntiSMASH analysis of the MCy9557 genome led to identification of a trans-AT PKS-NRPS hybrid gene cluster putatively responsible for pyxipyrrolone biosynthesis (see Figure 3).[10]
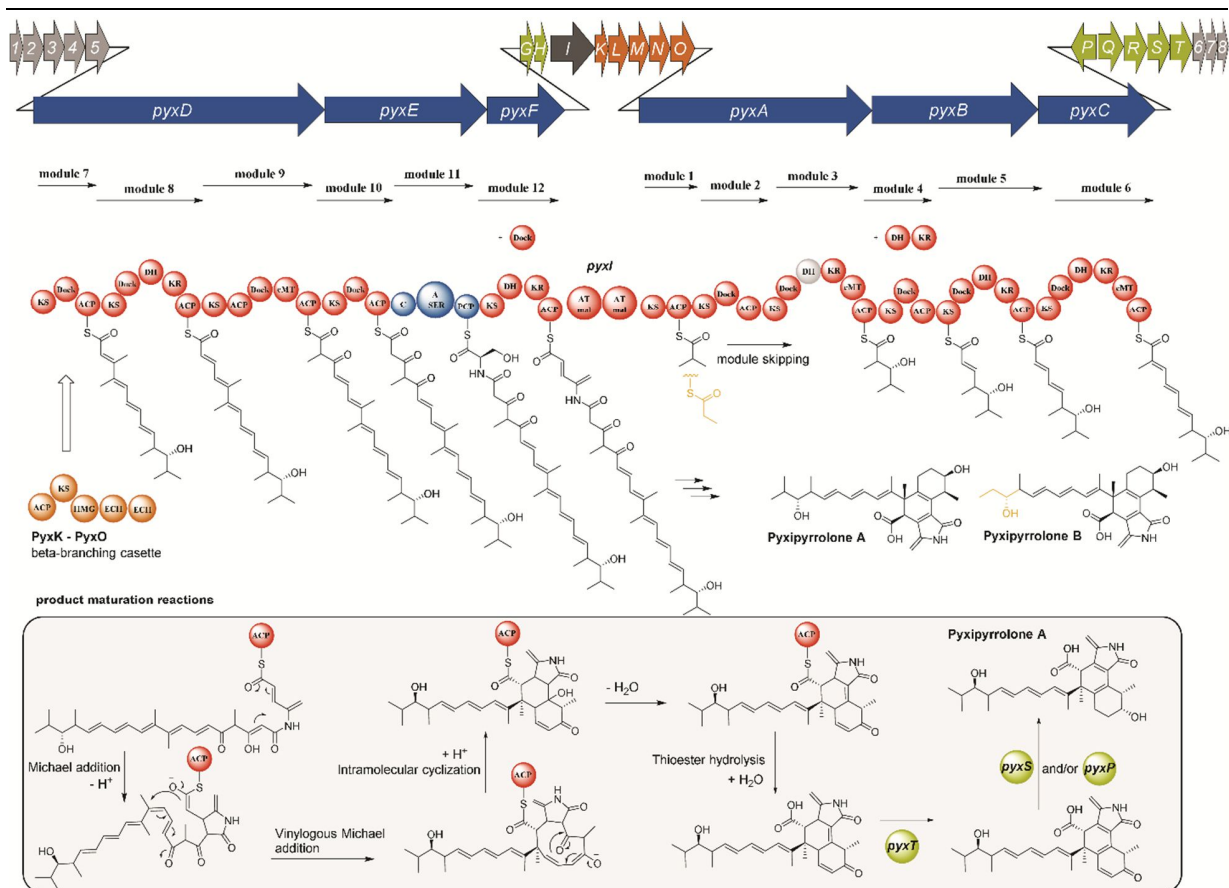
Figure 3. *Biosynthesis hypothesis for the pyxipyrrolones (A: adenylation domain; ACP: acyl carrier protein; AT: acyl transferase; C: condensation domain; cMT: carbon methyl transferase domain; DH: dehydratase domain; Dock: acyltransferase docking domain; ECH: enoyl-CoA hydratase; ER: enoyl reductase domain; HMG: hydroxymethylglutaryl-CoA synthase; KS: ketosynthase; mal: malonyl-CoA; PCP: peptidyl carrier protein; SER: serine; gene color code: blue: trans-AT PKS assembly line, black: in trans acyl transferase; orange: beta branching cassette; green: assigned and unassigned tailoring enzymes, respectively; gray: open reading frames in proximity to the pyxipyrrolone cluster). The box in gray shows the reaction cascade that might lead to synthesis of the mature product and the tailoring enzymes putatively involved in these transformations (see table S5.1).*

It contains eleven PKS modules and one NRPS module encoded on 6 genes (*pyxA-F*) as well as 13 additional genes encoded by *pyxG-T* (further info in section S5; GenBank accession number KY765914). The difference between the pyxipyrrolone analogs is the result of incorporation of different starter units. Module 2 seems to be skipped, as no malonyl-CoA incorporation for this module is seen in the final product, a behavior often observed for trans-AT PKS clusters, e.g. in leinamycin biosynthesis.[11] Module 3 contains an inactive dehydratase (DH) domain whereas module 4 has neither a keto reductase (KR) nor a DH domain that could explain the double bond at C-24/25 in the final product. We propose that the KR domain of module 3 acts twice during chain elongation in module 3 and 4, while the DH in module 3 acts only after ketosynthesis and ketoreduction in module 4. Not only would this explain the alcohol moiety at C-27, for which the KR stereochemistry predictor by Kitsche and Kalesse predicts the correct stereochemistry,[12] but also the double bond formed in module 4. The genes *pyxK-O* encode a β-branching

cassette which is likely to create the β-methyl branch in module 7 as it was described for myxovirescin.[13] The dehydroalanine moiety is created from a serine moiety incorporated in module 11, correctly predicted by Stachelhaus code based NRPS predictor 2[14] and confirmed by an isotopic labeling experiment (see above) although no candidate enzyme for this transformation is found in the cluster. A Bayesian KS domain alignment for trans-AT PKS structure prediction further validates this biosynthesis hypothesis as the predicted building blocks fit the final pyxipyrrolone structure (see section S5).[15] A series of cyclization and maturation steps occur to form the polycyclic end product. The linear chain holds a β-keto amide moiety that is likely to be partially enolized (C7/8) and thus might act as Michael donor in a cyclization event. Additionally, the α,β-unsaturated carboxylic acid at C-2/3 as well as the vinylogous system (C-12 to C-16) could act as Michael acceptors and therefore it is likely that the polyketide chain is pre-oriented and two subsequent Michael additions lead to the formation of the tricyclic ring system. There are alternative possibilities to explain the cyclization reactions, e.g. radical-based mechanisms, but based on the gene cluster and chemical considerations the reaction cascade presented here seems most likely. The C-10 stereocenter is probably created during the second ring formation and not during C methyl transfer in module 9, as it is part of a 1,3-diketo functionality that racemizes fast and the final product is solely *S*-configured. The final biosynthetic steps are probably reduction at C-12 to C-14 and dehydrogenation to form the second double bond in the C ring. The enzymes PyxT, -S, and -P are believed to perform the remaining redox steps needed (see section S5). Finally, the enzyme PyxR shows homology to myxobacterial ABC transporters and is therefore likely to be involved in compound export. Unfortunately and despite serious efforts we were unable to mutagenize MCy9557, and thus this biosynthesis hypothesis is based on *in silico* analysis.

Table 1. *Half-inhibitory concentrations (IC$_{50}$ in µM) for pyxipyrrolone A (**1**) and B (**2**) on a panel of cell lines.*

Cell line	1	2
A549 (human lung carcinoma)	> 200	190.3
HCT-116 (human colon carcinoma)	68.3	12.6
HEC-1-BG (human endometrial adenocarcinoma)	71.4	46.0
L929 (murine subcutaneous connective tissue fibroblasts)	116.1	10.3
SW480 (human colon adenocarcinoma)	34.3	7.9
THP1 (human acute monocytic leukemia)	45.3	8.3

Pyxipyrrolone B (**2**) displayed cytotoxic activity in the micromolar range against a panel of cell lines. In contrast, **1** showed much lower activity, which highlights the importance of the side chain for biological activity. In general, human colon and leukemia cells and murine fibroblasts were most sensitive towards

treatment with pyxipyrrolones. SW480 cells were used for microscopic studies to get first insights into the mechanism of action. Following treatment with the approximately 10-fold IC_{50} concentration of **2** for 1-4 days, the cells showed typical morphological changes, such as cellular enlargement and rounded-up (apoptotic) cells. Furthermore, staining with acridine orange revealed lysosomal membrane permeabilization (increased lysosomal pH), which is a hallmark of the lysosomal death pathway (Figure 4). Pyxipyrrolones **1** and **2** were inactive against most of the tested Gram-positive and Gram-negative bacteria, yeast and molds. However, **2** displayed some weak inhibitory effect (MIC 32-64 µg/mL) on *Mucor hiemalis* and *Staphylococcus aureus* (see Table S6 for details).
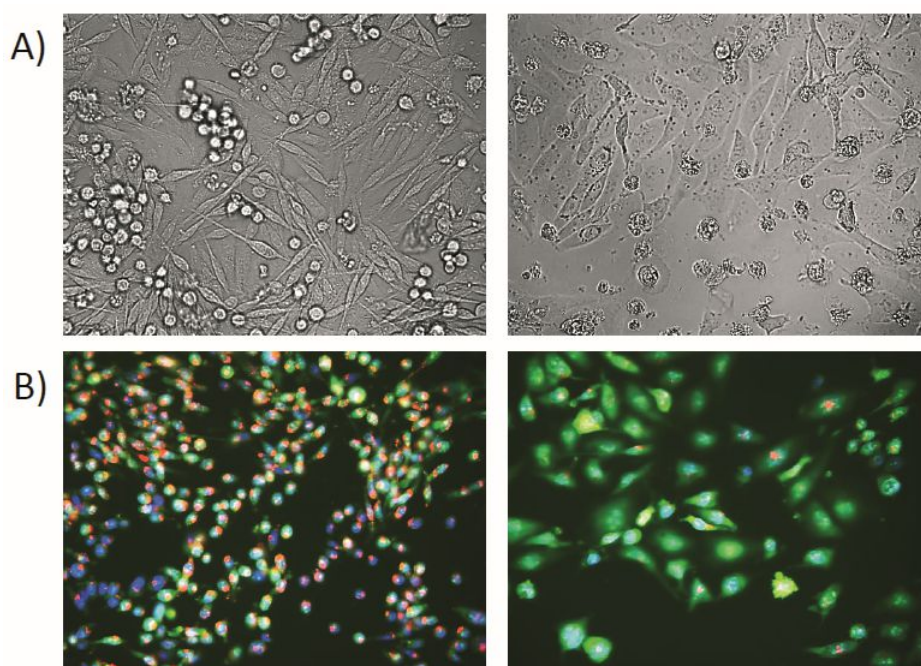
Figure 4. *Effects on SW480 cells after 4 days treatment with **2**. A) Light microscopy images of control (left) and treated cells (right). B) Fluorescence microscopy images of control (left) and treated cells (right) that were stained with acridine orange (red: acidic compartments, green: neutral compartments) and Hoechst33342 (blue: nuclei).*

5.4   Conclusion and Outlook

In conclusion, the findings summarized in this communication highlights myxobacteria as producers of novel molecules with attractive biological activities. We identified the pyxipyrrolones as new metabolites with unique chemical structures and report a first glimpse of their biological potential. Intriguingly the small difference of the methyl group at C-28 gives a remarkable change regarding biological activity. To fully characterize these structures and further investigate the biological role of the pyxipyrrolones, total synthesis should be considered, despite the inevitable challenges to be expected with this novel scaffold.

5.5   Acknowledgments

5.6   References

[1] a) D. J. Newman, *J. Med. Chem.* **2008**, *51*, 2589; b) G. M. Cragg, D. J. Newman, *Biochim. Biophys. Acta* **2013**, *1830*, 3670.

[2] J. Herrmann, A. A. Fayad, R. Müller, *Nat. Prod. Rep.* **2017**, *34*, 135.

[3] L. Vollbrecht, H. Steinmetz, G. Höfle, L. Oberer, G. Rihs, G. Bovermann, P. von Matt, *J. Antibiot.* **2002**, *55*, 715.

[4] P. Kancharla, W. Lu, S. M. Salem, J. X. Kelly, K. A. Reynolds, *J. Org. Chem.* **2014**, *79*, 11674.

[5] J. A. Dale, H. S. Mosher, *J. Am. Chem. Soc.* **1973**, *95*, 512.

[6] M. Ueno, T. Someno, R. Sawa, H. Iinuma, H. Naganawa, M. Ishizuka, T. Takeuchi, *J. Antibiot.* **1993**, *46*, 979.

[7] G. M. Happi, S. F. Kouam, F. M. Talontsi, S. Zuhlke, M. Lamshoft, M. Spiteller, *Fitoterapia* **2015**, *102*, 35.

[8] Y. Miyake, C. Ito, M. Itoigawa, T. Osawa, *Biosci. Biotechnol. Biochem.* **2007**, *71*, 2515.

[9] T. Yamamoto, Y. Tsunematsu, H. Noguchi, K. Hotta, K. Watanabe, *Org. Lett.* **2015**, *17*, 4992.

[10] T. Weber, K. Blin, S. Duddela, D. Krug, H. U. Kim, R. Bruccoleri, S. Y. Lee, M. A. Fischbach, R. Müller, W. Wohlleben et al., *Nucleic Acids Res.* **2015**, *43*, W237-W243.

[11] G. L. Tang, Y. Q. Cheng, B. Shen, *J. Nat. Prod.* **2006**, *69*, 387.

[12] A. Kitsche, M. Kalesse, *ChemBioChem* **2013**, *14*, 851.

[13] C. T. Calderone, D. F. Iwig, P. C. Dorrestein, N. L. Kelleher, C. T. Walsh, *Chem. Biol.* **2007**, *14*, 835.

[14] M. Rottig, M. H. Medema, K. Blin, T. Weber, C. Rausch, O. Kohlbacher, *Nucleic Acids Res.* **2011**, *39*, W362-7.

[15] T. Nguyen, K. Ishida, H. Jenke-Kodama, E. Dittmann, C. Gurgui, T. Hochmuth, S. Taudien, M. Platzer, C. Hertweck, J. Piel, *Nat. Biotechnol.* **2008**, *26*, 225.

Supporting Information

Pyxipyrrolones: Novel cytotoxic myxobacterial metabolites. Structure elucidation and biosynthesis proposal

Louise Kjaerulff[a] †, Ritesh Raju[a] †, Fabian Panter[a], Ullrich Scheid[a], Ronald Garcia[a,b], Jennifer Herrmann[a,b], and Rolf Müller*[a,b]

Affiliations

[a] Department Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmaceutical Biotechnology, Saarland University, Campus E8.1, 66123 Saarbrücken, Germany

[b] German Centre for Infection Research (DZIF), Partner Site Hannover-Braunschweig, D-38124, Braunschweig (Germany)

† These Authors contributed equally to the manuscript

The printed version of the Supporting Information does not contain data that cannot be visualized satisfactorily on paper. To access this data please refer to the enclosed storage medium or the on-line version of this research paper.

S 5.1 General experimentals

Chiroptical measurements ($[\alpha]_D$) were obtained on a Perkin Elmer (Model 341) polarimeter in a 100 × 2 mm cell at 20 °C. NMR spectra were acquired on a Bruker Ascend 700 MHz and a Bruker UltraShield 500 MHz equipped with 5 mm cryoprobes using standard pulse sequences. The signals of the residual solvent were used as internal reference (δ_H 3.31 and δ_C 49.0 ppm for methanol-d_4). LC-HRESI-DAD-MS were acquired on a Bruker maXis 4G mass spectrometer or a Bruker Amazon ion trap instrument, both coupled with a Dionex Ultimate 3000 RSLC system using a BEH C_{18} column (100 × 2.1 mm, 1.7 µm, Waters, Germany) with a gradient of 5-95% MeCN + 0.1% FA in H_2O + 0.1% FA at 0.6 ml/min and 45 °C over 18 min or a BEH C_{18} column (50 × 2.1 mm, 1.7 µm, Waters, Germany) with a gradient of 5-95% MeCN + 0.1% FA in H_2O + 0.1% FA at 0.6 ml/min and 45 °C over 9 min, with UV detection by a diode array detector at 200-600 nm. Mass spectra were acquired from 150 to 2000 m/z at 2 Hz.

S 5.2 Isolation and Identification of MCy9557

Pyxidicoccus sp. MCy9557 was isolated in 2010 from a soil sample taken at the Saarland University campus, Saarbrücken, Germany. The strain was recognized for fruiting bodies on a standard isolation medium containing bacterial prey.[1,1,2] After successive transfers of the colony swarm edge, the myxobacterium was finally purified. The identification was performed by morphology characterization of growth stages on buffered VY/2 agar,[3] and together by molecular phylogenetics using 16S rRNA gene sequencing.

>MCy9557 complete 16S rRNA gene sequence

AGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGCGTGCCTAACACATGCAAGTCGAGCGCGAATAGGGGCAAC
CCTTAGTAGAGCGGCGCACGGGTGCGTAACACGTGGATAATCTGCCTGGATGCCCGGGATAACCAGTCGAAAGAT
TGGCTAATACCGGATAAGCCCACGGCCTCTTCGGAGACTGAGGGAAAAGGTGGCCTCTGTATACAAGCTATCACA
ACCAGATGAGTCCGCGGCCCATCAGCTAGTTGGCGGGGTAATGGCCCACCAAGGCGACGACGGGTAGCTGGTCT
GAGAGGACGATCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATTTTGC
GCAATGGGCGAAAGCCTGACGCAGCAACGCCGCGTGTGTGATGAAGGTCTTCGGATTGTAAAGCACTTTCGACCG
GGACGAAAACCCGTAGCCCAACACGCTACGGCTTGACGGTACCGGGAGAAGAAGCACCGGCTAACTCTGTGCCAG
CAGCCGCGGTAATACAGAGGGTGCAAGCGTTGTTCGGAATTATTGGGCGTAAAGCGCGTGTAGGCGGCGTGACA
AGTCGGGTGTGAAAGCCCTCAGCTCAACTGAGGAAGTGCGCCCGAAACTGTCGTGCTTGAGTGCCGGAGAGGGT
GGCGGAATTCCCCAAGTAGAGGTGAAATTCGTAGATATGGGGAGGAACACCGGTGGCGAAGGCGGCCACCTGGA
CGGTAACTGACGCTGAGACGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAAC

GATGAGAACTAGGTGTCGTGGGAGTTGACCCCCGCGGTGCCGAAGCTAACGCATTAAGTTCTCCGCCTGGGAAGT
ACGGTCGCAAGACTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGACG
CAACGCGCAGAACCTTACCTGGTCTTGACATCCTTGGAATCCCTCAGAGATGAGGGAGTGCCCGCAAGGGAACCA
AGAGACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCC
TCGCCTTTAGTTGCCACGCAAGTGGATCTCTAGAGGGACTGCCGGTGTTAAACCGGAGGAAGGTGGGGATGACGT
CAAGTCCTCATGGCCTTTATGACCAGGGCTACACACGTGCTACAATGGCCGGTACAGAGCGTTGCCAACCCGCGAG
GGGGAGCTAATCGCATAAAACCGGTCTCAGTTCAGATTGGAGTCTGCAACTCGACTCCATGAAGGCGGAATCGCT
ATTAATCGCAGATCAGCACGCTGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTC
GATTGCTCCAGAAGTCATCTCACCAAGAGATGCCCAAGGAGTGGTCGGTAACTGGGGTGAAGTCGTAACAAGGTA
GCCGTAGGGGAACCTGCGGCTGGATCACCT

S 5.2.1   Extraction Protocol for Pyxidicoccus sp. SBCy002 (MCy9557) and compound optimization

For cultivation of MCy9557, the strain was grown in 10 L TS-6 medium[1] containing 2% amberlite resin XAD-16 (Sigma) and seeded with 5% bacterial culture. The inoculum was prepared by cultivating the strain in the same medium for 4 days at 180 rpm and 30 °C. The culture was grown for 10-12 days at 30 °C, 160 rpm. At the end of fermentation, both resin and cells were harvested together by centrifugation (8000 rpm, 4 °C, 30 min).

S 5.3 Isolation of pyxipyrrolone A (1) and B (2)

The adsorber resin XAD-16 and cell pellet were extracted four times with 300 ml methanol. The crude extract was filtered, concentrated *in vacuo*, and subjected to solvent partitioning with 300 ml 90% MeOH in H_2O and 100 ml heptane. The polar fraction was concentrated *in vacuo* and partitioned between 150 ml H_2O and 250 ml EtOAc and the aqueous phase was further washed with 100 ml EtOAc to yield 0.7 g of organic extract containing pyxipyrrolones. This was wet loaded on C_{18} and dried under nitrogen before it was fractionated on C_{18} (120 g SNAP, Biotage). After equilibration with 400 ml of 10% MeOH/H_2O, a multistep gradient was run at 50 ml/min (10% MeOH for 5 min, then 10-90% MeOH over 26 min and 90% MeOH for 10 min). Fractions of 20 ml were collected, and pyxipyrrolones eluted after about 30 min over seven fractions which were pooled to give 22 mg of pyxipyrrolone-containing extract. Final purification by RP-HPLC (Luna II C_{18}, Phenomenex, 250 × 10 mm, 100 Å) using H_2O/MeCN with 0.1% formic acid (25-50% MeCN over 40 min, 4 ml/min, 45 °C) yielded pure pyxipyrrolone B (29 min, 0.5 mg) and pyxipyrrolone A (34 min, 1.1 mg).

Several 10 L cultures were produced to obtain sufficient material for analyses and derivatizations.

Pyxipyrrolone A (1): weakly yellow amorphous solid; $[\alpha]_D - 270$ (*c* 0.1, MeOH); UV (MeOH) λ_{max} 222, 270, 281, 292, 350 nm; NMR (500 MHz, methanol-d_4) see Table S4 and section 7.1; HRESIMS *m/z* 482.2897 [M+H]$^+$ (calcd for $C_{29}H_{40}NO_5$, 482.2901, Δ = 0.8 ppm).

Pyxipyrrolone B (2): weakly yellow amorphous solid; $[\alpha]_D - 500$ (*c* 0.1, MeOH); UV as **1**; NMR (500 MHz, methanol-d_4) see Table S4 and section 7.2; HRESIMS *m/z* 468.2743 [M+H]$^+$ (calcd for $C_{28}H_{38}NO_5$, 468.2745, Δ = 0.4 ppm).

S 5.4 Structure elucidation of pyxipyrrolones
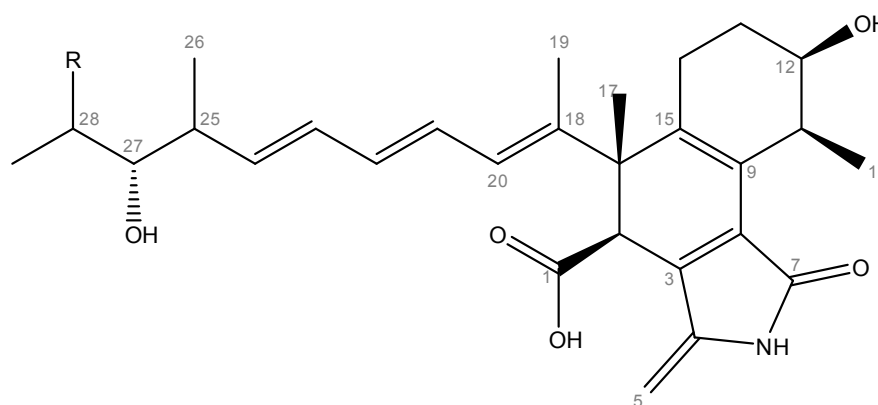
Figure S4a. *Atom numbering for pyxipyrrolone A (R = CH₃) and B (R = H).*

Table S4. *NMR data for pyxipyrrolone A (1) and B (2) at 700 MHz in methanol-d₄.*

Pyxipyrrolone A (1)				
#	δ_H, multiplicity, (*J* in Hz)	δ_C	HMBC	ROESY
1		173.8		
2	3.62, 1H, s	47.8	1, 3, 4, 8, 15, 16, 17, 18	5, 17, 19, 20
3		143.3*		
4		140.1*		
5	4.89, 2H, s	94.7	3, 4, (7), (9)	2
6-N				
7		171.3		
8		130.6		
9		131.6		
10	3.57, 1H, m	34.2	(8), 9, 11, 12, 13, 15	
11	1.08, 3H, d (6.8)	14.1	9, 10, 12	13a
12	3.97, 1H, ddd (10.9, 5.1, 3.6)	69.9	11, (13)	14b, 20
13a	1.83, 1H, m	26.4	10, 12, 14, 15	11
13b	1.73, 1H, m	26.4		
14a	2.40, 1H, m	26.5	9, 12, 13, 15	17
14b	2.02, 1H, m	26.5	9, 13, 15	12, 20
15		138.0		
16		48.6		
17	1.36, 3H, s	21.2	2, (14), 15, 16, 18	2, 14a, 19, 20

#	δH, multiplicity, (J in Hz)	δC	HMBC	ROESY
18		138.4		
19	1.76, 3H, s	12.8	16, 18, 20	2, 17, 21
20	5.90, 1H, d (11.0)	126.3	16, 19, 22	2, 12, 14b, 17, 22
21	6.31, 1H, dd (14.5, 11.0)	128.3	18, 20, 23	19, 23
22	6.03, 1H, dd (14.5, 10.6)	134.1	20, 24	20, 24
23	6.10, 1H, dd (14.9, 10.6)	131.1	21, 25	21, 25
24	5.57, 1H, dd (14.9, 8.6)	138.6	22, 25, 26, 27	22, 26, 27
25	2.31, 1H, m	41.8	23, 24, 26, 27	23, 29, 30
26	1.03, 3H, d (6.6)	16.2	24, 25, 27	24, 27
27	3.07, 1H, dd (7.0, 4.8)	80.8	24, 25, 26/30, 28, 29	24, 26, 29, 30
28	1.71, 1H, m	31.7	27, 29, 30	
29	0.91, 3H, d (6.8)	20.3	27, 28, 30	25, 27
30	0.86, 3H, d (6.7)	16.2	27, 28, 29	25, 27

Pyxipyrrolone B (2)

#	δH, multiplicity, (J in Hz)	δC	HMBC	ROESY
1		175.0		
2	3.59, 1H, s	48.9	1, 3, 4, 8, 9, 15, 16, 17, 18	5, 17, 19, 20
3		143.5ª		
4		141.2ª		
5	4.89ᵇ, 2H	94.6	3, 4, (7)	2
6-N				
7		171.6		
8		130.1		
9		131.2		
10	3.54, 1H, m	34.2	(8), 9, 11, 12, 13, 15	13b
11	1.10, 3H, d (6.8)	14.2	9, 10, 12	13a
12	3.96, 1H, ddd (10.7, 5.3, 3.5)	70.0	11	14b, 20
13a	1.84, 1H, m	26.7	(12)	11
13b	1.71, 1H, m	26.7		10
14a	2.37, 1H, m	26.2	(9)	17
14b	1.99, 1H, m	26.2	(9), (15)	12, 20
15		137.9		
16		48.7		
17	1.36, 3H, s	21.1	2, 15, 16, 18	2, 14a, 19, 20
18		138.9		
19	1.77, 3H, s	12.8	16, 18, 20	2, 17, 21, 22
20	5.89, 1H, d (11.0)	126.1	16, 19, 22	2, 12, 14b, 17
21	6.31, 1H, dd (14.3, 11.0)	128.6	18, 20, 23	19, 24
22	6.02, 1H, dd (14.3, 10.6)	133.8	20, 24	19
23	6.10, 1H, dd (14.8, 10.6)	131.6	21, 25	25, 26
24	5.59, 1H, dd (14.8, 8.4)	137.9	22, 25, 26, 27	21, 26, 27
25	2.21, 1H, m	44.2	23, 24, 26, 27, 28	23, 28ab
26	1.02, 3H, d (6.8)	16.1	24, 25, 27	23, 27
27	3.25, 1H, m	77.6	(24), 26, 29	24, 26, 29
28a	1.53, 1H, m	28.2	25, 27, 29	24, 25
28b	1.31, 1H, m	28.2	25, 27, 29	25
29	0.93, 3H, t (7.4)	10.5	27, 28	27

* The assignment of C3 and C4 may be interchanged.
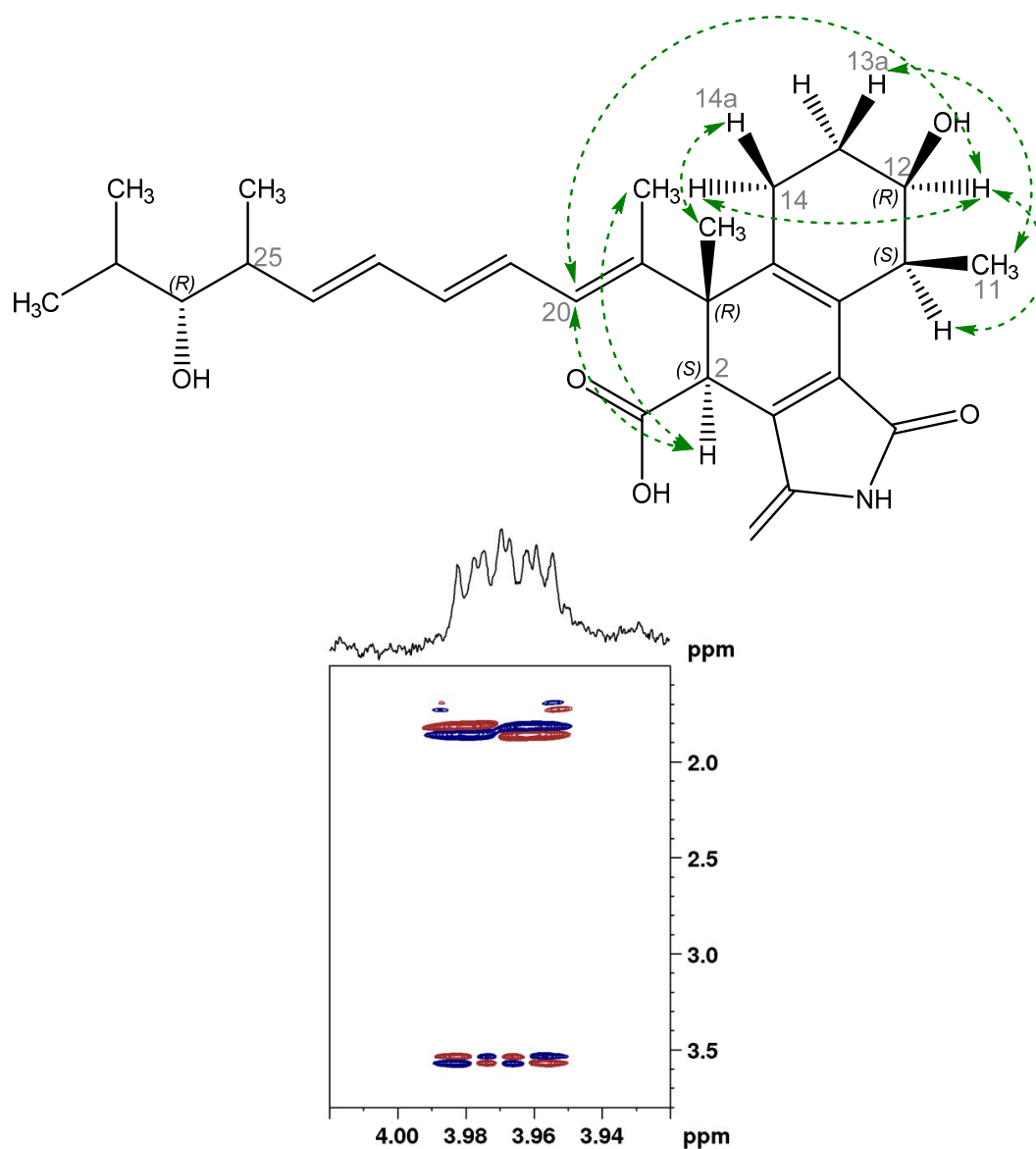
Figure S4b. *ROESY and DQF-COSY correlations for H-12 used to find the absolute stereochemistry of C-2, C-10, and C-16.*

From DQF-COSY NMR data it is evident that the large coupling constant for H-12 is to H-13a, and a smaller coupling constant is seen to H-11 and the correlation to H-13b is very weak, in good agreement with the trans-relationship from H-12 to H-13a and cis-relationship to H-11 and H-13b supported by the ROESY correlations shown. A correlation from H-12 to H-20 establishes the triene-chain as pointing down. Strong ROE correlations from H-2 to H_3-19 and H-20 firmly establish the stereochemical relationship between C-2 and C-16 as shown. This is supported by a ROE correlation between H_3-17 (pointing up) and H-14a as well as the correlation between H-14b and H-12 (pointing down).

S 5.4.1   Analytical scale synthesis of the methyl ester of 1

15 µg of **1** (1 µg/µl) was reacted with TMS-diazomethane in MeOH (2.1 µl of 2 M TMS-diazomethane in ether diluted with 200 µl MeOH, then 5 µl of diluted reagent added), and although the conversion was not complete, the produced methyl ester of pyxipyrrolone A could be analyzed by LCMS.
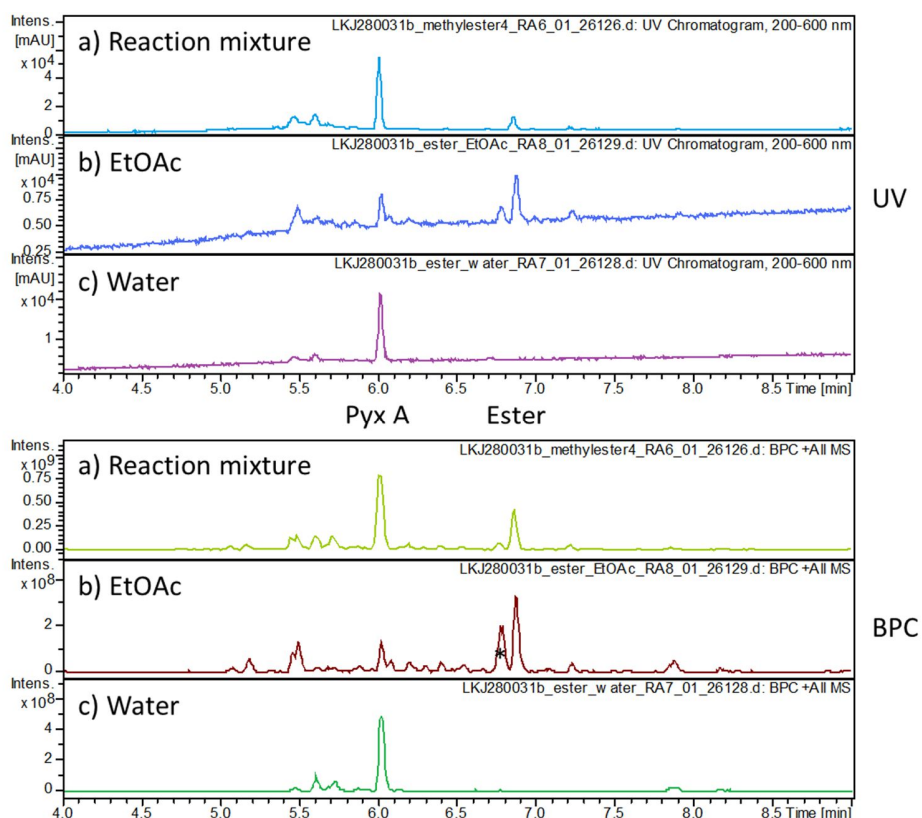
Figure S4.1a. *LCMS after esterification reaction.*

LC-UV chromatograms (top) and BPCs (bottom) of a) the esterification reaction mixture before liquid-liquid extraction, b) the EtOAc extract redissolved in MeOH containing the methyl ester (6.9 min) and a bit of pyxipyrrolone A (6.0 min), and c) the NaHCO3-water phase containing primarily pyxipyrrolone A (6.0 min). The fact that pyxipyrrolone A can be extracted with weak base in water supports the presence of a carboxylic acid moiety in the structure. The peak at 6.8 min (before the ester) is an unrelated impurity (m/z 415).
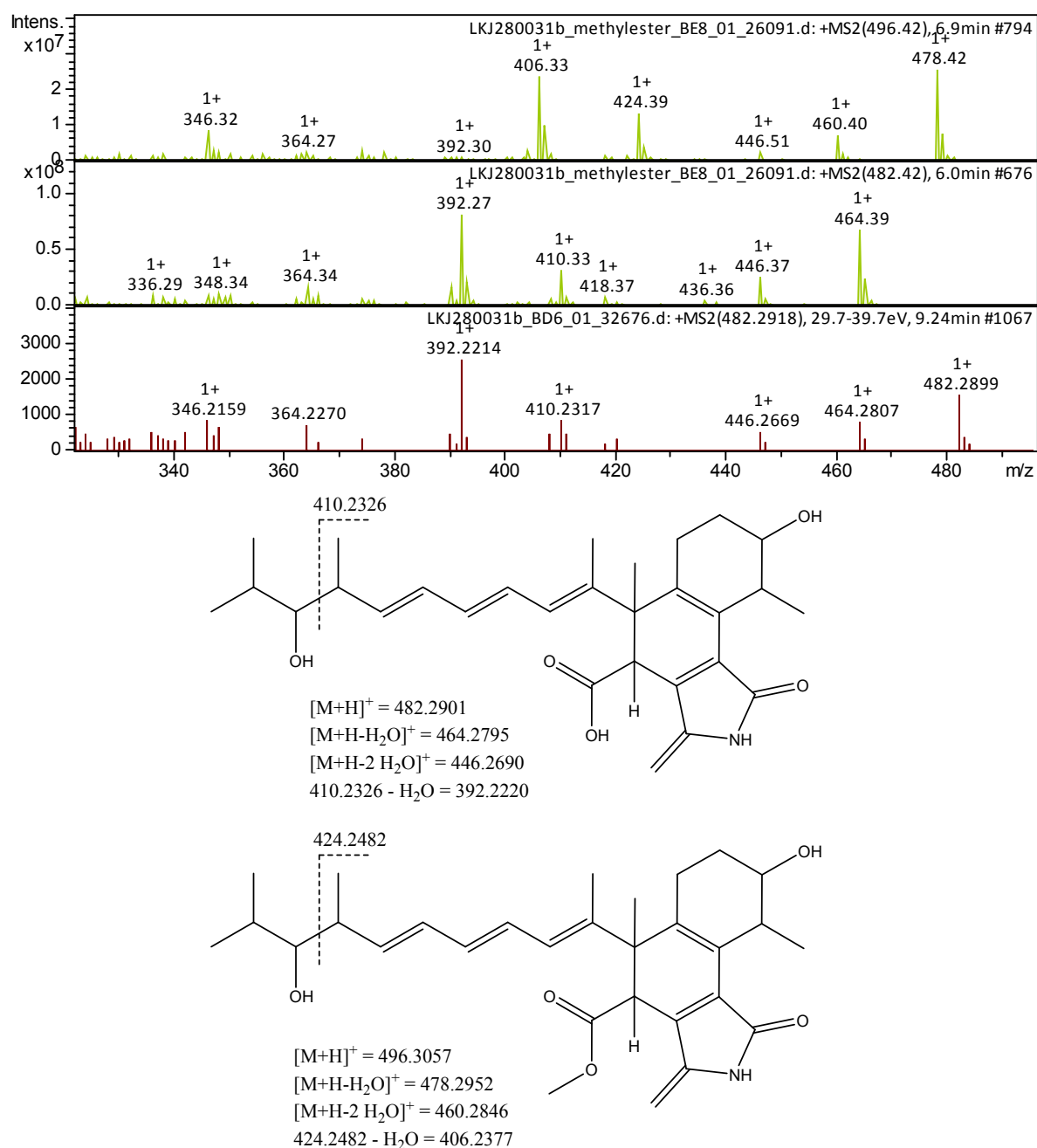
Figure S4.1b. Positive-mode MS2 spectra and interpretation of fragmentation patterns for pyxipyrrolone A and its methyl ester.

The bottom spectrum is HR-MS^2 (maXis) of pyxipyrrolone A, whereas the two upper MS^2 spectra (Amazon) are from after the esterification reaction, showing the ester (top) and unreacted pyxipyrrolone A with a relative mass shift of 14 Da.
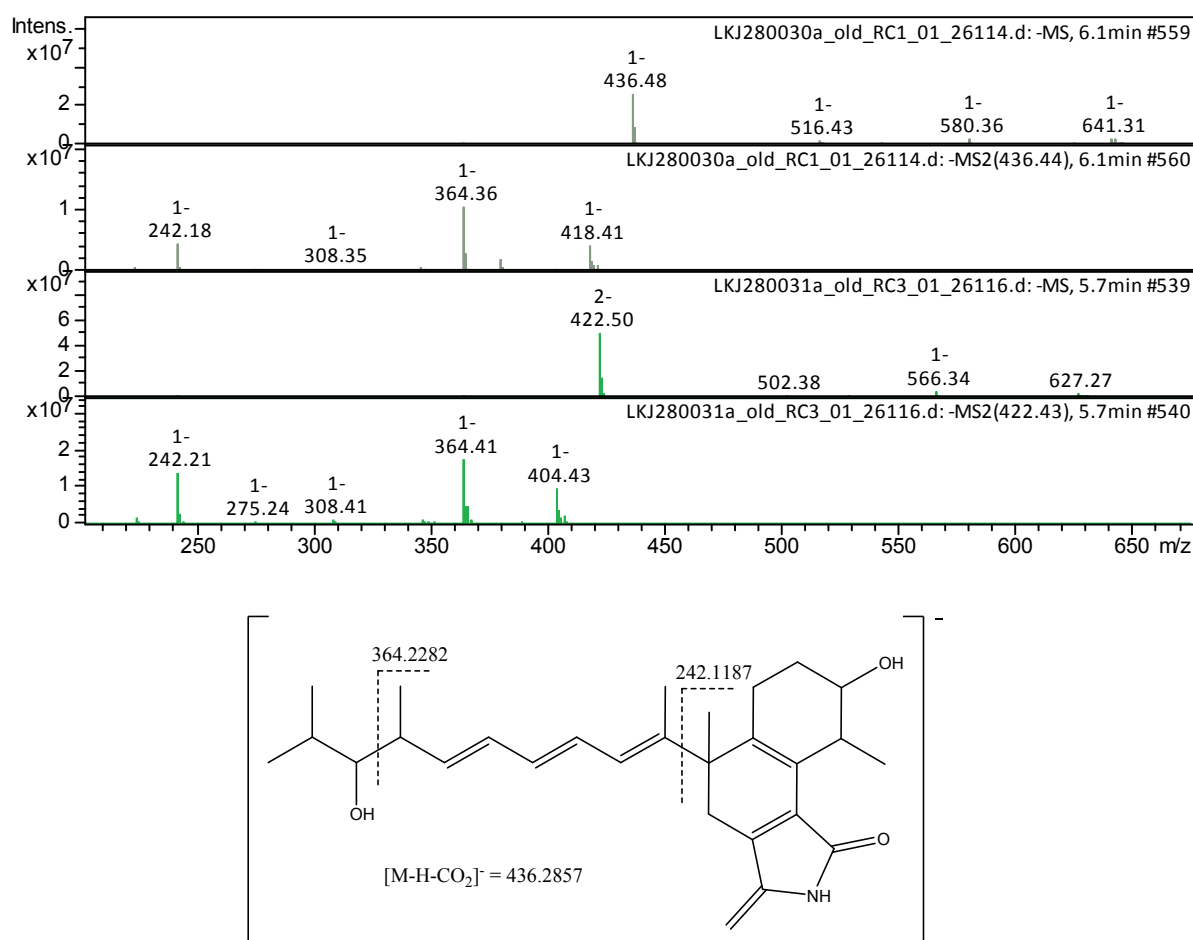
Figure S4.1c. *Negative-mode ESI-MS and MS2 spectra of pyxipyrrolone A (grey, top) and pyxipyrrolone B (green, bottom), showing a prominent decarboxylation fragment (m/z 436/422), then loss of water. The low molecular weight fragments are the same for pyxipyrrolone A and B because they involve fragmentation of the chain (see structure of decarboxylated pyxipyrrolone A).*

S 5.4.2   Isotopically labelled feeding with d_3-Ser

Two 50 ml cultures of MCy9557 were grown in 2SWT medium (Bacto tryptone 0.3%, Bacto soytone 0.1%, glucose 0.2%, soluble starch 0.2%, maltose monohydrate 0.1%, cellobiose 0.2%, $CaCl_2·2H_2O$ 0.05%, $MgSO_4·7H_2O$ 0.1%, HEPES 10 mM, pH adjusted to 7.0 with KOH before autoclaving) and for four consecutive days 50 µl of a 0.1 M solution of [2,3,3-d_3]-L-serine in sterile water was added to one of them (the other used as control). After 10 days at 30 °C and 160 rpm, the cultures were harvested by centrifugation at 8000 rpm for 5 min, the supernatant was discarded, and the residual material extracted with 50 ml MeOH/acetone (1:1) for 60 min. This was then filtered and concentrated *in vacuo*, redissolved in 1 ml MeOH and analyzed by HRMS.
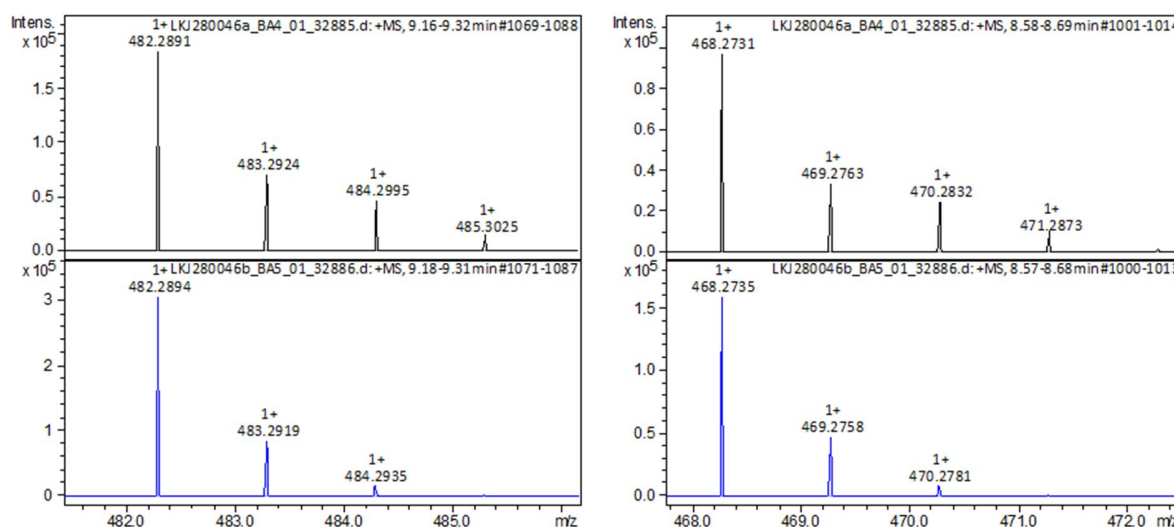
Figure S4.2. *Partial ESI+MS spectra for pyxipyrrolone A (left) and B (right) with [2,3,3-d3]-L-serine feeding (top) and control (bottom). The isotopic patterns of the pyxipyrrolones for the d_3-Ser fed culture shows intensified peaks at m/z 484 and 485, corresponding to a mass shift of +2 Da, caused by the incorporation of deuterium-labelled Ser in the form of dehydroalanine.*

S 5.4.3   Pyxipyrrolone A Mosher analysis

The absolute configuration of the secondary alcohols of **1** was determined by preparation of (*R*)- and (*S*)-Mosher's ester derivatives and assignment of the ^{1}H chemical shift differences, $\Delta\delta^{SR}$. Treatment of **1** with (*R*)- and (*S*)-MTPA (α-methoxy-α-trifluoro-methylphenylacetic acid), DCC and DMAP gave the (*R*)- and (*S*)-MTPA *bis*-Mosher ester derivatives (1a and 1b) (Figure S4.3a). ^{1}H NMR and COSY data was acquired (see section 7.3), assigned and chemical shift differences were calculated as $\Delta\delta_{SR} = \delta_S - \delta_R$ (Figure 4.3b).

S 5.4.3.1   Preparation:

A solution of **1** (1.2 mg, 2.5 µmol), (*S*)-MTPA (3.6 mg, 15.4 µmol), DCC (3.7 mg, 17.5 µmol) and DMAP (1.0 mg, 8.2 µmol) in dry CH_2Cl_2 (1 mL) was stirred at room temperature for 16 h. A second reaction was performed in an analogous manner using (*R*)-MTPA in place of (*S*)-MTPA. The products were semi-purified by HPLC (Phenomenex, Synergi 4µm, Fusion-RP, 250 × 10 mm, gradient from 10 – 100% ACN – H_2O over 30 mins with a hold at 100% ACN for 20 mins) to afford pyxipyrrolone A (*R*)-Mosher ester (**1a**) (t_R = 35.4 , 0.4 mg) and pyxipyrrolone A (*S*)-Mosher ester (**1b**) (t_R = 35.7 , 0.4 mg). (Note: Two repeated batch fermentation (10 L) and isolation of pyxipyrrolone A was performed to generate pyxipyrrolone A (*R*) and (*S*)-Mosher esters (**1a** and **1b**) under similar conditions.)
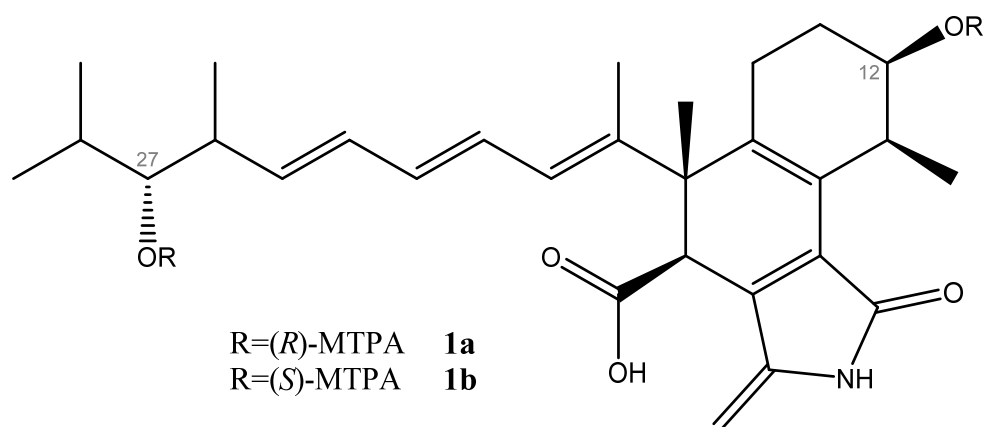
Figure S4.3a. Structure of bis-Mosher's ester derivatives of pyxipyrrolone A, 1a and 1b.
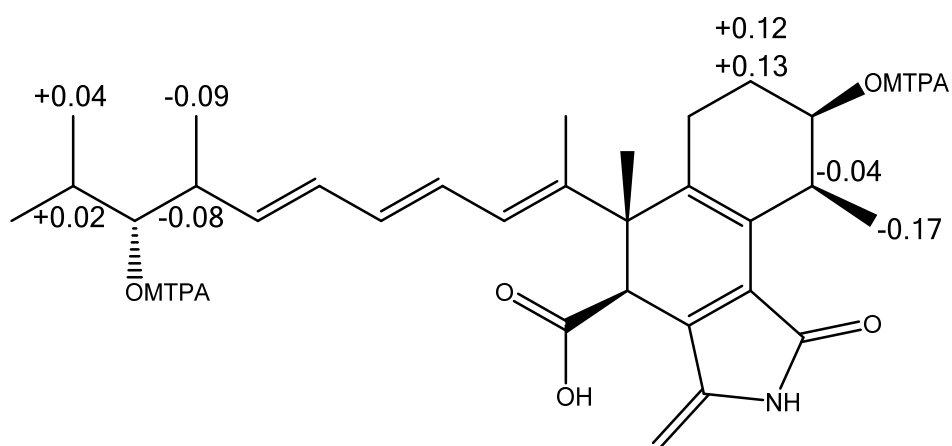
Figure S4.3b. ΔδSR values of the bis-Mosher's ester of pyxipyrrolone A (1).

S 5.5 Gene cluster organization and proposed biosynthesis

S 5.5.1 Genes encoded in the pyxipyrrolone biosynthetic gene cluster

Table S5.1. *Table of all open reading frames belonging to the putative pyxipyrrolone cluster with proposed function and closest homologue according to a blastp search against the nr (non-redundant protein sequences) protein database at NCBI*

Name	Length [bp]	proposed function	closest homologue	Identity [%] and length of alignment [AA]	Accession Nr.
PyxA	12162	polyketide biosynthesis	-	-	-
PyxB	6441	polyketide biosynthesis	CorL	39.8 / 1230	ADI59534
PyxC	7128	polyketide biosynthesis	MalonylCoA acyl carrier protein transacylase	38.1 / 2038	AKQ22681
PyxD	11478	polyketide biosynthesis	-	-	-
PyxE	5652	polyketide biosynthesis and serine incorporation	trans-AT hybride polyketide synthase	34.0 / 1911	ADA69239
PyxF	3633	polyketide biosynthesis	NosB	61.6 / 508	AAF15892
PyxG	723	unknown	hydroxyacylglutathione hydrolase	45.9 / 207	AEB63405
PyxH	477	unknown	hypothetical protein	29.0 / 162	AFC321425
PyxI	3585	transAT acyl transfer	S-Malonyltransferase	45.4 / 1197	WP_037678075
PyxK	255	beta branching	acyl carrier protein	39.1 / 87	SFQ69131
PyxL	1254	beta branching	beta ketoacyl synthase	51.8 / 417	WP_018792602
PyxM	1239	beta branching	3-hydroxy-3-methylglutaryl ACP synthase	60.7 / 412	KJS64697
PyxN	807	beta branching	enoyl-CoA hydratase	35.8 / 265	WP_076891482
PyxO	741	beta branching	enoyl-CoA hydratase	53.9 / 253	WP_030547341
PyxP	636	keto reduction	Short chain dehydrogenase/Reductase	68.8 / 199	WP_002609846
PyxQ	1473	-	carotenoid oxygenase	48.0 / 487	WP_047858220
PyxR	1344	export	ABC transporter	34.2 / 450	WP_071059359
PyxS	1458	carbon-carbon bond reduction	Lycopene cyclase / geranylgeranyl reductase	80.0 / 484	WP_047854616
PyxT	1542	aromatization of the middle ring	Isorenierate synthase	80.2 / 510	WP_075335863
Orf1	288	-	transposase	46.7 / 92	WP_023173142
Orf2	228	-	HTH domain containing protein	49.3 / 74	SFU95209
Orf3	405	-	transposase	86.6 / 134	SEU39003
Orf4	381	-	transposase	96.0 / 126	SEL21147
Orf5	1734	-	transposase	88.3 / 474	SEU38991
Orf6	882	-	endonuclease/exonuclease	61.9 / 289	WP_015346002
Orf7	666	-	sugar O-Acetyltransferase	43.9 / 223	ADY30429
Orf8	1119	-	sugar ABC transporter substrate binding protein	77.8 / 347	WP_014399543

The nucleotide sequence of the pyxipyrrolone gene cluster has been deposited in GenBank and is accessible under accession Nr. KY765914.

S 5.5.2   In detail description of the pyxipyrrolone biosynthesis model

Genomic DNA of *Pyxidicoccus* sp. MCy9557 was sequenced using PacBio technology,[4] yielding a complete genome of approximately 13 Mbp. After antiSMASH prediction of biosynthetic gene clusters encoded on the genome,[5] a candidate gene cluster for pyxipyrrolone biosynthesis is identified by retrobiosynthetic reasoning. As it was not possible to mutagenize *Pyxidicoccus* sp. MCy9557, the identification of the gene cluster as well as the following biosynthesis proposal are solely based on *in silico* analysis. The putative gene cluster responsible for pyxipyrrolone biosynthesis spans 19 open reading

frames (ORFs). The biosynthesis cluster that is identified to be a type 1 trans-acyltransferase (AT) polyketide synthase (PKS) – non-ribosomal peptide synthetase (NRPS) hybrid gene cluster contains 12 modules, eleven of which are PKS derived and one NRPS module (*pyxE),* encoded on 6 genes (*pyxA-F*). The biosynthetic gene cluster contains 13 additional genes encoded on the ORFs *pyxG-T.* The polyketide mega synthase genes lack acyltransferase domains identifying the biosynthesis to be part of the trans-AT PKS cluster family. The two corresponding in trans acting AT's are encoded on the gene *pyxI*, with a domain specificity of the loading AT for malonyl-CoA. One would expect one of the two AT domains to transfer the starter unit while the other AT would probably be responsible for malonyl-CoA extender loading. Unfortunately, it was impossible to predict which acyl transferase catalyzes which reaction as both of the AT domains have predicted substrate specificities for malonyl CoA. The biosynthesis of the pyxipyrrolones features some intriguing steps especially during post-PKS compound maturation. First of all, the pyxipyrrolones exist as two analogues called pyxipyrrolone A and pyxipyrrolone B. These two analogues differ only by a single methyl group at C-28. This difference is originating from alternative starter unit usage. In the case of pyxipyrrolone A the biosynthesis starts with an isobutyryl starter unit while pyxipyrrolone B is extended from a propionyl starter unit. This broadened specificity for multiple starter units is often seen in polyketide derived natural products as for example in myxalamid biosynthesis in *M. xanthus* and *S. aurantiaca*.[6] The second module is likely to be skipped as no malonyl-CoA incorporation for this module is seen in the final product. The consecutive modules 3 and 4 are also peculiar. While module 3 possesses a dehydratase (DH) domain that is inactive in this elongation step, module 4 has neither a keto reductase nor a dehydratase domain that could explain the appearance of a double bond at position C-24 to C-25 in the final product. One could propose that the ketoreductase (KR) domain of module 3 acts twice during chain elongation in module 3 and 4 while the dehydratase present in module 3 only acts after ketosynthesis of module 4 has already been performed. There is evidence that KRs can have a broader substrate specificity spectrum as the bacillaene KR in the pksJ gene is able to reduce α- and β-keto groups.[7] This hypothesis is further supported by the fact that the motif containing the catalytic histidine residue as well as the motif containing the catalytic aspartate residue of the DH domain is present indicating it to be catalytically active. Additionally, the reduction of the carbonyl to a secondary alcohol at position C-27 has to produce a D-alcohol (according to Fischer projection with respect to the acyl carrier protein bound substrate) to fit the stereochemistry observed in the final structure. This is predicted with a score Diff of 59.87, using the profile HMM tool by Kitsche and Kalesse, [8] indicating a good prediction. Furthermore, the pyxipyrrolone structural formulae contain an uncommon β-methyl branch that is attached to the polyketide chain during chain elongation in module 7. This type of β-branching has already been shown to take place in myxobacteria, for example in the corallopyronin pathway.[9] The genes for the

corresponding β-branching cassette are located on *pyxK-O* encoding an acyl carrier protein (ACP), a ketosynthase (KS), a Hydroxymethylglutaryl-CoA Ligase (HMG) and two enoyl-CoA-hydratases (ECH). The assembly of this β-branching cassette is in accord with the β-branching cassette found in myxovirescin.[10] The next uncommon feature of this biosynthetic pathway is the incorporation of a serine moiety during chain elongation in module 11 that is dehydrated to form the dehydroalanine moiety present in the final molecule. Stachelhaus code based NRPS predictor 2 [11] predicts incorporation of serine into the polyketide chain and [2,3,3-d_3]-L-serine feeding experiments performed with the native producer *Pyxidicoccus* sp. MCy9557 confirm this finding. Unfortunately it was not possible to determine a candidate enzyme responsible for dehydration of the incorporated serine meaning this biosynthesis step remains elusive but dehydroalanine formation from serine moieties is well known from the biosynthesis of lantibiotics such as nisin.[12] It was not possible to detect an AT docking domain nor an AT in module 12 although one can see in the final product that malonyl-CoA is integrated to form C-1 and C-2. But since the KR and DH domains of this module are putatively active and the formed double bond is vital for any cyclization reaction during product maturation the AT docking domain in this region is likely to have a somewhat uncommon sequence and is not detected by *in silico* tools.   The next biosynthetic steps in pyxipyrrolone biosynthesis are the cyclisation reactions leading to the final pyxipyrrolone scaffold. It is unclear whether the cyclisation reactions that essentially involve two subsequent Michael additions as well as one intramolecular cyclisation reaction happen on the assembly line or after release of the product but as the thioester has a higher carbonyl activity that facilitates the nucleophilic attack in the first Michael addition it is likely that this reaction cascade already happens on the assembly line. The enzyme PyxQ, which is an iron dependent dioxygenase-like protein, might take part in isomerizing the compound in order to pre-orient the compound for the cyclization reaction cascade. These enzymes have already been characterized to be able to catalyze these kinds of isomerization reactions.[13] The structure possesses one β-keto amide function that is likely to be mostly present as an enol at C-8 to C-9. This moiety can act as a Michael donor while the α,β-unsaturated carboxylic acid at C-1 to C-2 and the vinylogous Michael system from C-12 to C-16 can act as Michael acceptors. Therefore, there is high probability that a protein pre-orients the linear polyketide chain, and two subsequent Michael additions lead to formation of the uncommon tricyclic ring system. The configuration of the stereo center at position C-10 is also most likely determined during the second ring closure, and not during C methyl transfer in module 9, as it is part of a 1,3-diketo function that racemizes very fast before this reaction and it was experimentally confirmed by Mosher's analysis that the methyl group's stereocenter in the final product is only *S*-configured. The last steps in the pyxipyrrolone biosynthesis are probably the reduction at C-12 to C-14 and the dehydrogenation to form the second double bond in the central pyxipyrrolone ring. The enzyme PyxT encodes an isorenierate synthase-like

protein that catalyzes dehydrogenations in six-membered rings. This enzyme might be involved in dehydrogenation of the central ring moiety. Moreover, the genes *pyxS* and *pyxP* encode reductase-like proteins that might play a role in final reduction of the α,β-unsaturated ketone at C-12 to C-14. Finally, the enzyme PyxR shows homology to myxobacterial ABC transporters and is therefore likely to be involved in compound export.

S 5.5.3 *In silico* prediction of the KS substrate specificity using the Bayesian KS alignment method

The structural outcome of trans-AT polyketide systems can be predicted by alignment of the KS domains into a Bayesian tree that classifies the precursor selectivity of the decarboxylative Claisen condensation of a ketosynthesis reaction according to the primary amino acid sequence of the KS domain.[14] According to Nguyen *et al*. KS of trans-AT PKS clusters fall into 16 distinct categories (I to XVI) which leads to some predictability if one looks at the synthesis of the nascent polyketide chain. The pyxipyrrolone biosynthesis comprises 12 KS domains named *pyxKS1* to *pyxKS12* as well as the stand-alone KS on the gene *pyxL* from the β-branching cassette called *pyxKSbetabranching.* In our analysis (Figure S5.3) the substrate specificity of the KS domains in the putative pyxipyrrolone gene cluster fits perfectly the predicted nascent polyketide chain (preferred substrate structure formulae are shown for the KS domain categories that the pyxipyrrolone KS domains fall into).
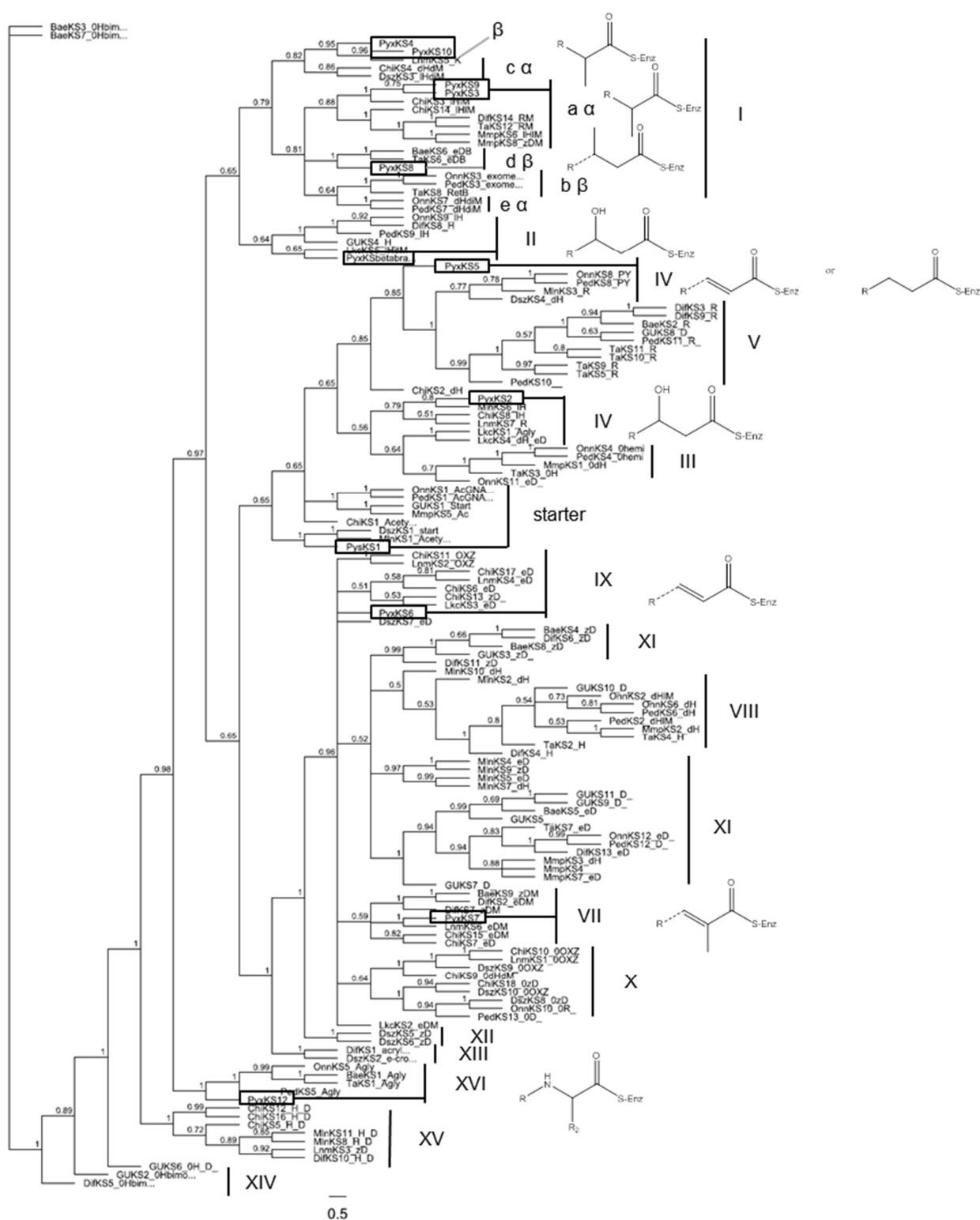
Figure S5.3. KS domain alignment tree containing the pyxipyrrolone KS domains according to Nguyen et al.[14]

S 5.5.4 Stereochemistry prediction of the secondary alcohol moiety during chain elongation in PKS module 3

Profile hidden Markov models of KR domains in PKS assembly lines make it possible to predict the stereochemical outcome of the KR domain in module 3. Therefore the profile HMM online tool under https://akitsche.shinyapps.io/profileHMM_App/ (Accessed February 2017) by Kitsche and Kalesse is used.[8]

The query sequence used for stereochemistry prediction is the translation of the KR domain in module 3 on the *pyxA* gene (FASTA format):

>pyxA KR

LEGEPLECAVFFSSAQSFLGNVGASNYAAGCTFQDAYAAWLREEKGVPTYVLNWGSWG

In silico prediction score for the D-form of secondary alcohols is + 59.87.

The D-form of the secondary alcohol at position C-27 of the pyxipyrrolone scaffold which is synthesized by the KR domain in module 3 fits the structure experimentally determined by Mosher's method (see SI 4.3).

S 5.5.5 In silico analysis of the dehydratase domain in PKS module 3

The situation of the dehydratase domain in module 3 of the PyxA protein is peculiar as it does not act directly in module 3 but might be responsible for dehydration of the nascent polyketide chain in module 4 where there is no dehydratase present. Therefore the DH domains of PyxA, PyxB, PyxC the dehydratase domain on the CorL protein of the corrallopyronin biosynthesis [9] that is predicted to be the nearest neighbor of the pyxipyrrolone assembly line by blast analysis and the dehydratase from the erythromycin biosynthesis that has been crystallized by keatinge and clay are aligned to obtain an insight into the catalytic residues.[15]

Table S5.5a. Conserved region containing the catalytic histidine residue of the DH domains

Motiv	X	X	H *	X	X	X	G	X	X	X	X	P	X
DH *pyxA*	N	E	H *	R	V	R	G	H	A	I	L	P	G
DH4 *ery*	A	E	H *	V	V	G	G	R	T	L	V	P	G
DH *corL*	T	Q	H *	Q	V	H	G	Q	P	V	M	P	G
DH *pyxB*	R	D	H *	V	I	Q	G	R	R	V	L	P	G
DH *pyxC*	A	E	H *	R	V	R	G	S	A	M	L	P	G

Table S5.5b. *Conserved region containing the catalytic aspartate residue of the DH domains*

Motiv	X	X	X	X	D *	X	X	X	Q	X	X	X	X
DH *pyxA*	P	G	M	L	D *	G	A	M	Q	T	V	-	-
DH4 *ery*	P	V	L	L	D *	A	V	A	Q	T	L	S	-
DH *corL*	P	V	L	L	D *	S	V	L	Q	S	V	S	L
DH *pyxB*	P	S	L	V	D *	G	A	I	Q	S	L	-	-
DH *pyxC*	P	G	M	L	D *	G	A	I	Q	S	I	-	-

As both the catalytic histidine as well as the catalytic aspartate are present, the DH domain of PyxA module 3 is possibly active.

S 5.6 Biological activity testing

Cytotoxicity testing: Cell lines were obtained from the German Collection of Microorganisms and Cell Cultures (*Deutsche Sammlung für Mikroorganismen und Zellkulturen*, DSMZ) or the American Type Culture Collection (ATCC). All cell lines were cultured under conditions recommended by the depositor. Media were purchased from Sigma, fetal bovine serum (FBS Gold) from PAN, and other reagents were from Sigma. Cells were seeded at 5×10^4 cells/mL of 96-well plates in 180 μL complete medium and treated with pyxipyrrolones in serial dilution after 2 h of equilibration. After 5 d incubation, 20 μL of 5 mg/mL MTT (thiazolyl blue tetrazolium bromide) in PBS (phosphate-buffered saline; pH 7.4) was added per well and it was further incubated for 2 h at 37 °C. The medium was then discarded and 100 μL 2-propanol/10 N HCl (250:1) was added in order to dissolve formazan granules. The absorbance at 570 nm was measured using a microplate reader (Tecan M200Pro) and cell viability was expressed as percentage relative to the solvent control. IC_{50} values were determined by sigmoidal curve fitting.

Microscopy studies: SW480 cells were harvested and seeded at 5×10^3 cells/well in 96-well imaging plates (BD Falcon) and after 2 d, cells were treated with pyxipyrrolones. The cellular morphology was checked every day and after 4 d treatment images were recorded. The cells were washed with PBS and 75 μL of a staining solution (5 μg/mL acridine orange and 10 μg/mL Hoechst33342 in PBS) was added for 15 min. After staining at 37 °C, cells were washed with PBS and imaged on an automated microscope (BD Pathway 855) in 200-fold magnification with appropriate filter sets for green/red fluorescence and Hoechst.

Antimicrobial testing: Pyxipyrrolones were tested in microbroth dilution assays on the following panel of microorgansism: *Chromobacterium violaceum* DSM-30191, *Escherichia coli* DSM-1116, *E. coli* (TolC-deficient), *Pseudomonas aeruginosa* PA14, *Bacillus subtilis* DSM-10, *Micrococcus luteus* DSM-1790, *Mycobacterium smegmatis* mc²155, *Staphylococcus aureus* Newman, *Mucor hiemalis* DSM-2656, *Candida*

albicans DSM-1665, and *Pichia anomala* DSM-6766. Overnight cultures were prepared from cryopreserved cultures and were diluted to achieve a final inoculum of 10^4-10^5 cfu/mL. Serial dilutions of compounds were prepared in sterile 96-well plates in the respective test medium. The cell suspension was added and microorganisms were grown for 18-48 h at either 30 °C or 37 °C. Growth inhibition was assessed visually.

Table S6. *Biological activities for pyxipyrrolone A (1) and B (2) against a panel of microbial organisms, cancer cell lines, and Plasmodium falciparum. A parenthesis indicates partial inhibition.*

Indicator	1	2
Microorganisms	MIC [µg/mL]	
E. coli DSM-1116	> 64	> 64
E. coli (TolC-deficient)	> 64	> 64
P. aeruginosa PA14	> 64	> 64
M. luteus DSM-1790	> 64	> 64
M. smegmatis mc^{2}155	> 64	> 64
C. violaceum DSM-30191	(64)	(64)
S. aureus Newman	(64)	32-64
B. subtilis DSM-10	> 64	> 64
M. hiemalis DSM-2656	(64)	32
P. anomala DSM-6766	> 64	> 64
C. albicans DSM-1665	(64)	64
Cell lines	IC$_{50}$ [µM]	
HCT-116	68.3	12.6
L929	116.1	10.3
HEC-1-BG	71.4	46.0
THP1	45.3	8.3
A549	> 200	190.3
SW480	34.3	7.9
Parasites	IC$_{50}$ [µM]	
P. falciparum 3D7	> 2.6	> 2.6

S 5.7 References

[1] R. Raju, R. Garcia, R. Muller, *J. Antibiot.* **2014**, *67*, 725.

[2] R. Garcia, R. Müller in *The Prokaryotes*. *Deltaproteobacteria and Epsilonproteobacteria* (Eds.: E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, F. Thompson), Springer Berlin Heidelberg, Berlin, Heidelberg, s.l., **2014**, pp. 191–212.

[3] R. O. Garcia, D. Krug, R. Müller, *Methods Enzymol.* **2009**, *458*, 59.

[4] A. Rhoads, K. F. Au, *Genomics Proteomics Bioinformatics* **2015**, *13*, 278.

[5] T. Weber, K. Blin, S. Duddela, D. Krug, H. U. Kim, R. Bruccoleri, S. Y. Lee, M. A. Fischbach, R. Müller, W. Wohlleben et al., *Nucleic Acids Res.* **2015**, *43*, W237-W243.

[6] H. B. Bode, P. Meiser, T. Klefisch, N. S. Cortina, D. Krug, A. Göhring, G. Schwär, T. Mahmud, Y. A. Elnakady, R. Müller, *ChemBioChem* **2007**, *8*, 2139.

[7] C. T. Calderone, S. B. Bumpus, N. L. Kelleher, C. T. Walsh, N. A. Magarvey, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 12809.

[8] A. Kitsche, M. Kalesse, *ChemBioChem* **2013**, *14*, 851.

[9] Ö. Erol, T. F. Schäberle, A. Schmitz, S. Rachid, C. Gurgui, M. El Omari, F. Lohr, S. Kehraus, J. Piel, R. Müller et al., *ChemBioChem* **2010**, *11*, 1235.

[10] a) C. T. Calderone, D. F. Iwig, P. C. Dorrestein, N. L. Kelleher, C. T. Walsh, *Chem. Biol.* **2007**, *14*, 835; b) V. Simunovic, J. Zapp, S. Rachid, D. Krug, P. Meiser, R. Muller, *ChemBioChem* **2006**, *7*, 1206.

[11] M. Rottig, M. H. Medema, K. Blin, T. Weber, C. Rausch, O. Kohlbacher, *Nucleic Acids Res.* **2011**, *39*, W362-7.

[12] B. Li, J. P. J. Yu, J. S. Brunzelle, G. N. Moll, van der Donk, Wilfred A, S. K. Nair, *Science* **2006**, *311*, 1464.

[13] P. J. Harrison, T. D. H. Bugg, *Arch. Biochem. Biophys.* **2014**, *544*, 105.

[14] T. Nguyen, K. Ishida, H. Jenke-Kodama, E. Dittmann, C. Gurgui, T. Hochmuth, S. Taudien, M. Platzer, C. Hertweck, J. Piel, *Nat. Biotechnol.* **2008**, *26*, 225.

[15] A. Keatinge-Clay, *J. Mol. Biol.* **2008**, *384*, 941.

Chapter 6

Homospermidine Lipids: A Compound Class Specifically Formed during Fruiting Body Formation of *Myxococcus xanthus* DK1622

Michael Hoffmann[a]†, David Auerbach[a]†, Fabian Panter[a], Thomas Hoffmann[a,b], Pieter C. Dorrestein[b] and Rolf Müller[a]*

Affiliations

[a] Department of Microbial Natural Products (MINS), Department of Microbial Natural Products (MINS), Helmholtz Institute for Pharmaceutical Research Saarland (HIPS) - Helmholtz Centre for Infection Research (HZI) and Institute for Pharmaceutical Biotechnology, Saarland University, 66123 Saarbrücken, Germany

[b] Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, California 92093, United States

† These Authors contributed equally to the manuscript

Contributions and Acknowledgements

Author's effort:

The author significantly contributed to the conception of this study, designed and performed experiments, evaluated and interpreted resulting data. The author acquired all LC-MS and LC-MS² data from the strain producing the homospermidine lipids in liquid culture, performed PCA based statistical analysis on the results and extracted and purified the homospermidine lipids Cmp-522, Cmp-564 and Cmp-578 of the four mentioned Homospermidine lipids. The author also conceived the Homospermidine lipid biosynthesis and backed up the proposal with stable isotope labelling experiments.

Contributions by others:

The authors Michael Hoffmann and David Auerbach contributed to the conception of this study, designed and performed experiments, evaluated and interpreted resulting data. The author contributed substantially to the development of a feasible workflow to determine discriminating metabolites depending on the type of cultivation. Cultures on agar were prepared, harvested and extracted by the author. LC-MS analyses and evaluation of the resulting data as well as structure elucidation of Cmp-552 by NMR analysis were conducted by the author. Furthermore, the author contributed equally to conceiving and writing of the manuscript. All MALDI-MSI experiments described were performed, evaluated and interpreted by Thomas Hoffmann. Rolf Müller and Pieter C. Dorrestein contributed by supervision of the project and conceiving, editing and proofreading of the manuscript.

6.1   Abstract

The fascinating ability of myxobacteria to form multicellular spore filled fruiting bodies under starvation conditions was widely studied as a model for cooperative microbial behavior. The potential of a life cycle induced change of secondary metabolism, as a means to discover novel natural products, remains largely underexplored. We therefore studied the model organism *Myxococcus xanthus* DK1622 under submersed and solid cultivation conditions to find putatively life-cycle related compounds by applying statistical analysis on analytical data. Utilizing the advantageous characteristics of LC-MS, LC-MS/MS, and MALDI-MSI allowed the identification of compounds unambiguously associated with myxobacterial fruiting bodies. Our screening effort resulted in the purification and structure elucidation of a novel compound, the homospermidine lipid, from cultures that had undergone the fruiting process. A combination of molecular networking and targeted LC-MS/MS in conjunction with our in-house metabolomics database subsequently revealed alternative producers of the respective compound as well as a number of compounds belonging to the same structural class. Three further members of this compound class were isolated from an alternative producer and structurally elucidated by NMR. Insights into the biosynthesis of this novel compound class was gained by feeding of isotopically labeled substrates and *in silico* analysis.

6.2   Introduction

Microbial secondary metabolism is a long-standing source of novel structural scaffolds used in drug discovery, especially concerning antibiotic agents.[1] Among microbes, myxobacteria are remarkably well suited to provide a variety of secondary metabolites.[2] Details concerning these outstanding characteristics can be found in the literature.[3] Briefly, myxobacteria have exceptional biosynthetic potential due to their large genomes harboring dozens of secondary metabolite encoding gene clusters.[4] Advances in genome sequencing technologies and bioinformatics tools allowed assignment of biosynthetic gene clusters in a large number of complete genome sequences,[5,6] but the number of these clusters is often contrasted by a far lower quantitiy of known secondary metabolites.[5,7] A common explanation for the "absence" of secondary metabolites is that gene clusters might be tightly regulated under laboratory growth conditions leading to low or zero basal expression levels due to missing environmental signals required for activation. A pronounced change in secondary metabolism in response to culturing conditions could be shown for a range of marine actinomycetes in terms of detected molecular networks.[8] For the myxbacterial model species *M. xanthus* DK1622, five compound families and 18 presumed PKS/NRPS

biosynthetic gene clusters are known, but combined transcriptomics and proteomics efforts indicated that 17 out of the 18 clusters were transcriptionally active. Detection of transcription of these clusters with unknown products (on the transcriptome level) is another clear indication that tightly regulated biosynthetic gene clusters require a stimulus to raise their otherwise low basal levels of transcripts.[9] Even though a gene cluster might be functionally expressed, the detection and identification of the corresponding metabolite remains challenging. It is common sense and well documented in the literature that instrumental setup as well as sample preparation have a major impact on the coverage of compounds.[8,10] To harness the biosynthetic potential of bacteria, it is thus mandatory to have a sound analytical platform that allows identifying target compounds from metabolome data with the important additional demand of dereplication to avoid rediscovery of known compounds. More information regarding dereplication of metabolome data can be found in recent reviews.[11,12] As argued above, discovery of novel natural products essentially requires an easy and reliable cultivation method. Over the course of the past decades, cultivation processes in a shaking flask using liquid media with added adsorbents has become the method of choice for myxobacteria. It remains a matter of debate, though, whether this cultivation approach is ideal for soil-dwelling bacteria, which pass through various life cycle stages in their natural environment.[13] This is supported by the fact that many isolated myxobacterial species could not yet be successfully cultivated in liquid media. It can easily be imagined that external stimuli required to activate certain biosynthetic gene clusters (*e.g.,* interspecies cell–cell interaction) are missing in shake flask cultivation. Only limited data of studies aiming at the comparison of liquid and solid cultivation on a larger scale are available. It could be shown, however, that a substantial difference between submerged and solid cultivation can be observed for a range of 30 *actinomycetes*, with only a minimal overlap of 6.8%.[8] It is thus clear that any study aiming to gain insights into the fruiting body formation on a metabolome level requires an analysis technique that connects phenotypic changes to the metabolome to avoid false positive hits caused by differences in cultivation. Hence, we applied mass spectrometry imaging (MSI) as a tool to correlate the appearance of fruiting body specific metabolites for cultures grown on agar in a time dependent manner. In particular, we were interested in metabolic changes during fruiting body formation to find presently unknown natural products as a follow-up study that addressed semiquantitative changes of known metabolites.[14] We decided to use the well-studied *M. xanthus* DK1622, which shows favorably fast growth alongside reproducible fruiting body formation. We studied changes in the metabolome based on extracts of cultures after fruiting body formation grown on agar and cultures grown in suspension. Imaging mass spectrometry served as a tool to filter false positives by correlating the appearance of fruiting body specific metabolites for cultures grown on agar in a time dependent manner.

6.3   Results and Discussion

Fruiting body formation in myxobacteria was extensively studied on the genome and transcriptome level to gain a better understanding of the underlying regulatory and signaling mechanisms, which lead to cell aggregation and sporulation.[15] Effects of fruiting body formation on the metabolome level to increase the structural diversity of experimentally accessible structures however is underexplored. We thus designed an experiment to assess the prospects of fruiting body formation for the discovery of novel secondary metabolites. As a first hypothesis-generating step, cultures grown in suspension were compared to sporulated cultures grown on a solid medium. Possible novel target compounds putatively connected to the myxobacterial life cycle were obtained by statistical comparison of five biological replicates grown in liquid medium and five biological replicates grown on agar (each biological replicate consisted of five agar plates). The resulting distinguishing features can be attributed both to the growth condition as well as to the life cycle, since the presence of vegetative cells on the extracted agar plates cannot be fully excluded. A MALDI mass spectrometry imaging (MALDI-MSI) approach allows achieving this distinction by correlation of putative candidate *m/z* values with fruiting body formation observed on light microscopic images over a time course covering the development from vegetative cells to fruiting body formation.

6.3.1   Submersed Cultures versus Fruiting Bodies

Comparison of both cultivation conditions first required a methodology for extraction of cultures grown on agar plates. Submersed cultures were treated following the well-established approach of cultivation in the presence of an adsorber resin.[16] To promote fruiting body formation, *M. xanthus* DK1622 was cultivated on a minimal agar medium as well as on cellulose membranes placed on top of agar plates containing minimal medium. Liquid cultures used to inoculate plates were washed with minimal medium prior to inoculation. Cultivation on membranes placed on the agar surface was dropped at an early stage due to poor growth characteristics and the lack of fruiting body formation. The presence of adsorber resin randomly spread over agar plates to increase the production of secondary metabolites to avoid feedback inhibition of production and self-toxicity was found to be nonbeneficial. Small amounts of XAD-16 beads on the agar surface favored the extraction of media compounds while signals related to bacterial growth were under-represented as compared to plates without adsorber resin (Figure S1). Apparently, production and extraction efficiency of secondary metabolites cannot be improved for solid cultivation by this approach. The workup of solid cultures was performed by scraping off cells and fruiting bodies using a glass microscope slide (Figure 1). Extraction of complete cultures including agar media

turned out to be inadequate, as the concentration of extracted media compounds is orders of magnitude higher than the compounds related to bacterial growth (data not shown).
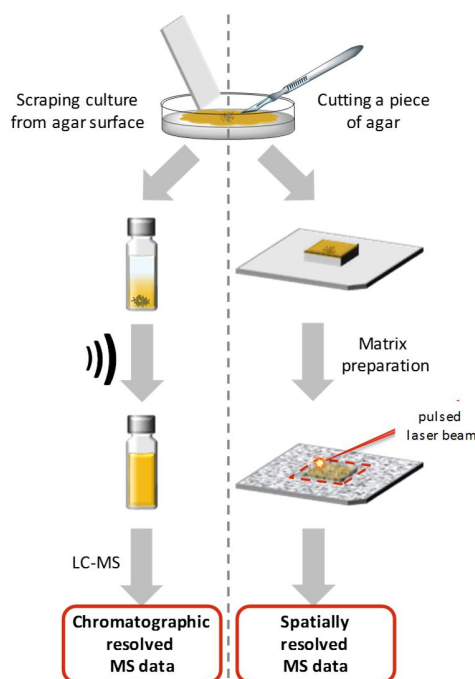
Figure 1. Overview of the two workflows applied for detection and verification of fruiting body related metabolites. To obtain chromatographic resolved MS data, vegetative cells and fruiting bodies were scraped from the agar surface using a microscope slide, extracted with methanol, and measured using LC-MS. Spatially resolved MS data were received by following a MALDI-IMS approach. An agar piece including cells and fruiting bodies was transferred to a MALDI target, dried, had a matrix applied, and measured using MALDI-ToF MS.

Extracts obtained from five biological replicates for both conditions (solid and submersed) were subsequently screened as two technical replicates by LC-MS to generate a set of m/z values putatively connected to changes in metabolism. The highly reduced background observed for extracts from agar plates allowed the highlighting of many signals exclusive to fruiting body samples by mere visual comparison of the base peak chromatograms (BPC; Figure 2). A statistically sound list of candidates was generated using principal component analysis (PCA) in conjunction with t tests to reveal significant differences in the metabolic profiles. Compounds found in at least four out of five biological replicates were considered as a significant hit, as long as no appearance could be detected in media blanks. A total of 62 discriminating molecular features (RT, m/z, intensity) were detected by this means (Figure 3).
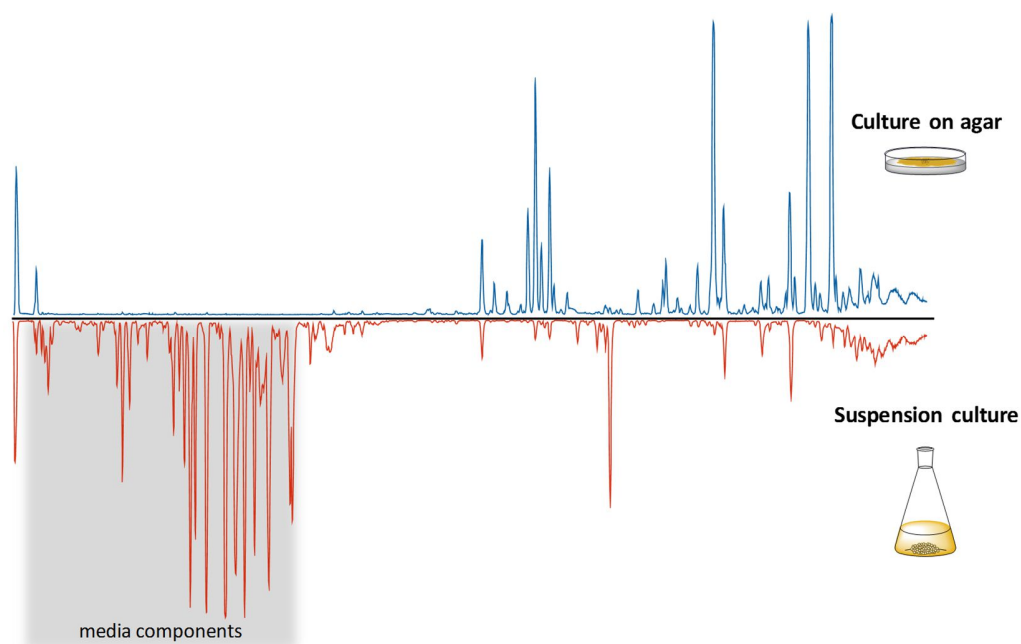
Figure 2. *Comparative base peak chromatograms (150–2500 m/z) of an extract derived from* M. xanthus *DK1622 cultivation on an agar plate (blue) and one derived from cultivation in a liquid medium with adsorber resin (red). Time range with predominantly media component is marked in gray (based on Figure S2).*
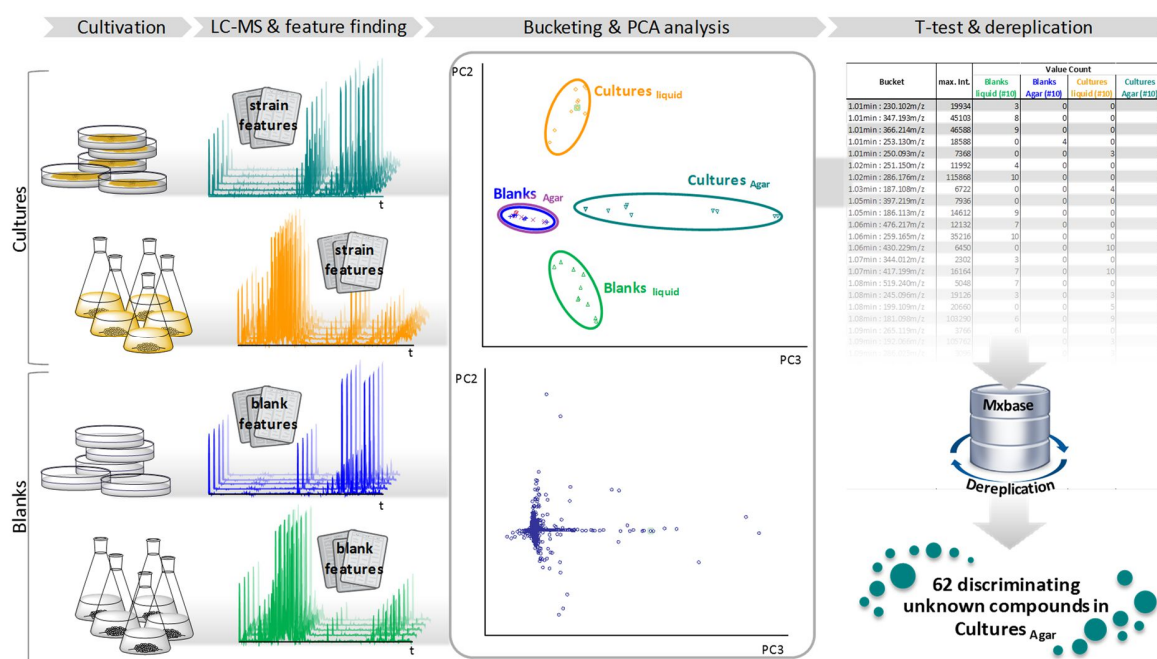
Figure 3. *Workflow toward the detection of discriminating compounds, which are verifiably due to cultivation on solid media, possibly also fruiting body formation. Five replicates of culture as well as five replicates of blanks were prepared for both cultivation methods. Subsequently, all extracts were injected into the LC-MS twice, and the molecular features were determined. On the basis of the molecular features, a principle component analysis and a t test were performed to reveal the discriminating peaks for cultures grown on agar. Dereplication of these data led to 62 compounds, which are directly related to cultivation on agar.*

Although cultivation of *M. xanthus* DK1622 on minimal agar over 5–7 days mainly results in the formation of fruiting bodies, the presence of vegetative cells cannot be excluded.[17] A comparison to vegetative cells grown on agar would be required to further specify the obtained data, which can hardly be achieved using minimal medium, as the presence of fruiting bodies can also not be fully excluded. We thus regarded the set of 62 features as a hypothetical data set that needed to be tested by a suitable additional analysis technique.

6.3.2   MALDI-MSI of *M. xanthus* DK1622

Mass spectrometry imaging of bacterial cultures became well accepted by the research community due to its inherent strength to provide complementary information to other well established molecular analysis workflows.[18,19] MSI excels at providing spatially resolved molecular information in combination with light microscopic images. This allows direct linking to visually observed phenotypic changes to the metabolome in the form of *m/z* values. MALDI-MSI was thus the method of choice to correlate the 62 features obtained from LC-MS measurements with the development of fruiting bodies. The overall workflow is illustrated in Figure 1, while details regarding the method can be found in the literature.[20]

Four time points covering different life cycle stages, including the fruiting body formation of *M. xanthus* DK1622, were analyzed for the presence of the 62 features in question. Mass spectral images were analyzed toward the presence of the 62 features as $[M + H]^+$, $[M + Na]^+$, $[M + NH_4]^+$, and $[M + K]^+$ ions. Surprisingly, only two out of the 62 features could be detected in mass spectral images as sodium and potassium adducts. One of the substances was specifically found in areas covered by fruiting bodies (590.4669 *m/z*), while the other was characteristic for the surrounding cells (724.3979 *m/z*). The mass spectral image of *m/z* 724.3979 shown in Figure 4 illustrates the necessity of an analysis technique that connects phenotypic changes and molecular information, as this compound is clearly specific for vegetative cells grown on agar. The sodiated species with *m/z* 590.4669 corresponds to the compound eluting at 15.9 min with a mass to charge ratio of 552.5104 (Cmp-552) in LC-MS runs. The predominance of cationization observed for these two compounds is likely owed to the presence of salts in the agar medium—a phenomenon that is well-known for biological samples of all sorts.[21] Possible causes for the inability to detect the remaining 60 compounds found by LC-MS are numerous but can only be speculated about. Some of the analytes might simply be present in amounts below the detection limit, which can be additionally impaired by ion suppression as often encountered in MALDI measurements of biological samples.[22] Another important influence on the outcome of MALDI-MSI is the efficiency of analyte extraction, as low efficiency results in a lack of analyte-doped matrix crystals.[23]
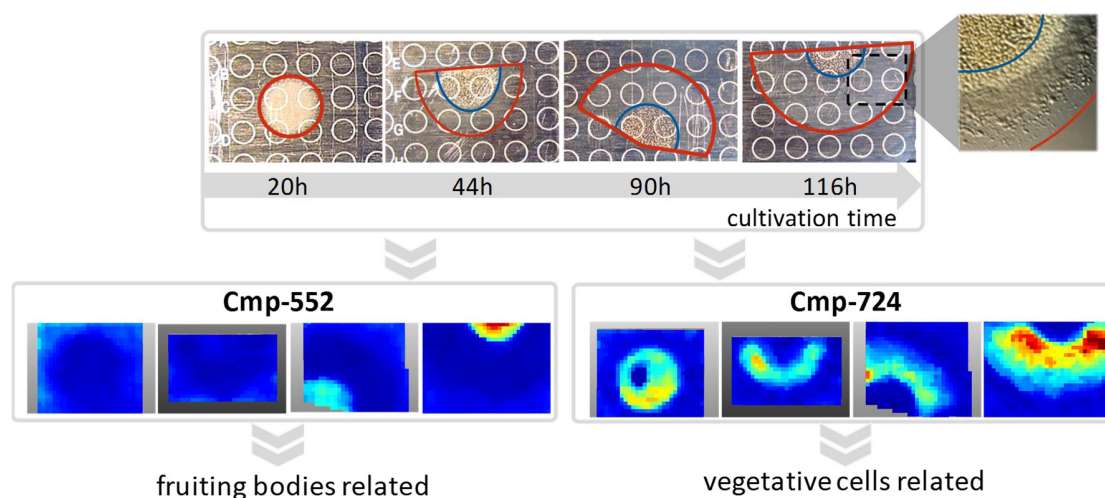
Figure 4. Cultivation time dependent MALDI-MSI data of Cmp-552 and Cmp-686 (both shown as [M+K]⁺). The light microscopic images in the upper part of the figure show the respective samples on a MALDI target prior to matrix application. The red borders are indicating the area of vegetative cells, while the blue borders indicate areas where fruiting bodies are observed.

Several of the distinguishing *m/z* values found by LC-MS are likely specific for fruiting bodies as well, since they share similar MS/MS patterns and belong to the same molecular network (for details, see below). Unambiguous evidence, however, could not be obtained for the respective candidates. Correlation of the MALDI images of *m/z* 590.4669 and the respective light microscopic images revealed that the biosynthesis of Cmp-552 starts at a later stage of fruiting body formation since the signal intensity was the highest after 116 h of cultivation, although fruiting bodies could already be observed after 44 h (Figure 4). We next set out to purify Cmp-552 as the only fruiting body specific compound that could be confirmed by MALDI imaging.

6.3.3   Structure Elucidation and Identification of Derivatives

Purification and structure elucidation of microbial compounds discovered by imaging mass spectrometry is a challenging task. Only one metabolite obtained from MSI data of solid cultivation could be successfully purified and structurally confirmed by NMR.[18,24] Owing to the limited amount of biomass per agar plate, a substantial amount of Petri dishes was necessary to provide sufficient crude extract for subsequent purification attempts.

We could isolate 300 μg of Cmp-552 by semipreparative LC from a methanolic extract of fruiting bodies of *M. xanthus* DK1622 grown on roughly 500 Petri dishes. The structure of Cmp-552 was elucidated by NMR and turned out to be a sym-homospermidine core structure with fatty acyl residues attached to the secondary and the two primary amine groups (Figure 5). Since the compound is reminiscent of a lipid, it was named homospermidine lipid 552. As a descriptive shorthand notation based on the lipidmaps

nomenclature, the following form is suggested: HS(iso5:0/5:0/iso15:0), where HS stands for homospermidine and the three substituents are given as fatty acyl moieties on the two primary amines of sym-homospermidine followed by the substituent on the secondary amine. According to the lipidmaps classification system, the compound belongs to the subclass of N-acyl amines (FA0802).[25]

Cmp-552; **1**
HS(iso5:0/5:0/iso15:0)
$C_{33}H_{65}N_3O_3$
Exact Mass: 551.50259

Cmp-564; **2**
HS(5:0/iso5:0/iso15:0(6E))
$C_{34}H_{65}N_3O_3$
Exact Mass: 563.50259

Cmp-578; **3**
HS(5:0/5:0/iso17:0(6E))
$C_{35}H_{67}N_3O_3$
Exact Mass: 577.51824

Cmp-522; **4**
HS(2:0/2:0/iso19:0(6E))
$C_{31}H_{59}N_3O_3$
Exact Mass: 521.45564
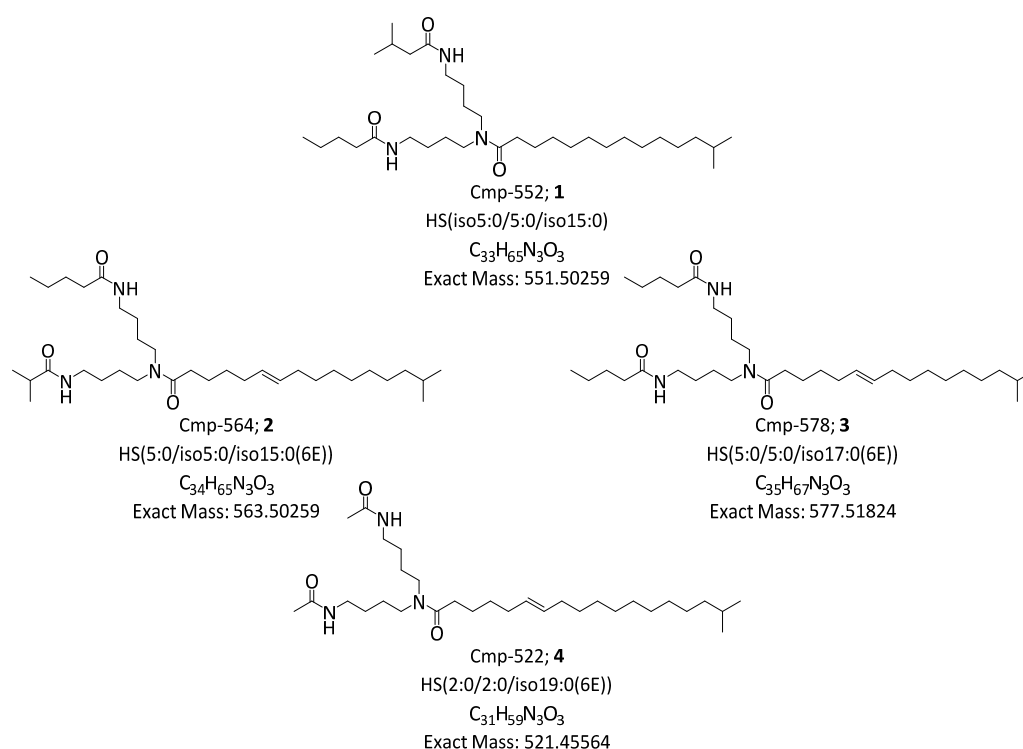
Figure 5. *Chemical structures of the homospermidine lipids (HS). Compounds were named based on the lipidmaps guidelines.*

Use of our in-house metabolome database allowed identification of *Myxococcus fulvus* MCy9290 as a strain that seemed to form the homospermidine lipid in liquid cultures.[12] Since the results shown above prove the relation between the biosynthesis of Cmp-552 and fruiting body formation of *M. xanthus* DK1622, we assume that fundamental differences in regulation lead to an activation of the respective gene cluster during cultivation of *M. fulvus* MCy9290 in a liquid medium. The lipid-like structure of HS(iso5:0/5:0/iso15:0) led to the expectation that further derivatives might be found among the 64 putatively fruiting body related compounds as well as in extracts of *M. fulvus* MCy9290. A powerful and very well received tool to identify such structurally related compounds based on MS/MS similarity is the molecular networking workflow provided within the Global Natural Product Social Molecular Networking (GNPS) platform.[26] Molecular networking is perfectly suited to identify compound classes and is increasingly implemented in dereplication and discovery campaigns.[27]

Extracts of five biological replicates of *M. fulvus* MCy9290 and fruiting bodies of *M. xanthus* DK1622 were obtained from liquid and solid cultivation, respectively. Each sample was measured twice, and results were subjected to PCA analysis with subsequent *t* tests as described above. The resulting culture specific features were used for the acquisition of targeted MS/MS data as an input for spectral networking. The resulting MS/MS networks are shown in Figure 6.
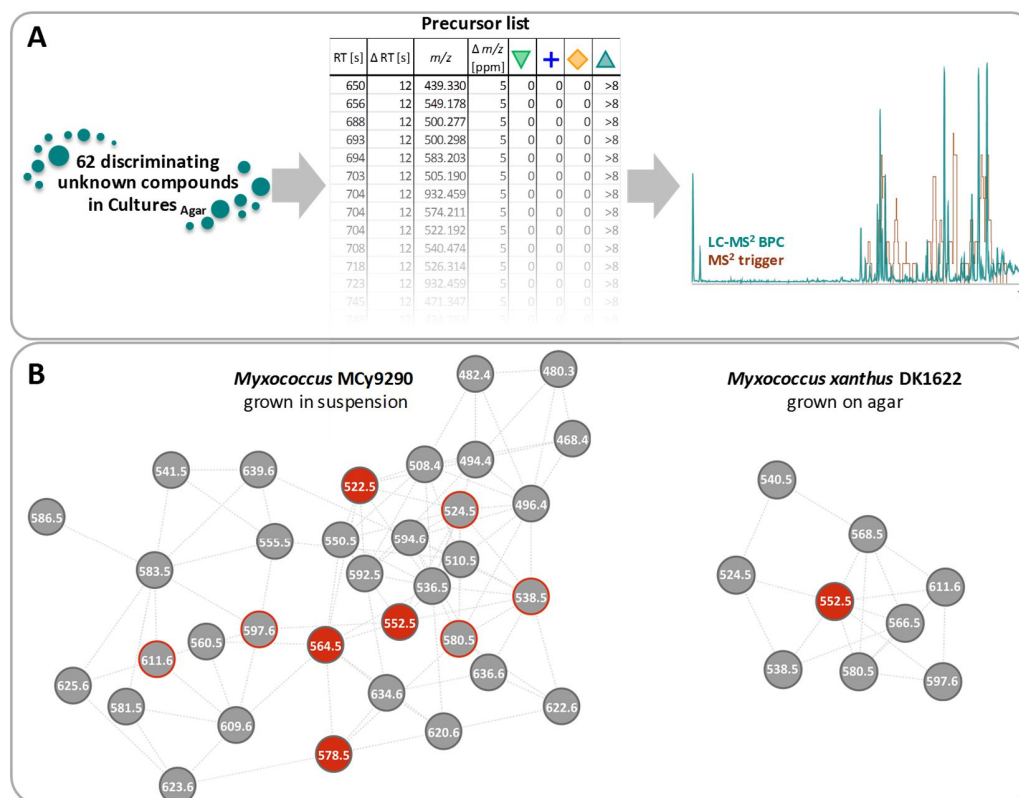
Figure 6. *A precursor directed MS/MS based on PCA and* t *test results exemplified by the 62 discriminating unknown compounds in agar grown cultures. B Cmp-552 including MS/MS networks derived from extracts of* M. fulvus *MCy9290 (grown in suspension) and* M. xanthus *DK1622 (grown on agar). Red colored nodes indicating derivatives, which were isolated, and structures elucidated. Nodes bordered in red marking derivatives of Cmp-552, which were also found in extracts of* M. xanthus *DK1622.*

As liquid cultures are more accessible for preparing larger amounts of extract, we cultivated *M. fulvus* MCy9290 on a larger scale to isolate further derivatives of the novel compound class. We were able to isolate three additional main derivatives differing mainly by alkyl chain length and saturation degree while showing similar bioactivities (Figure 5 and Table 1). The configuration of the double bonds was determined to be *trans* due to the allylic carbon shifts. The vinylic hydrogens form second order signals that prevent the evaluation of coupling constants.[28]

Table 1. *Minimum Inhibitory Concentration (MIC) of Homospermindine Lipids 1–4*

test organism	MIC/IC$_{50}$
Staphylococcus aureus	16 µg/mL
Bacillus subtilis	16 µg/mL
Micrococcus luteus	8 µg/mL
Mucor hiemalis	64 µg/mL
Chinese hamster ovary CHO-K1	10–15 µg/mL

As observed activities do not differ for the individual derivatives, values are not listed separately.

6.3.4 Bioactivity and Biosynthesis of Homospermidine-Lipids

Testing 24 compounds for bioactivity against a panel of microorganisms and cell lines revealed cytotoxicity values in the range of 20 µM and moderate activity against some Gram-positive bacteria ranging from 14 to 28 µM (Table 1). Since the purpose of fruiting body formation is to endure times of low nutrient supply, we hypothesize that the organism tries to defend the fruiting body against predators and thereby increases its chances of survival. Although comparably high, MIC values might represent a spore defense mechanism, as the local concentrations of the lipids in the fruiting bodies is relatively high. Another biological function of these lipids might be carbon and nitrogen storage during the low nutrient supply.
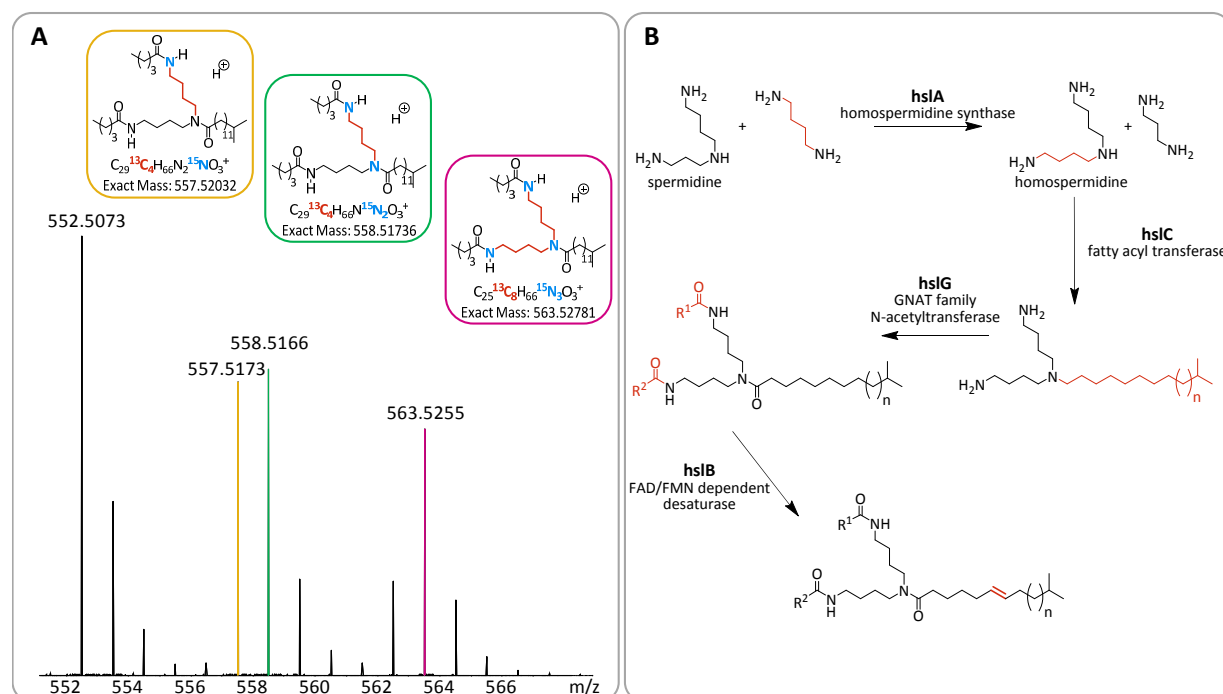
Figure 7. *(A) MS spectrum of feeding experiment with l-arginine-$^{13}C_6$,$^{15}N_4$ confirms the incorporation of up to two arginine building blocks in Cmp-552. (B) Putative biosynthesis of the homospermidine lipids according to in silico analysis.*

The biosynthesis of the homospermidine lipids was elucidated based on the assumption that the homospermidine lipids derive from sym-homospermidine and thereby originate from arginine *via* the putrescine pathway (Figure 7).[29] For this reason, $[^{13}C_6,^{15}N_4]$-l-arginine was fed to a liquid culture of *M. fulvus* MCy9290, and the obtained extracts were checked for the corresponding mass shift of the compounds. LC-MS analysis confirmed the incorporation of two arginine moieties, substantiating the homospermidine origin of this substance class (Figure 7). Since the genome sequence of *M. xanthus* DK1622 is available, we searched for the responsible biosynthetic gene cluster. A genetic island containing the homospermidine synthase, an *N*-acyltransferase, and a fatty acid acyltransferase was identified by searching for the homospermidine synthase analogues known as deoxyhypusine synthases available on NCBI Protein using BLAST.[30] More details regarding the potential biosynthesis can be found in the Supporting Information.

6.4   Conclusion

In this work, we present the first secondary metabolomics study devoted to the influence of fruiting body formation in a myxobacterium. Application of a broad scope of measurement and analysis techniques allowed unambiguous identification, purification, and structural elucidation of a novel compound that is directly connected to fruiting body formation of *M. xanthus* DK1622. We present one of the very few cases where imaging mass spectrometry leads to the isolation of a compound obtained from cultures grown on a solid medium. The data obtained from mass spectral images impressively show the necessity of applying a molecular imaging technique that allows connection of spatial information with visually observed phenotypic changes to establish clear-cut evidence for the localization of a candidate *m/z* value on fruiting bodies of *M. xanthus* DK1622. Further database and spectral networking based analysis resulted in the identification of a new compound family and allowed us to isolate three more derivatives from an alternative and deregulated producer capable of forming homospermidine lipids in liquid culture. Data obtained from spectral networking indicates that other members related to HS(iso5:0/5:0/iso15:0) are also specific for fruiting bodies, but confirmation by MALDI MSI was not possible due to the inability to detect the respective *m/z* values.

Owing to the observed cytotoxicity and activity against several Gram-positive bacteria, these compounds might be part of the bacterium's defense strategy or play a role in carbon storage. As the activities of the compounds were rather nonselective, it can be assumed that they are part of a more general defense against competitors. We believe that we have shown a very promising workflow that will not only lead to the identification of additional novel natural compounds but also offer unique

opportunities to gain fundamental insights into metabolic processes taking place during fruiting body formation.

6.5 References

[1] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2016**, *79*, 629.

[2] J. Herrmann, A. A. Fayad, R. Müller, *Nat. Prod. Rep.* **2017**, *34*, 135.

[3] S. C. Wenzel, R. Müller, *Mol. Biosyst.* **2009**, *5*, 567.

[4] N. Zaburannyi, B. Bunk, J. Maier, J. Overmann, R. Müller, *Appl. Environ. Microbiol.* **2016**, *82*, 1945.

[5] S. C. Wenzel, R. Müller, *Nat. Prod. Rep.* **2009**, *26*, 1385.

[6] a) M. H. Medema, M. A. Fischbach, *Nat. Chem. Biol.* **2015**, *11*, 639; b) N. Ziemert, M. Alanjary, T. Weber, *Nat. Prod. Rep.* **2016**, *33*, 988.

[7] P. R. Jensen, *Trends Microbiol.* **2016**, *24*, 968.

[8] M. Crüsemann, E. C. O'Neill, C. B. Larson, A. V. Melnik, D. J. Floros, R. R. da Silva, P. R. Jensen, P. C. Dorrestein, B. S. Moore, *J. Nat. Prod.* **2016**.

[9] P. J. Rutledge, G. L. Challis, *Nat. Rev. Microbiol.* **2015**, *13*, 509.

[10] a) J.-L. Wolfender, *Planta medica* **2009**, *75*, 719; b) L. J. Barkal, A. B. Theberge, C.-J. Guo, J. Spraker, L. Rappert, J. Berthier, K. A. Brakke, C. C. C. Wang, D. J. Beebe, N. P. Keller et al., *Nature communications* **2016**, *7*, 10610.

[11] a) J. R. Doroghazi, J. C. Albright, A. W. Goering, K.-S. Ju, R. R. Haines, K. A. Tchalukov, D. P. Labeda, N. L. Kelleher, W. W. Metcalf, *Nat. Chem. Biol.* **2014**, *10*, 963; b) K. M. Fisch, T. F. Schäberle, *Arch. Pharm. (Weinheim)* **2016**, *349*, 683; c) P.-M. Allard, G. Genta-Jouve, J.-L. Wolfender, *Current opinion in chemical biology* **2017**, *36*, 40; d) B. C. Covington, J. A. McLean, B. O. Bachmann, *Nat. Prod. Rep.* **2017**, *34*, 6.

[12] D. Krug, R. Müller, *Nat. Prod. Rep.* **2014**, *31*, 768.

[13] R. Garcia, D. Krug, R. Müller in *Methods in Enzymology*. *Complex Enzymes in Microbial Natural Product Biosynthesis*. Part A: Overview Articles and Peptides (Ed.: David A. Hopwood), **2009**, pp. 59–91.

[14] C. Volz, C. Kegler, R. Müller, *Chem. Biol.* **2012**, *19*, 1447.

[15] a) S. Huntley, N. Hamann, S. Wegener-Feldbrügge, A. Treuner-Lange, M. Kube, R. Reinhardt, S. Klages, R. Müller, C. M. Ronning, W. C. Nierman et al., *Mol. Biol. Evol.* **2011**, *28*, 1083; b) F. J. Marcos-Torres, J. Pérez, N. Gómez-Santos, A. Moraleda-Muñoz, J. Muñoz-Dorado, *Nucleic acids research*

2016, *44*, 5571; c) L. Kroos, *Trends Genet.* **2017**, *33*, 3; d) D. G. Lloyd, D. E. Whitworth, *Front. Microbiol.* **2017**, *8*, 439.

[16] a) K. Gerth, N. Bedorf, G. Höfle, H. Irschik, H. Reichenbach, *J. Antibiot.* **1996**, *49*, 560; b) C. Jahns, T. Hoffmann, S. Müller, K. Gerth, P. Washausen, G. Höfle, H. Reichenbach, M. Kalesse, R. Müller, *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 5239; c) L. Keller, A. Plaza, C. Dubiella, M. Groll, M. Kaiser, R. Muller, *J. Am. Chem. Soc.* **2015**, *137*, 8121.

[17] S. A. Kraemer, M. A. Toups, G. J. Velicer, *FEMS microbiology ecology* **2010**, *73*, 226.

[18] T. Luzzatto-Knaan, A. V. Melnik, P. C. Dorrestein, *The Analyst* **2015**, *140*, 4949.

[19] S. J. B. Dunham, J. F. Ellis, B. Li, J. V. Sweedler, *Accounts of chemical research* **2017**, *50*, 96.

[20] T. Hoffmann, P. C. Dorrestein, *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1959.

[21] a) E. B. Monroe, B. A. Koszczuk, J. L. Losh, J. C. Jurchen, J. V. Sweedler, *International Journal of Mass Spectrometry* **2007**, *260*, 237; b) D. Gode, D. A. Volmer, *The Analyst* **2013**, *138*, 1289.

[22] a) M. Petkovic, J. Schiller, M. Müller, S. Benard, S. Reichl, K. Arnold, J. Arnhold, *Anal. Biochem.* **2001**, *289*, 202; b) R. Knochenmuss, *Annual review of analytical chemistry (Palo Alto, Calif.)* **2016**, *9*, 365.

[23] a) L. Crossman, N. A. McHugh, Y. Hsieh, W. A. Korfmacher, J. Chen, *Rapid Commun. Mass Spectrom.* **2006**, *20*, 284; b) S. Shimma, Y. Sugiura, *Mass spectrometry (Tokyo, Japan)* **2014**, *3*, S0029; c) L. Signor, E. Varesio, R. F. Staack, V. Starke, W. F. Richter, G. Hopfgartner, *J. Mass Spectrom.* **2007**, *42*, 900.

[24] a) W.-T. Liu, Y.-L. Yang, Y. Xu, A. Lamsa, N. M. Haste, J. Y. Yang, J. Ng, D. Gonzalez, C. D. Ellermeier, P. D. Straight et al., *PNAS* **2010**, *107*, 16286; b) C.-J. Shih, P.-Y. Chen, C.-C. Liaw, Y.-M. Lai, Y.-L. Yang, *Nat. Prod. Rep.* **2014**, *31*, 739.

[25] E. Fahy, S. Subramaniam, R. C. Murphy, M. Nishijima, C. R. H. Raetz, T. Shimizu, F. Spener, G. van Meer, M. J. O. Wakelam, E. A. Dennis, *J. Lipid Res.* **2009**, *50 Suppl*, S9-14.

[26] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan et al., *Nat. Biotechnol.* **2016**, *34*, 828.

[27] P.-M. Allard, T. Péresse, J. Bisson, K. Gindro, L. Marcourt, C. van Pham, F. Roussi, M. Litaudon, J.-L. Wolfender, *Analytical chemistry* **2016**, *88*, 3317.

[28] F. D. Gunstone, M. R. Pollard, C. M. Scrimgeour, H. S. Vedanayagam, *Chemistry and physics of lipids* **1977**, *18*, 115.

[29] R. Kuttan, A. N. Radhakrishnan, *Biochem. J.* **1972**, *127*, 61.

[30] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403.

Supporting Information

Homospermidine Lipids: A Compound Class Specifically Formed during Fruiting Body Formation of *Myxococcus xanthus* DK1622

Previously published in:

Michael Hoffmann[a]†, David Auerbach[a]†, Fabian Panter[a], Thomas Hoffmann[a,b], Pieter C. Dorrestein[b] and Rolf Müller[a]*

ACS Chem. Biol. 2018 Jan 19; **13**(1):273-280

DOI: 10.1021/acschembio.7b00816

Affiliations

[a] Department of Microbial Natural Products (MINS), Department of Microbial Natural Products (MINS), Helmholtz Institute for Pharmaceutical Research Saarland (HIPS) - Helmholtz Centre for Infection Research (HZI) and Institute for Pharmaceutical Biotechnology, Saarland University, 66123 Saarbrücken, Germany

[b] Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, California 92093, United States

† These Authors contributed equally to the manuscript

The printed version of the Supporting Information does not contain data that cannot be visualized satisfactorily on paper. To access this data please refer to the enclosed storage medium or the on-line version of this research paper.

S 6.1   Methods

S 6.1.1   Extract Preparation - Cultures grown in suspension.

Pre-cultures of *Myxococcus xanthus* DK1622 and *Myxococcus fulvus* MCy9290 were grown at 30 °C for 3 days in 25–50 mL CTT medium (2 g L^{-1} MgSO$_4$ x 7H$_2$O, 10 g L^{-1} casitone, 1.2 g L^{-1} TRIS, 136 mg L^{-1} K$_2$HPO$_4$; pH 7.6) and VY/2 medium (5 g L^{-1} bakers yeast, 50 mg L^{-1} CaCl$_2$ x 2H$_2$O, 1.2 g L^{-1} HEPES, 10 g L^{-1} starch; pH 7.0), respectively. Main cultures were inoculated by the well-grown pre-cultures (2 % (v/v)) and cultivated at 30 °C for 7 days. Two days after inoculation, XAD-16 absorber resin was added to the culture (1 % (w/v)). Extracts were prepared by extracting the resin including cells with methanol, evaporation of solvent and re-dissolving in methanol. 1 µL of crude extract was analyzed by LC-hrMS. For purification purposes of *M. fulvus* MCy9290, a total of 7.5 L suspension culture including XAD-16 was grown and extracted in accordance to the method described above.

S 6.1.2   Extract Preparation - Cultures grown on agar.

All cultures were cultivated in standard Petri dishes on HS agar medium (1 g L^{-1} MgSO$_4$ x 7H$_2$O, 1 g L^{-1} KNO$_3$, 1.5 g L^{-1} casitone, 2 g L^{-1} TRIS, 4 g L^{-1} Glucose, 8 mg L^{-1} Fe-EDTA, 75 mg L^{-1} CaCl$_2$ x 2H$_2$O, 6.25 mg L^{-1} K$_2$HPO$_4$, 2 % agar (w/v); pH 7.2) at 30 °C for 7 days. Pre-cultures, cultivated in liquid CTT medium, were washed with CF medium (2 g L^{-1} MgSO$_4$ x 7H$_2$O, 150 mg L^{-1} casitone, 1.2 g L^{-1} TRIS, 174 mg L^{-1} K$_2$HPO$_4$, 0.2 mg L^{-1} (NH$_4$)SO$_4$, 1 g L^{-1} sodium pyruvate, 2 g L^{-1} sodium citrate; pH 7.6) and 10-fold concentrated. Petri dishes were inoculated with 150 µLplate^{-1} of the concentrated pre-culture. Extracts were prepared by scraping off the cells and fruiting bodies from the agar surface, extraction with methanol, evaporation and re-dissolving in methanol.

In total, approximately 500 Petri dishes were prepared, harvested, and extracted to obtain sufficient amounts of Cmp-552.

S 6.1.3   LC-hrMS Methods.

All measurements of crude extracts were performed on a Dionex Ultimate 3000 RSLC comprising a high pressure gradient pump (HPG-3400RS) with a 150 µl mixing chamber. All LC connections are realized by 0.13 µm stainless steel capillaries before the column. The method is based on separation with BEH C18, 100 x 2.1 mm, 1.7 µm dp column (Waters, Eschborn, Germany). One µl sample is separated by a linear gradient from (A) H$_2$O + 0.1 % (v/v) formic acid to (B) ACN + 0.1 % (v/v) formic acid at a flow rate of 600 µL min^{-1} and 45 °C. The gradient is initiated by a 0.5 min isocratic step at 5 % B, followed by an increase to 95 % B in 18 min to end up with a 2 min step at 95 % B before reequilibration under the initial conditions. UV spectra are recorded by a DAD in the range from 200 to 600 nm with 2 nm slit width. The LC flow is

split to 75 µL min^{-1} before entering the maXis 4G hr-ToF mass spectrometer (Bruker Daltonics, Germany) using the Apollo ESI source. A fused silica (f.s.) capillary (20 cm, 100 µm ID, 360 µm OD) coming from the mass spectrometer's switching valve is connected to an Upchurch PEEK microtight T-junction. Another f.s. capillary of 20 cm and 75 µm ID is connected to the inlet of the mass spectrometer whereas an f.s. capillary of 10 cm and 100 µm ID is directed to waste. The split is approximately 1:8 based solely on the f.s. capillary ratio. The real split is even more than 1:8 as the effect of the ESI needle is not taken into account here. The ion source parameters are: capillary, 4000 V; endplate offset, -500 V; nebulizer, 1 bar; dry gas, 5 L min^{-1}; dry gas temperature, 200 °C. Ion transfer parameters were: funnelRF, 350 Vpp; multipoleRF, 400 Vpp; quadrupole ion energy, 5 eV @ low *m/z* 120. Collision cell is set to 5 eV with a collisionRF of 1100 Vpp; transfer time, 110 µs; pre puls storage, 5 µs in full scan mode. Mass spectra are acquired in centroid mode ranging from 150–2500 *m/z* at a 2 Hz scan rate in full scan positive ESI mode. Each measurement is started with the injection of a 20 µL plug of basic sodium formate solution, which is introduced by a loop that is connected to the system's 6-port switching valve. The resulting peak is used for automatic internal *m/z* calibration. In addition, three lock masses (622, 922, 1222 *m/z*) are present in the ion source all the time and allows recalibration of each spectrum.

Fractions obtained by semi-preparative purification were measured on the same setup using a 50 mm instead of a 100 mm column of the same series. Hence, only the gradient was slightly adjusted. Initiation by a 1 min isocratic step at 5 % B, followed by an increase to 95 % B in 6 min to end up with a 1.5 min step at 95 % B before reequilibration under the initial conditions.

Statistical evaluation of the data was performed using *ProfileAnalysis* V2.1 (Bruker Daltonics, Germany)

S 6.1.4   MALDI-MSI Method.

Experimental details regarding strain cultivations, sample preparation, and instrumental parameters can be found in literature.[1] All MS imaging data were evaluated and visualized with *SCiLS Lab* 2015 (Bruker Daltonics, Germany).

S 6.1.5   MS^n Experiments and Molecular Networking.

All LC-MS/MS measurements were based on the already described 18 min gradient separation using the U3000-maXis 4G setup. Fragmentation spectra were achieved either by auto-MS/MS measurements or by targeted precursor selection. Selected precursor list (SPL) were generated by taking into account the dereplication and PCA results respectively. Fixed MS/MS parameters were an *m/z*-adjusted precursor isolation width and fragmentation energy. Values are interpolated in between the set points [300 *m/z* / 4 *m/z* / 30 eV], [600 *m/z* / 6 *m/z* / 35 eV], [1000 *m/z* / 8 *m/z* / 45 eV], and [2000 *m/z* / 10 *m/z* / 55 eV]. The

precursor intensity threshold is always set to 5000. All methods are set up with an active precursor exclusion, which puts precursors after two spectra on an exclusion list. Each precursor *m/z* remains there for 0.2 min.

A molecular network was created using the online workflow at GNPS (Global Natural Products Social Molecular Networking). The data was filtered by removing all MS/MS peaks within +/- 17 Da of the precursor m/z. MS/MS spectra were window filtered by choosing only the top 6 peaks in the +/- 50 Da window throughout the spectrum. The data was then clustered with MS-Cluster with a parent mass tolerance of 0.05 Da and a MS/MS fragment ion tolerance of 0.1 Da to create consensus spectra . Further, consensus spectra that contained less than 1 spectra were discarded. A network was then created where edges were filtered to have a cosine score above 0.65 and more than 4 matched peaks. Further edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. Data were visualized with *Cytoscape* V 3.5.0.[2]

MS^n fragmentation experiments were conducted by direct infusion MS using a solariX XR (7T) FT-ICR mass spectrometer (Bruker Daltonics, Germany) equipped with an Apollo ESI source. In the source region, the temperature was set to 200 °C, the capillary voltage was 4500 V, the dry-gas flow was 4.0 L min^{-1} and the nebulizer was set to 1.1 bar. While MS/MS spectra were obtained by collision-induced dissociation (CID) in the collision cell, all further fragmentations were obtained by isolating the respective precursor in the ICR cell and applying sustained off resonance irradiation (SORI) CID. Parameters were individually adjusted for each measurement.

S 6.1.6   Purification of HS(iso5:0/5:0/iso15:0) from *M. xanthus* DK1622.

Purification was carried out on a Thermo Dionex (Germering, Germany) Ultimate 3000 low pressure gradient system, equipped with SR3000 solvent rack, LPG-3400SD pump module, WPS-3000TSL autosampler, Knauer (Berlin, Germany) Jetstream column oven, DAD-3000 photodiode array detector with a cell volume of 10 μL and an AFC 3000 fraction collector. LC separations were performed using 50 μL injections on a Phenomenex Kinetex Biphenyl column (10 x 250 mm, 5 μm) at 40°C and a flow rate of 5 mL min^{-1} with the following gradient conditions: 0–55 min, 5–95 % B; 55–59 min, 95 % B; 59–59.5 min, 95–5 % B; 59.5–70 min, 5 % B. Mobile phases were (A) water + 0.1 % (v/v) formic acid and (B) acetonitrile + 0.1 % (v/v) formic acid. Fractions of 18 s width were collected from 20–55 min. The resulting fractions were combined, evaporated to dryness and reconstituted in 1 mL of methanol.

S 6.1.7   Purification of HS-derivatives from *M. fulvus* MCy9290.

First separation was done on a Sephadex LH20 gravitation column and follow-up separation was done on Waters preparative HPLC using a Waters X-Bridge C-18, 150 x 19 mm, 5 μm dp column. Separation was

performed using (A) water + 0.1 % (v/v) formic acid and (B) acetonitrile + 0.1 % (v/v) formic acid and a flow rate of 5 mL min^{-1} with the following gradient: 0–3 min, 5 % B; 3–7 min, 5–55 % B; 7–32 min, 55–98 % B; 32–35 min, 98 % B; 35–36 min, 98–5 % B; 36–40 min, 5 % B.

Residual impurities were removed by an additional semi-preparative chromatography step on a Phenomenex Synergi Fusion-RP, 250 x 10 mm, 4 µm dp column. Separation was performed using the same eluent as described above and a flow rate of 5 mL min^{-1} by using the following gradient: 0–3 min, 5 % B; 3–7 min, 5–65 % B; 7–27 min, 55–85 % B; 27–28 min, 85–98 % B; 28–31 min, 98 % B; 31–32 min, 98–5 % B; 32–36 min, 5 % B.

S 6.1.8   Antimicrobial Assay.

All microorganisms were handled according to standard procedures and were obtained from the German Collection of Microorganisms and Cell Cultures (*Deutsche Sammlung für Mikroorganismen und Zellkulturen*, DSMZ) or were part of our internal strain collection. For microdilution assays, overnight cultures of Gram-positive bacteria in Müller-Hinton broth (0.2 % (w/v) beef infusion, 0.15 % (w/v) corn starch, 1.75 % (w/v) casein peptone; pH 7.4) were diluted in the growth medium to achieve a final inoculum of ca. 10^6 cfu ml^{-1}. Serial dilutions of homospermidine lipids were prepared from Methanol stocks in sterile 96-well plates. The cell suspension was added and microorganisms were grown on a microplate shaker (750 rpm, 37°C and 16 h). Growth inhibition was assessed by visual inspection and given MIC (minimum inhibitory concentration) values are the lowest concentration of antibiotic at which no visible growth was observed.

S 6.1.9   Cytotoxic activity.

Cell lines were obtained from the German Collection of Microorganisms and Cell Cultures (Deutsche Sammlung für Mikroorganismen und Zellkulturen, DSMZ) or were part of our internal collection and were cultured under conditions recommended by the depositor. Cells were seeded at 6 x 103 cells per well of 96-well plates in 180 µl complete medium and treated with homospermidine lipids in serial dilution after 2 h of equilibration. Each compound was tested in duplicate as well as the internal solvent control. After 5 d incubation, 20 µl of 5 mg ml^{-1} MTT (thiazolyl blue tetrazolium bromide) in PBS was added per well and it was further incubated for 2 h at 37°C. The medium was then discarded and cells were washed with 100 µl PBS before adding 100 µl 2-propanol/10 N HCl (250:1) in order to dissolve formazan granules. The absorbance at 570 nm was measured using a microplate reader (Tecan Infinite M200Pro), and cell viability was expressed as percentage relative to the respective methanol control. IC$_{50}$ values were determined by sigmoidal curve fitting.

S 6.2   Supplementary Figures & Tables
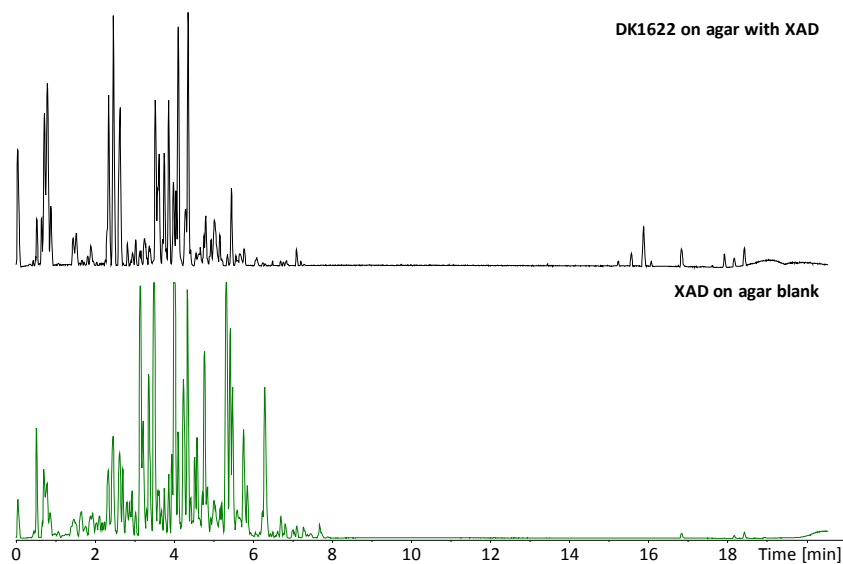
Figure S 1. *Base peak chromatograms (150–2500 m/z) showing the effect of XAD addition during cultivation on agar.*
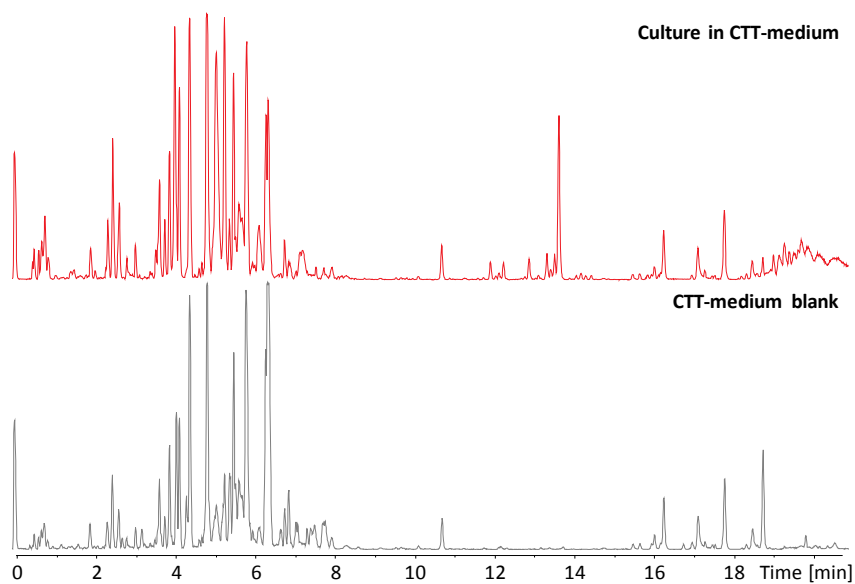
Figure S 2. *Comparative base peak chromatograms (150 – 2500 m/z) of a M. xanthus DK1622 extract originating from suspension culture in CTT medium and the respective blank.*
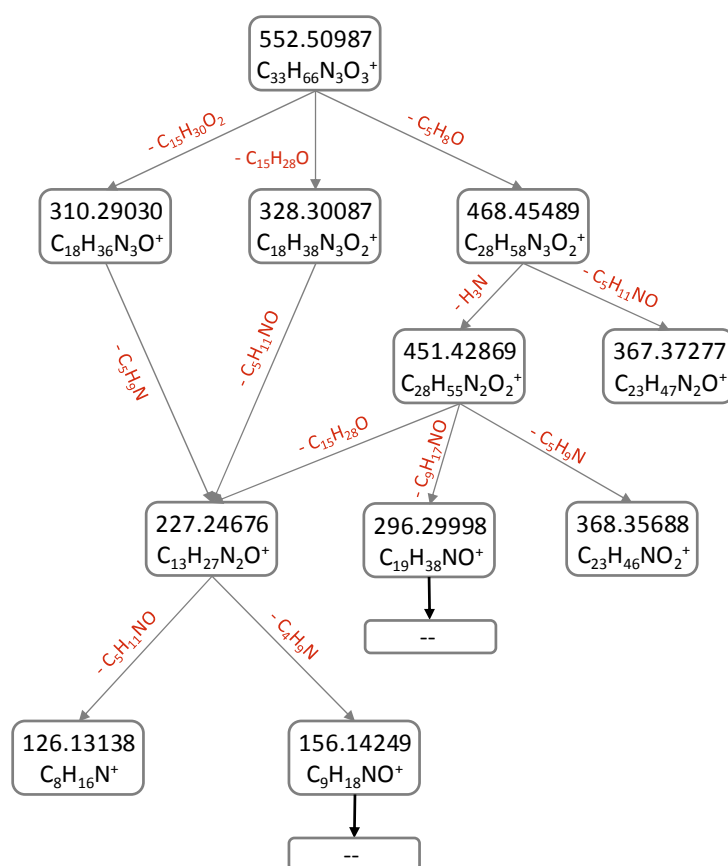
Figure S 3. *MS fragmentation tree of Cmp-552 based on MS^n experiments performed on FT-ICR.*
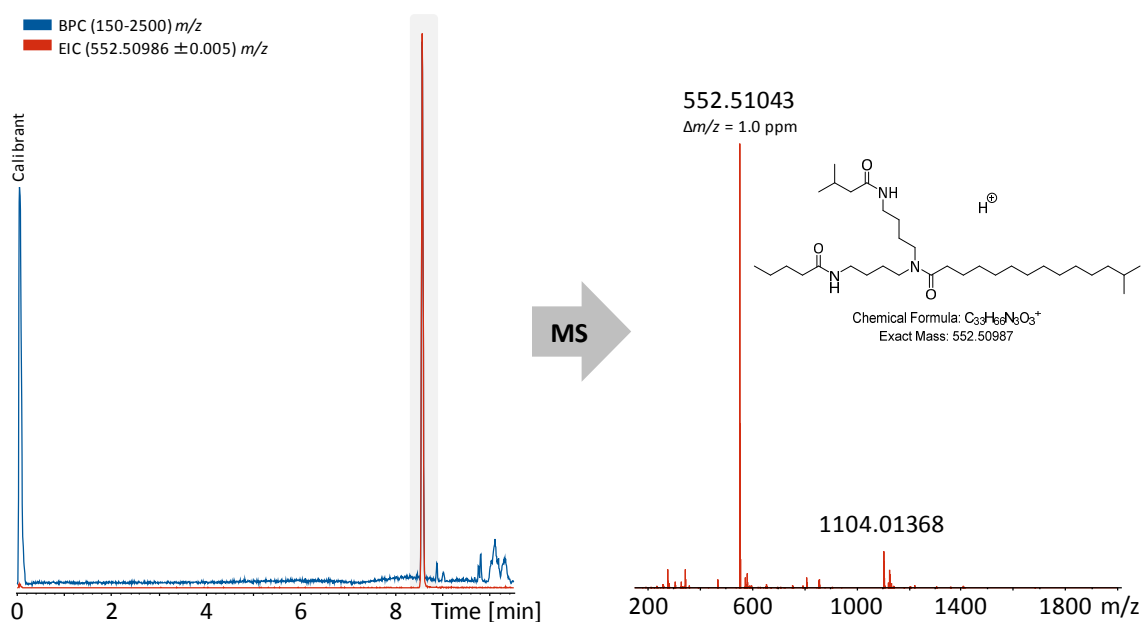
Figure S 4. *Overlaid base peak chromatogram and extracted ion chromatogram (552.50986 ±0.005 m/z) of the Cmp-552 containing fraction indicated sufficient purity for NMR analysis. The respective fraction was received by semi-preparative isolation effort of M. xanthus DK1622 extracts from cultivation on agar.*
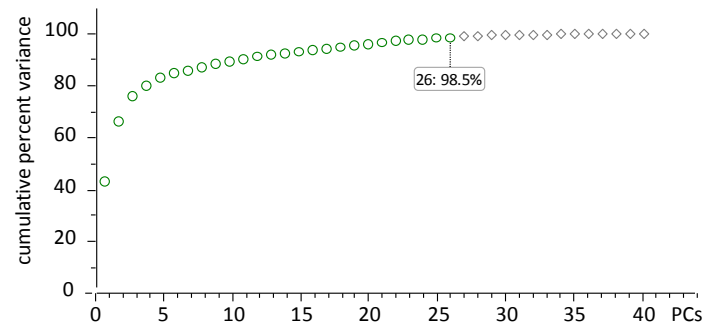
Figure S 5. *Explained variance plot of the PCA model considered in Figure 3.*

Table S 1. *All open reading frames (orfs) belonging to the putative homospermidine lipid biosynthesis cluster with proposed function and closest homologue according to a blastp search against the non-redundant protein database and the position of the orfs on the published M. xanthus DK1622 genome (NCBI reference sequence NC_008095.1)*

Name	Length [bp]	Proposed function	Closest homologue	Identity [%] and length of alignment [AA]	Accession of closest homologue	Locus tag on the DK1622 genome
hslA	1023	Homospermidine synthesis	Homospermidine synthase (Myxococcus virescens)	98,5 / 339	SDF06447	MXAN_1678
hslB	1896	fatty acyl chain dehydrogenation	oxidase – FAD/FMN dependent (Myxococcus fulvus)	97,2 / 631	WP_043712221	MXAN_1676
hslC	2607	fatty acyl chain transfer	glycerol-3-phosphate acyl transferase (Myxococcus virescens)	99.5 / 868	SDF06388	MXAN_1675
hslD	990	unknown	NAD dependent epimerase (Myxococcus virescens)	98.5 / 329	SDF06360	MXAN_1674
hslE	1542	unknown	hypothetical adventurous gliding motility protein N (Myxococcus xanthus)	98.8 / 344	AAO66306	MXAN_1673
hslF	1005	unknown	hypothetical Protein (Stigmatella aurantiaca)	71.6 / 334	WP_075004684	MXAN_1687
hslG	897	N-acylation of the homospemidine moiety	Acetyltransferase GNAT family (Myxococcus virescens)	93.6 / 298	SDF06648	MXAN_1686
hslH	174	unknown	hypothetical protein (Myxococcus virescens)	92.9 / 56	SDF06632	MXAN_1685
hslI	558	regulation	histidine kinase (Myxococcus fulvus)	75.9 / 191	WP_013938788	MXAN_1684
hslK	2280	unknown	Dipeptidyl aminopeptidase (Myxococcus virescens)	98.3 / 747	SDF06517	MXAN_1682
hslL	798	unknown	Dienelactone hydrolase (Myxococcus virescens)	97.7 / 265	SDF06504	MXAN_1681
hslM	675	regulation	DNA binding response regulator	99.6 / 224	SDF06483	MXAN_1680
hslN	1167	regulation	signal transduction histidine kinase	96.6 / 338	SDF06456	MXAN_1679
hslO	2610	compound sensing	tonB like protein (Myxococcus virescens)	96.8/ 869	SDF06711	MXAN_1688
hslP	1269	unknown	hypothetical protein (Myxococcus virescens)	98.7 / 422	SDF06729	MXAN_1689
hslQ	945	unknown	esterase (Myxococcus fulvus)	79.9 / 313	WP_013935160	MXAN_1690
hslR1	525	regulation	TetR family regulator (Myxococcus virescens)	99.4 / 174	SDF06422	MXAN_1677
hslR2	906	regulation	LysR family regulator (Myxococcus virescens)	99.3 / 301	SDF06540	MXAN_1683
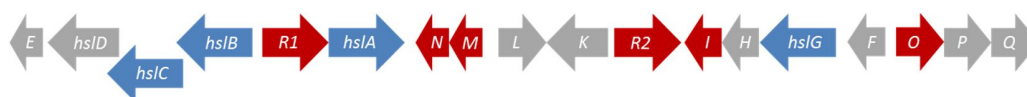
Figure S 6. *Gene organization of the putative homospermidine lipid biosynthesis cluster. Biosynthesis genes depicted in blue; regulatory elements depicted in red; genes of unknown function or genes putatively not involved in homospermidine lipid biosynthesis depicted in gray.*

Cmp-552; **1**
$C_{33}H_{65}N_3O_3$
Exact Mass: 551.50259

Cmp-564; **2**
$C_{34}H_{65}N_3O_3$
Exact Mass: 563.50259

Cmp-578; **3**
$C_{35}H_{67}N_3O_3$
Exact Mass: 577.51824

Cmp-522; **4**
$C_{31}H_{59}N_3O_3$
Exact Mass: 521.45564
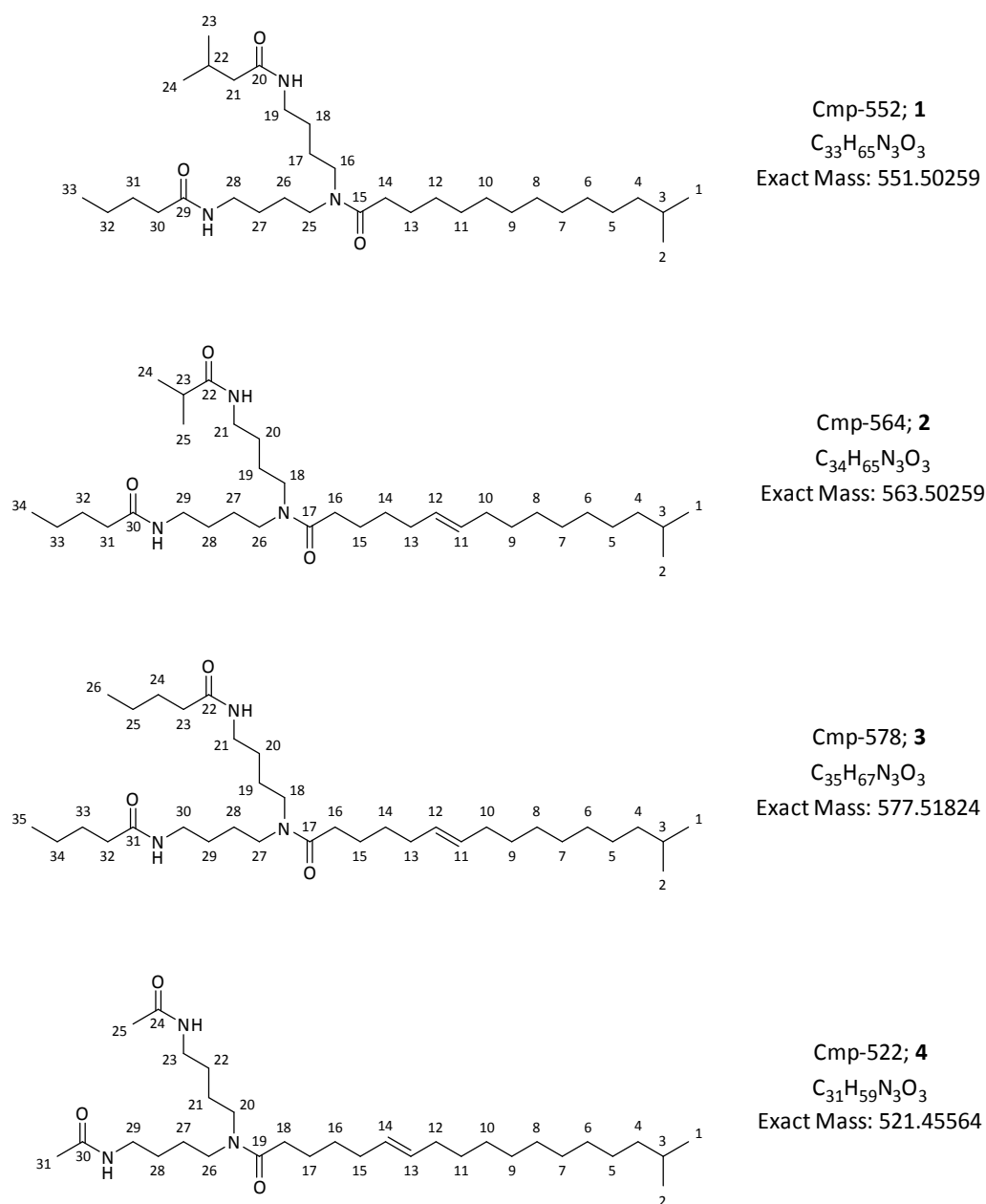
Figure S 7. *Chemical structures of homospermidine-lipid derivatives with numbered carbon atoms corresponding to the NMR-data tables (Table S 2 to Table S 5)*

Table S 2. *NMR data for Cmp-552, 1; HS(iso5:0/5:0/iso15:0); ([a–c] HSQC signals overlapping)*

Carbon number	δ ^{1}H [ppm]	Multiplicity, J [Hz] and proton number	δ ^{13}C [ppm]	COSY correlations	HMBC correlations
1/2	0.84	d, 6.73, 6	23.0	3	1/2, 3, 4
3[a]	1.49	-	28.2	1/2, 4	1/2, 4
4[b]	1.13	-	39.4	3, 5, 6	1/2, 4, 5, 6
5/6/7[c]	1.23	-	30.1	4, 8/9/10	4, 8/9/10
8/9/10[a]	1.58	-	26.0	5/6/7, 8/9/10, 11/12	5/6/7, 8/9/10, 11/12
11/12[c]	1.29	-	30.1	8/9/10, 13	8/9/10, 11/12, 13, 14
13	1.59	-	26.1	12, 14	11, 12, 14, 15
14	2.25	t, 7.59, 2	33.6	12, 13	12, 13, 15
15	-	-	173.5	-	14, 16, 25
16	3.30	-	45.5	17	15, 17, 25
17[a]	1.52	-	25.1	17, 18	16, 18, 19
18[a]	1.47	-	26.7	17, 19	16, 17, 19
19	3.27	-	39.0	18	17, 18, 20
20	-	-	177.7	-	19, 21, 22
21	2.35	m	33.8	22	22, 23/24
22	2.35		36.0	21, 23/24	20, 21, 23/24
23/24[b]	1.13	t, 6.69, 6	20.0	21, 22	20, 21, 22, 23/24
25	3.23	t, 7.59, 2	48.0	26	15, 16, 26/27
26/27[a]	1.54	-	26.1	25, 27	25, 26/27, 28
28	3.27	-	39.0	26/27	25, 26/27, 29
29	-	-	173.7	-	28, 30, 31
30	2.14	t, 7.64, 2	39.1	31	31, 32, 33
31	1.64	m	19.5	30, 32, 33	30, 32, 33
32[c]	1.29	-	25.9	31, 33	30, 31, 33
33	m	-	14.3	31, 32	31, 33

Table S 3. *NMR data for Cmp-564, 2; HS(5:0/iso5:0/iso15:0(6E))*

Carbon number	δ ¹H [ppm]	Multiplicity, J [Hz] and proton number	δ ¹³C [ppm]	COSY correlations [ppm]	HMBC correlations [ppm]
1 + 2	0.91	d 6.7 Hz 2H	23.58	1.55	29.50, 40,69
3	1.55	m 1H	29.7	0.91, 1.21	23.57, 40.32
4	1.21	m 2H	40	1.55, 1.31	23.4, 29.6, 32.0
5	1.31	m 2H	32	1.2, 1.3	31.0, 30
6	ca 1.3	m	ca. 30	ca 1.3	ca. 30
7	ca 1.3	m	ca. 30	ca 1.3	ca. 30
8	ca 1.3	m	ca. 30	ca 1.3	ca. 30
9	1.37	m 2H	30.9	2.0, 1.34	132.0, 33.8
10	2	m 2H	33.8	1.37, 5.43	132.0, 30.9
11	5.43	m 1H	132	2.0, 5.37	-
12	5.37	m 1H	131.1	2.07, 5.43	28.8
13	2.07	m 2H	28.8	1.42, 5.34	131.1, 31.3
14	1.42	m 2H	31.3	2.07, 1,65	131.1, 33.4, 26.8, 28.1
15	1.62	m 2H	26.9	1.42, 2.26	31.3, 175.0
16	2.27	t 2H	36.9	1.62	26.9, 31.3, 179.5
17	-	-	176.5	-	-
18	3.34	m 2H	48.07	1.58, 3.2	26.5, 176.5
19	1.58	m 2H	26.5	1.508, 3,34	48.1,28.5
20	1.508	m 2H	28.5	1.58, 3.34	26.5, 40.3
21	3.21	t 2H 7.19 Hz	40.3	1.51, 1.58	28.5, 180.3
22			180.3		
23	2.454	dq 1H	36.9	1.133	20.3, 180.3
24+25	1.133	CH3 2 mal	20.36	2.467	36.9, 20.28, 180.0
26	3.35	m 2H	47.26	1.58, 3.20	26.5, 176.5
27	1.58	m 2H	26.5	1.51, 3.35	47.3, 28.5
28	1.507	m 2H	28.5	1.58, 3.20	26.5, 36.9
29	3.22	t 2H 7.19 Hz	36.9	1.50, 1.58	28.5, 176
30			175.93		
31	2.378	t 2H	34.45	1.6442,	26.7, 30.9
32	1.6442	m 2H	26.7	1.347, 3,36	
33	1.35	m 2H	30.1		
34	0.934	1 3H 6.7 Hz	14.8	1.347	

Table S 4. NMR data for Cmp-578, 3; HS(5:0/5:0/iso17:0(6E))

Carbon number	δ ¹H [ppm]	Multiplicity, J [Hz] and proton number	δ ¹³C [ppm]	COSY correlations [ppm]	HMBC correlations [ppm]
1 + 2	0.9	d 6.7 Hz 6H	23.67	1.54	29.08, 40.8
3	1.55	m 1H	29.67	0.9, 1.2	23.37, 28.11, 40.71
4	1.2	m 2H	40.72	1.54	23.6, 29.6, 33.75
5	1.31	m 2H	33.75	1.2, 1.3	40.7, ~30
6	ca 1.3	m	ca. 30	ca 1.3	ca. 30
7	ca 1.3	m	ca. 30	ca 1.3	ca. 30
8	ca 1.3	m	ca. 30	ca 1.3	ca. 30
9	1.38	m 2H	30.3	1.3, 2.0	34.1, 132.0
10	2	m 2H	34.1	1.38, 5.45, 5,34	30.3, 131.1, 132.0
11	5.45	m 1H	132	2.0, 5,34, 2.07	34.1, 132.0, 33
12	5.375	m 1H	131.1	2.07, 2.2, 5.45	31.0, 131.1, 132.0
13	2.07	m 2H	31	1.42, 5.375	30.9, 131.5, 132.0
14	1.42	m 2H	30.9	2.07, 1.62,	131.1, 31, 26.89, 33.78
15	1.6	m 2H	26.8	1.42, 2.31	30.9, 33.68,176.3
16	2.375	t, 2H, 7.2Hz	33.7	1.63	26.8, 31.0, 176.3
17	-	-	176.3	-	-
18	3.36	m 2H	47.04	1.58, 3.2	26.6, 176.5
19	1.57	m 2H	26.65	1.508, 3,34	47.0, 28.4
20	1.5	m 2H	28.38	1.58, 3.34	26.7, 40.2
21	3.21	t 2H 7.19 Hz	40.19	1.51, 1.58	28.3, 178.5
22			178.3		
23	2.3	t 2H	35.62	1.62	28.8, 31, 178.5
24	1.62		26.8	1.34	35.46
25	1.34		33.4	0.93	35.4, 14.7
26	0.93		14.7	1.34	33.4
27	3.36	m 2H	47.04	1.58, 3.20	26.6, 176.5
28	1.57	m 2H	26.65	1.51, 3.35	47.0, 28.4
29	1.5	m 2H	28.38	1.58, 3.20	26.7, 40.2
30	3.21	t 2H 7.19 Hz	40.19	1.50, 1.58	28.3, 178.5
31			178.3		
32	2.3	t 2H	35.62	1.62	28.8, 31, 178.5
33	1.62	m 2H	26.8	1.34	35.46
34	1.34	m 2H	33.4	0.93	35.4, 14.7
35	0.93	1 3H	14.7	1.34	33.4

Table S 5. *NMR data for Cmp-522, 4; HS(2:0/2:0/iso19:0(6E))*

Carbon number	δ 1H [ppm]	Multiplicity, J [Hz] and proton number	δ 13C [ppm]	COSY correlations [ppm]	HMBC correlations [ppm]
1 + 2	0.88	d, 6.7 Hz, 6H	22.9	1.51	29.0, 39.8
3	1.51	m, 1H	29	0.88, 1.18	39.8, 22.9, 28.3
4	1.18	m, 2H	39.8	1.29, 1.51	22.9, 29.0, 28.3
5	1.29	m, 2H	28.3	1.18, 1.3	30, 39.8
6	ca. 1.3	m, 2H	ca. 30	ca. 1.3	ca. 30
7	ca. 1.3	m, 2H	ca. 30	ca. 1.3	ca. 30
8	ca. 1.3	m, 2H	ca. 30	ca. 1.3	ca. 30
9	ca. 1.3	m, 2H	ca. 30	ca. 1.3	ca. 30
10	ca. 1.3	m, 2H	ca. 30	ca. 1.3	ca. 30
11	1.35	m, 2H	30.4	1.98, ca.1.30	131.6, 33.3, 30
12	1.98	m, 2H	33.3	1.35, 5.45, 5.39	30.4, 131.6, 130.6
13	5.45	m, 1H	131.6	1.98, 2.01, 5.39	30.4, 33.3,130.6
14	5.39	m, 1H	130.6	2.01, 1.98, 5.45	33.0, 30.0, 131.6
15	2.01	m, 2H	33	1.40, 5.39	30.2, 33.3, 130.6, 131.6
16	1.4	m, 2H	30.2	1.59, 2.35, 2.01	33.0, 26.0
17	1.59	m, 2H	26	1.40, 2.35	33.7, 30.2 175.2
18	2.35	t, 7.6 Hz, 2H	33.7	1.59, 1.40, 3.32	30.2, 26.0, 175.2
19	-	-	175.2	-	-
20	3.32	m, 2H	46.3	1.56, 3.35, 2.35	25.9, 48.5, 175.2
21	1.56	m, 2H	25.9	3.32, 1.48	46.3, 27.1
22	1.48	m, 2H	27.1	3.17, 1.56	25.9, 46.3, 39.6
23	3.17	t, 6.2 Hz, 2H	39.6	1.48, 1.92	27.1, 25.9, 173.3
24	-	-	173.3	-	-
25	1.92	s, 3H	22.1	3.17	39.6, 173.3
26	3.35	m, 2H	48.5	1.59, 3.32, 2.35	27.0, 46.3, 175.2
27	1.59	m, 2H	27	3.35, 1.52	48.5, 27.5
28	1.52	m, 2H	27.5	3.20, 1.59	48.5, 27.0, 39.4
29	3.2	t, 6.2 Hz, 2H	39.4	1.52, 1.94	27.5, 27.0, 173.5
30	-	-	173.5	-	-
31	1.94	s, 3H	22.1	3.2	39.4 ,173.5
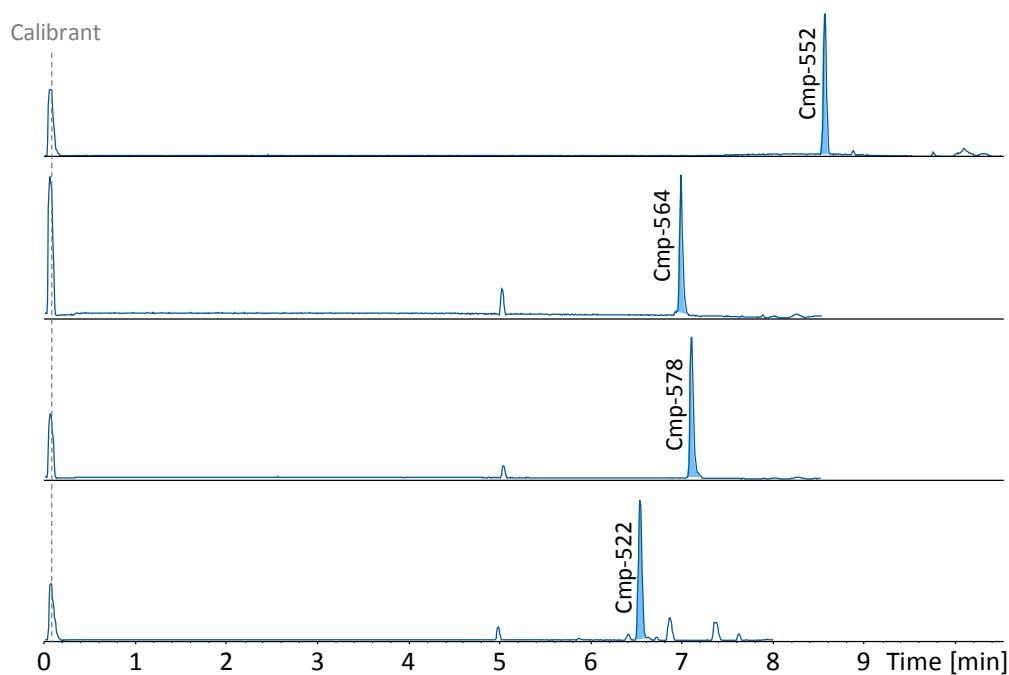
Figure S 12. *Base Peak Chromatograms of pure homospermidine lipids for NMR analysis*

S 6.3   References

[1]   T. Hoffmann, P. C. Dorrestein, *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1959.

[2]   a) M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan et al., *Nat. Biotechnol.* **2016**, *34*, 828; b) W.-T. Liu, Y.-L. Yang, Y. Xu, A. Lamsa, N. M. Haste, J. Y. Yang, J. Ng, D. Gonzalez, C. D. Ellermeier, P. D. Straight et al., *PNAS* **2010**, *107*, 16286.

Chapter 7

Predicting the Presence of Uncommon Elements in Unknown Biomolecules from Isotope Patterns

Marvin Meusel[a], Franziska Hufsky[a,c], Fabian Panter[b], Daniel Krug[b], Rolf Müller[b] and Sebastian Böcker*[a]

Affiliations

[a] Chair for Bioinformatics, Friedrich Schiller University Jena, 07743 Jena, Germany

[b] Department of Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland, Helmholtz Centre for Infection Research and Pharmaceutical Biotechnology, Saarland University, 66123 Saarbrücken, Germany

[c] RNA Bioinformatics and High Throughput Analysis, Friedrich Schiller University Jena, 07743 Jena, Germany

Contributions and Acknowledgements

Author's effort:

The author contributed to performing experiments and the evaluation and interpretation of resulting data. The author acquired all used mass spectrometry data acquired by direct infusion on our maxis4G spectrometer, on the one hand an evaluation data set for evaluating noise distribution around m/z values and relative isotope peak intensities as well as the myxobacterial natural product spectra set used during algorithm training. The author also contributed to editing and proofreading of the manuscript.

Contributions by others:

The authors Marvin Meusel and Franziska Hufsky contributed to the conception of this study, designed experiments, evaluated and interpreted resulting data. The authors conceptualized and programmed the machine learning based software and implemented the program into SIRIUS. The author also contributed significantly to conceptualizing, editing and proofreading of the manuscript. Daniel Krug contributed to editing and proofreading of the manuscript. Rolf Müller and Sebastian Böcker contributed by supervision of the project and conceiving, editing and proofreading of the manuscript.

7.1   Abstract

The determination of the molecular formula is one of the earliest and most important steps when investigating the chemical nature of an unknown compound. Common approaches use the isotopic pattern of a compound measured using mass spectrometry. Computational methods to determine the molecular formula from this isotopic pattern require a fixed set of elements. Considering all possible elements severely increases running times and more importantly the chance for false positive identifications as the number of candidate formulas for a given target mass rises significantly if the constituting elements are not prefiltered. This negative effect grows stronger for compounds of higher molecular mass as the effect of a single atom on the overall isotopic pattern grows smaller. On the other hand, hand-selected restrictions on this set of elements may prevent the identification of the correct molecular formula. Thus, it is a crucial step to determine the set of elements most likely comprising the compound prior to the assignment of an elemental formula to an exact mass. In this paper, we present a method to determine the presence of certain elements (sulfur, chlorine, bromine, boron, and selenium) in the compound from its (high mass accuracy) isotopic pattern. We limit ourselves to biomolecules, in the sense of products from nature or synthetic products with potential bioactivity. The classifiers developed here predict the presence of an element with a very high sensitivity and high specificity. We evaluate classifiers on three real-world data sets with 663 isotope patterns in total: 184 isotope patterns containing sulfur, 187 containing chlorine, 14 containing bromine, one containing boron, one containing selenium. In no case do we make a false negative prediction; for chlorine, bromine, boron, and selenium, we make ten false positive predictions in total. We also demonstrate the impact of our method on the identification of molecular formulas, in particular on the number of considered candidates and running time. The element prediction will be part of the next SIRIUS release, available from https://bio.informatik.uni-jena.de/software/sirius/.

7.2   Introduction

Hyphenated high-resolution mass spectrometry, mostly coupled to liquid chromatography (LC–MS) or gas chromatography (GC/MS), is the predominant experimental platform for untargeted metabolomics and also plays an important role in other analytical fields requiring high information content and increased sample throughput, such as natural products research. Because of the underlying study designs, these applications regularly bring about high numbers of unidentified mass spectral features, leading to the analytical challenge to identify as many as possible of the corresponding unknown compounds. One of the decisive steps investigating unknown compounds is to determine its molecular formula, which can serve

as a starting point for the structural elucidation. High-throughput molecular formula annotation workflows are required for the analysis of complex biological samples. The vast numbers of unknowns detected in mass spectral data sets acquired from these samples necessitate efficient methods for formula generation, since the process is computationally expensive and error-prone. As an example, a bacterial extract could contain 600–700 unknown substances,[1] and studies of the metabolomes of higher organisms exceed these numbers: the human serum metabolome contains more than 4000 metabolites visible by LC–MS analysis.[2] Especially for nonmodel organisms, an astounding number of metabolites to date remain uncharacterized with respect to their structure and function. Contrary to proteins and other biopolymers, which are constructed from a well-defined set of building blocks, the structure of metabolites is much less defined and thus the structure elucidation process is labor-intensive and usually requires additional techniques like NMR spectroscopy. The capability to efficiently compute high-confidence molecular formulas facilitates this objective.

In the following, we will limit ourselves to biomolecules; that is, molecules that are products of nature or synthetic products with potential bioactivity. Many biomolecules, among them an overwhelming number of substances found ubiquitously across the domains of life, are composed of six elements, i.e., carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur.[3] In contrast, secondary metabolites and other biomolecules, such as drugs or pesticides, occasionally contain less frequently occurring elements, which we refer to as uncommon elements throughout this paper. To name a few examples, marine organisms and some terrestrial bacteria produce halogenated compounds incorporating bromine or chlorine;[4,5] antibiotics containing boron have been reported;[6] metabolites of higher plants can contain selenium;[7] finally, fluorine and iodine have been detected in the metabolism of certain organisms.[8,9] Microbial secondary metabolites are the subject of natural product screening workflows forming the basis for the discovery of novel drug candidates, thus strengthening our motivation to improve computational tools for their identification.[10] However, it is understood that methods developed for biomolecule identification are likely to find application in other analytical areas too, since the underlying fundamental challenge of unknown characterization is common to many varieties of small-molecule investigations using mass spectrometry, ranging from in vitro drug metabolism studies to pesticide screening.

Mass spectrometric instrumentation has seen significant improvement in terms of resolution and mass accuracy over the last 2 decades; however, exact (monoisotopic) mass alone is insufficient to determine the molecular formula of a compound even for sub ppm mass accuracy.[11] To this end, several approaches use the natural isotopic distributions of elements to improve molecular formula determination,[12] assuming a mass accuracy of about 10 ppm or better. Certain approaches limit computation to molecular formulas present in a database;[13] but since many natural compounds are

absent from any database, this restriction is unacceptable. Other approaches compute all candidate molecular formulas (considering a fixed set of elements) that are sufficiently close to the measured peak mass,[14] simulate an isotope pattern for each candidate molecular formula,[11,15–17] and compare it to the measured one.[17–20] Here, the isotope pattern may also contain accurate masses of isotope peaks. Some approaches additionally take the fragmentation pattern of compounds into account.[21,22,23] All of these approaches require the researcher to specify the set of elements to be considered.

For compounds above 400 Da, the number of molecular formulas increases rapidly,[11,17,23] in particular when considering a set of elements beyond CHNOPS. This severely increases not only running time but also chances for false identifications. On the other hand, manually adjusted restrictions on elements, or the maximum allowed number of atoms of a particular element, could exclude the correct molecular formula from the candidate set[24] and hence prevent the discovery of an interesting compound featuring an uncommon element. The presence of certain uncommon elements can be inferred by manual inspection of the isotope pattern in mass spectra, see Figure 1 in ref [25] for an instructive example. Transforming such observations into robust classifiers is nevertheless a challenging problem, as both masses and intensities are perturbed in real-world measurements. Previous efforts along these lines have, for example, been related to estimating the number of chlorine and bromine atoms from electron ionization mass spectra as part of the NIST Mass Spectral Search Program.[26] Here, we present a fully automated method to robustly determine the elements for the compound under study from the isotope pattern. The set of elements determined by our method then serves as input for the next step of the analysis, where accurate masses, isotope pattern (and potentially other information) are used to determine the molecular formula of a compound under investigation. The four most abundant elements in living organisms, carbon, hydrogen, nitrogen, and oxygen, form our basic set of elements. We add phosphorus to this set, as it is relatively common in biocompounds (e.g., nucleotides, ATP) and, in addition, has only a single stable isotope, thus cannot be predicted on the basis of isotope pattern (the latter applies also to iodine and fluorine). We present a method to predict the presence of the elements sulfur (S), chlorine (Cl), bromine (Br), boron (B), and selenium (Se) from isotope patterns.

Our method uses Machine Learning, that is, algorithms that can learn from data, and make predictions based on data. We employ Supervised Machine Learning where the computer is presented with example inputs and desired outputs (the training data), and the goal is to learn rules that map inputs to outputs. For each element, we use a binary classifier that classifies the data into two groups (contains the element vs does not contain the element). For classification, we use random forests.[27] For each element, we present three classifiers based on the number of observed isotope peaks (three, four, and five or more peaks). We find that the more unique the isotopic distribution of an element is, the more

precise our predictions are. Furthermore, availability of more isotopic peaks from the mass spectrum also improves prediction quality. Evaluating the classifiers on three real-world data sets of a total of 663 isotope patterns measured on three different instruments results in no false negative predictions and few false positive predictions for all elements but sulfur. We show that our method significantly decreases the number of molecular formulas that have to be considered by subsequent analysis steps, resulting in a massive decrease of running time, while at the same time also slightly improving identification rates. The method presented here will be integrated in an upcoming release of the SIRIUS 3 software package.[17,23]

7.3   Background

The most common method to determine a molecular formula is to simulate an isotope pattern for each candidate molecular formula and compare it to the measured one.[17–20] This requires a fixed set of elements to be considered for the generation of candidate molecular formulas. Obviously, the number of candidates increases considerably with the size of the set of elements. For example, consider *rifampicin* with monoisotopic mass 822.405 Da and mass accuracy 6 ppm: we reach 2 358 candidates for the set of elements CHNOP and 117 029 candidates for CHNOPSClBr, an almost 50-fold increase caused by only three uncommon elements. (The above numbers ignore molecular formula filters such as the "Senior rules",[28](but the effect is comparable.) On the theoretical side, the number of decompositions of a certain mass can be approximated with high accuracy using a polynomial:[17] Assuming a relative error, this polynomial has degree k for k elements.

Isotopes are variants of an element with different numbers of neutrons. Different isotopes of the same element have nearly identical chemical properties but different mass. Most elements in nature have more than one stable isotope, and each of these isotopes occurs in nature with a certain abundance.[29] The totality of isotopes is the isotopic distribution of an element. The mass difference of successive isotopes differs from element to element, despite the fact that in all cases, one or more neutrons are added: For example, the mass difference of ^{12}C and ^{13}C is 1.00335 Da (Dalton), whereas the mass difference of ^{10}B and ^{11}B is 0.99637 Da. The different isotope abundances and masses of the isotopes that are contained in a compound generate a characteristic set of peaks, called an isotope pattern. In comparison to C, H, N, O, and P (referred to as "CHNOP" in the following), the elements S, Cl, Br, B, and Se lead to more "distinctive" isotope patterns, as the most abundant isotope of these elements has relative abundance less than 95%.

In this paper, we will not consider the isotopic fine structure (isotopologues) of a compound; instead, we limit ourselves to the mean peak masses[17] (or accurate masses[16]) of isotopic peaks, combining all

isotopologues with identical nominal mass. Throughout this paper, the isotopologue where each atom is the isotope with the lowest nominal mass is referred to as monoisotopic, see, for example, ref [30]. For certain elements such as boron or selenium, this is not the most abundant isotope. We refer to the peak at the monoisotopic mass as the monoisotopic peak or + 0 peak, which is followed by the + 1, + 2 ... peaks. Referring to a peak "before" ("after") another peak, means it has a smaller (higher, respectively) mass.

The isotopic distribution of elements influences the isotope pattern of the compound in both mass differences and intensities of peaks. Small compounds containing only CHNOP have an intense monoisotopic peak; the + 1 peak has much lower intensity, and intensity of subsequent peaks decreases further. The more "distinctive" isotopic distributions of S, Cl, Br, B, and Se are reflected in the shape of the isotope pattern of a compound, see Figure 1 and section S.1, Supporting Information.
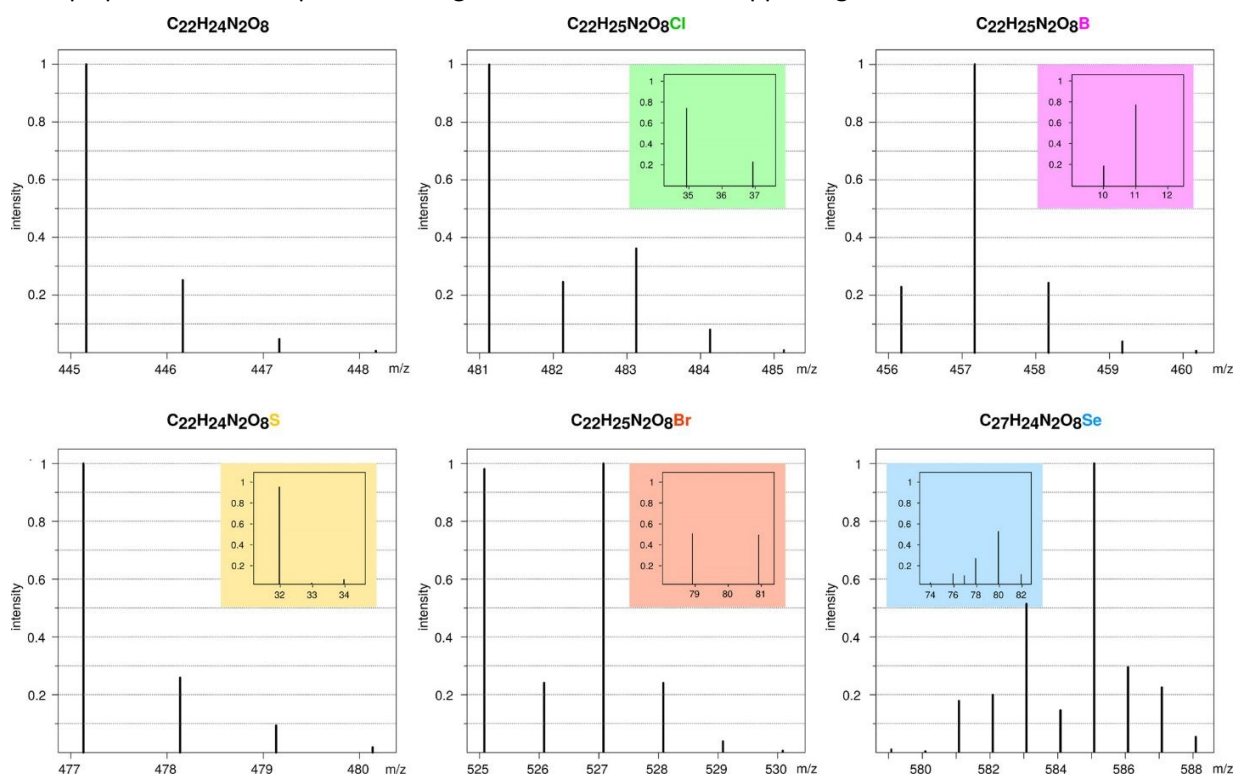
Figure 1. *Effect of the uncommon elements S (yellow), Cl (green), Br (red), B (magenta), and Se (blue) on the isotope pattern. [M + H]$^+$ ionization is assumed. Individual isotopic distributions of the corresponding uncommon elements are shown in the colored boxes. All molecular formulas come from the molecular structure databases.*

7.4   Methods

We use a set of binary classifiers to predict the presence of uncommon elements from isotope patterns of compounds using supervised machine learning. We create one classifier for the presence of each of the uncommon elements Cl, Br, B, S, and Se. Further, we create a classifier "CHNOPS" for compounds that contain only the elements C, H, N, O, P, and S; we use this classifier to demonstrate the

discriminative power of the isotope pattern for identifying uncommon elements. The classifier cannot rule out the presence of fluorine or iodine in a compound or any other element that has only a single stable isotope.

To learn and predict the presence of uncommon elements from isotope patterns, the patterns need to be transformed to a set of numerical features characterizing the data. Those features are based on the masses and intensities of the peaks. Measured isotope patterns vary in the number of detected peaks. In particular for small compounds, we get very short isotope patterns, as only a few peaks are intense enough to be detected as signal. Using classifiers for arbitrary peak number would result in missing (peak) features for smaller isotope patterns, making the training more complicated. Hence, we construct different classifiers for isotope patterns of size three, four, and five. In total, we train 18 classifiers (five uncommon elements plus CHNOPS times three isotope pattern sizes).

Isotope patterns with less than three peaks do not contain enough information for uncommon element prediction. However, for isotope patterns with less than three peaks, the presence of an uncommon element is very unlikely since the + 2 should be intense enough for detection if an uncommon element is present (see Figure 1). For isotope patterns with more than five peaks, we consider only the first five peaks. Classifiers for isotope patterns with more than five peaks did not improve the results significantly (data not shown).

7.4.1 Features

We use features based on the intensity and mass of the isotope peaks. The full list of features can be found in Table S.1. The number of features varies according to the number of peaks in the isotope pattern: There are 21 (38, 39) intensity features and 4 (7, 11) mass features for three (four, five, respectively) isotope peaks.

7.4.2 Intensity Features

We use the intensity of each peak as well as the minimal, maximal, and median peak intensity. Since the presence of some elements results in zigzag shaped isotope patterns, we further use the sums, minima, and maxima of intensities for even and odd peaks, respectively. Furthermore, we use the index of the most intense, second most intense, and third most intense peak. We use quotients and differences of all pairs of peak intensities for the first four peaks. Finally, we use all combinations of sums of peak intensities.

7.4.3 Mass Features

We use monoisotopic mass, and all pairwise mass differences between isotope peaks as features. Mass is included, since the influence of uncommon elements on the shape of the isotope pattern is stronger for small compounds than for large ones. The mass difference between consecutive peaks of the isotope pattern is influenced by elements with more than one stable isotope: For example, if Cl is present, the mass difference between the monoisotopic peak and + 2 peak decreases due to the mass difference of 1.99705 Da between ^{35}Cl and ^{37}Cl.

7.4.4 Random Forests

For classification we use random forests, a supervised machine learning method.[27] A random forest consists of a set of unpruned decision trees. Each tree is trained on a random sample (with replacement) of the training data. This technique is called bagging.[31] For the determination of a split (a node in a decision tree), we use a subset of $m \approx \sqrt{M}$ features, where M is the total number of features.[32] We use the random forest implementation of the Mahout library (version 0.9, https://mahout.apache.org/) for training.

We set $m = 5$ for isotope patterns of size three, and $m = 7$ for isotope patterns of size four and five. For each classifier, we generate a random forest consisting of 100 trees. For all other parameters we use the default values of the Mahout library. The classes DecisionTreeBuilder and Bagging are used to construct the trees. By default, the minimum set size for a split is 2 and the minimum variance proportion is 0.001. We use the OptIgSplit class to compute the best split.

For classification, the input is run through all of the decision trees in the random forest, and the final classification is done by voting: every decision tree votes for true or false. A threshold for the ratio of positive votes is used to determine the classification of the forest.

7.4.5 Data Sets

For the training and evaluation of the classifiers, we want to use a very large set of isotope patterns. Unfortunately, there is only limited experimental data available for isotope patterns of biomolecules. In this paper, we use three measured data sets from different instruments for evaluation; in addition, we use simulated isotope patterns of molecular formulas from several compound databases (see Table 1).

Table 1. *Overview of the Datasets Used for Training and Evaluation*

	simulated		measured		
	training	evaluation	*myxo*	*pesticide*	*CASMI*
all	1 128 059	51 097	88	43	532
CHNOPS	604 506	42 713	78	15	370
sulfur S	502 529	9 892	24	14	146
chlorine Cl	345 799	5 502	8	27	152
bromine Br	171 596	3 015	0	1	13
boron B	56 808	141	1	0	0
selenium Se	16 680	98	1	0	0

Total number of molecular formulas and numbers of molecular formulas that are positive examples for the different classifiers are given. In addition to the simulated evaluation set, the measured datasets are used exclusively for evaluation.

The *myxo* data set consists of 88 isotope patterns measured on a Bruker MaXis 2G qTOF spectrometer (Bremen, Germany). The corresponding compounds are secondary metabolites from *Myxobacteria*, with several exceptions mentioned below. Compounds range in mass from 192.009 to 2213.962 Da, with an average mass 623.277 Da and median mass 591.307 Da. From these compounds, 24 contain sulfur (20 contain a single sulfur atom, 2 contain two, and 2 contain three sulfur atoms), 8 contain chlorine (6 with one chlorine atom, 2 with two chlorine atoms), and one compound contains a single boron atom. We added selenomethionine to this data set as a positive example for a selenium-containing compound. Nine compounds have a measured isotope pattern of length three, 47 of length four, and 32 of length five or more (3 of length six, 5 of length seven, 1 of length nine). The myxo data set is available from https://bio.informatik.uni-jena.de/data/.

The *pesticide* data set is taken from Stravs et al.[22] The data set consists of isotope patterns from different pesticides measured on a LTQ Orbitrap XL from Thermo Fisher Scientific (San José, CA) with electrospray ionization in positive and negative modes. Compounds range in mass from 198.056 to 443.125 Da, with average mass 275.353 Da and median mass 260.016 Da. From the 60 mass spectra, 43 show an isotope pattern with at least three peaks. For the 18 shorter isotope patterns, none of the compounds contains any of the elements SClBrBSe. For the remaining 43 isotope patterns, 14 isotope patterns contain a single sulfur atom, 27 chlorine (22 with a single chlorine atom, 4 with two, and 1 with three chlorine atoms), and one isotope pattern contains a single bromine atom. The data set contains 18 isotope patterns with a measured isotope pattern of length three, 18 of length four, and 7 of length five or more (1 of length six).

The *CASMI* data set was measured by Martin Krauss (Helmholtz Centre for Environmental Research, Leipzig, Germany) and processed by Emma Schymanski (Eawag, Dübendorf, Switzerland) as part of the

Critical Assessment of Small Molecule Identification Challenge 2016 (http://www.casmi-contest.org/2016/) using RMassBank.[22] Measurements were performed on a Q Exactive Plus Orbitrap (Thermo Scientific) with electrospray ionization in positive and negative mode. MS1 spectra were extracted for substances with $[M + H]^+$ (positive) and $[M – H]^-$ (negative mode) ions. We removed one compound containing silicon. The remaining data set contains 628 independent mass spectra from 512 compounds. Compounds range in mass from 67.042 to 776.687 Da, with an average mass 259.500 Da and median mass 246.071 Da. From the 628 mass spectra, 532 show an isotope pattern with at least three peaks. For the 96 shorter isotope patterns, none of the compounds contain any of the elements SClBrBSe. For the remaining 532 isotope patterns, 146 patterns contain sulfur (128 with a single sulfur atom, 17 with two, and 1 with four sulfur atoms), 152 chlorine (84 with a single chlorine atom, 42 with two, 19 with three, 3 with four, and 3 with six chlorine atoms) and 13 bromine (11 with a single bromine atom and 2 with two bromine atoms). The data set contains 140 isotope patterns with a measured isotope pattern of length three, 220 of length four, and 172 of length five or more (49 of length six, 2 of length seven, and 1 of length 8).

For training and evaluation, we use simulated isotope patterns of molecular formulas of compounds from 11 compound databases: ChEBI (Chemical Entities of Biological Interest),[33] ChEMBL version 19,[34] DrugBank version 4.2,[35] HMDB (Human Metabolome Database) version 3.6,[36] Indofine (http://indofinechemical.com/), KEGG (Kyoto Encyclopedia of Genes and Genomes),[37] KNApSAcK,[38] MolMall (http://www.molmall.net/), PubChem,[39] UNPD (Universal Natural Product Database),[40] and ZINC (ZINC Is Not Commercial) version 12,[41] see section S.3 and Table S.2 for details.

Some of the databases, in particular PubChem, contain many records which are uncommon for biomolecules. To remove such molecular formulas, we filtered all data sets using the following ad hoc rules: (1) the compound has a monoisotopic mass between 100 and 1500 Da; (2) the compound contains only elements from CHNOPSClBBrSe and contains at least one carbon and one hydrogen atom; (3) the compound contains at most five atoms of sulfur, chlorine, boron, or bromine. For the training data only, we also discarded (4) charged or generic compounds or complexes. We stress that molecular formulas passing these restrictions do not necessarily correspond to biomolecules, and that biomolecule molecular formulas may fail one or more of these rules; these data are used to train and evaluate our method.

We split the databases into a large training set that is used to construct the classifiers and a smaller evaluation set. As the training set, we use PubChem, ChEMBL, and ZINC. PubChem is the largest data set and contains reasonably many compounds even for Se and B. It is particularly useful to train cases such as compounds incorporating a high number of uncommon elements or unlikely combinations of uncommon elements. We observe that molecular formulas from ChEMBL and ZINC are almost completely covered by

PubChem. As the evaluation set, we use HMDB, ChEBI, DrugBank, Indofine, KEGG, KNApSAcK, MolMall, and UNPD.

Several molecular formulas are contained in more than one database; clearly, we consider only a single occurrence of each. Further, the training set should be independent from the evaluation set. To this end, all molecular formulas from the evaluation set are removed from the training set. In total, the training set contains 1 128 059 molecular formulas, and the evaluation set contains 51 097 molecular formulas (see Table 1).

7.4.6 Isotope Pattern Simulation

For the simulation of isotope patterns, we use SIRIUS[17] to compute the exact isotope pattern. For the training and evaluation set, we add noise to the peak intensities to simulate measured isotope patterns. In Machine Learning, adding noise often improves classification results, as it forces the method to search for complex patterns and improves generalizability. It is not necessary that the added noise has the same characteristics as noise observed in the real-world data. For each peak, we calculate the simulated intensity $I = I_{exact} \mathcal{N}(1, \sigma_{IR}) + \mathcal{N}(0, \sigma_{IA})$ where $\mathcal{N}(1, \sigma_{IR})$ is a normally distributed relative noise with mean one and $\mathcal{N}(0, \sigma_{IA})$ is a normally distributed absolute noise with mean zero. After adding noise to all peaks of an isotope pattern, the peak intensities are normalized to 100%.

In addition, we add noise to the peak masses. For each peak, we calculate the simulated mass $m = m_{exact} + \mathcal{N}(0, \sigma_M)$, where $\mathcal{N}(0, \sigma_M)$ is an additive, normally distributed absolute noise with mean zero. We are aware that mass deviation is not constant throughout the mass range but rather depends on the mass of the compound. Nevertheless, we refrain from using a multiplicative mass noise: The features for the classifiers are built on mass differences between the peaks instead of peak masses. To our knowledge, the distribution of deviation of dependent mass differences has not been investigated. To make our classifiers robust, we prefer to consider the worst case scenario, which is an additive noise with the highest mass deviation we have observed.

7.4.7 Noise Parameters of the Evaluation Set

To find realistic noise parameters, we investigate measured isotope patterns from a Bruker Maxis 2G qTOF mass spectrometer, measured at different concentration levels. On the basis of the accuracy of these patterns, we define a standard noise profile with standard deviations $\sigma_{IA} = 0.0015$, $\sigma_{IR} = 0.04$, and $\sigma_M = 0.0013$ Da. In addition, we define a high noise profile with standard deviations $\sigma_{IA} = 0.006$, $\sigma_{IR} = 0.07$, and $\sigma_M = 0.0018$ Da to generate particularly hard cases. The reader is reminded that we measure mass

differences between isotope peaks. Both noise profiles are used for generating isotope patterns for the evaluation set.

7.4.8 Noise Parameters of the Training Set

Instead of using exact simulated isotope patterns, we found that adding noise to the training data improves classification results. We found that using relatively high noise during training, improves classification results even for the standard noise profile. Thus, for simulating the training set, we choose noise parameters that are higher than the standard noise profile of the evaluation set, namely, $\sigma_{IA} = 0.005$, $\sigma_{IR} = 0.05$, and $\sigma_M = 0.0015$ Da.

7.4.9 Learning Phase

For most classifiers, we have many more negative examples than positive examples in the training set. Because of the optimization function of random forests, unbalanced data leads to suboptimal classification results for the smaller class.[42] For each classifier, we generate a training subset by randomly drawing without replacement to reach the same number of positive and negative examples. The CHNOPS, S, Cl, Br, and B classifiers are trained using a subset of 100 000 molecular formulas (50 000 positive and 50 000 negative examples) and Se with 25 000 molecular formulas (12 500 positive and 12 500 negative examples).

There exist improved downsampling techniques to balance the data, such as distance based methods.[43] Because of the large number of compounds and high dimensionality of the training data, computation would be very time-consuming. We found that random downsampling delivers very good results and refrained from testing improved techniques.

7.5   Results

For each element, we present three classifiers based on the number of observed isotope peaks (3, 4, and 5 or more peaks). Classifiers are trained using simulated isotope patterns of more than a million compounds. The resulting classifiers give excellent separation, with the area under curve (AUC) above 0.99 for all elements but sulfur. We then concentrate on classifiers with very high sensitivity, so that we almost never miss an occurrence of an uncommon element. In contrast, we put less focus on the specificity of our classifiers, resulting in more false positive predictions for certain elements. The reasoning for doing so is that our classifiers are meant as a first step of an automated pipeline for isotope pattern analysis: After generating and scoring candidates, the downstream method can sort out cases where our classifiers made

a false positive prediction. In contrast, if our classifiers make a false negative prediction, wrongly stating that some uncommon element is absent from a compound, the downstream analysis has no means to correct this.

Training of all 18 classifiers required 3.5 h on two Intel Sandy Bridge processors at 2.3 GHz with six cores and 64 GB RAM. We find that training requires more time and the decision trees are larger, the less unique the isotopic distribution of the element is (see Tables S.3 and S.4). For example, in the random forest of the four peaks sulfur classifier, the resulting trees have a median of 9 450 nodes (median tree depth 47). Construction of this random forest required 23 min. In contrast, for the four peaks boron classifier, trees contain only a median of 13 nodes (depth four) and learning required only 7 min. For an example of a decision tree, see Figure 2.
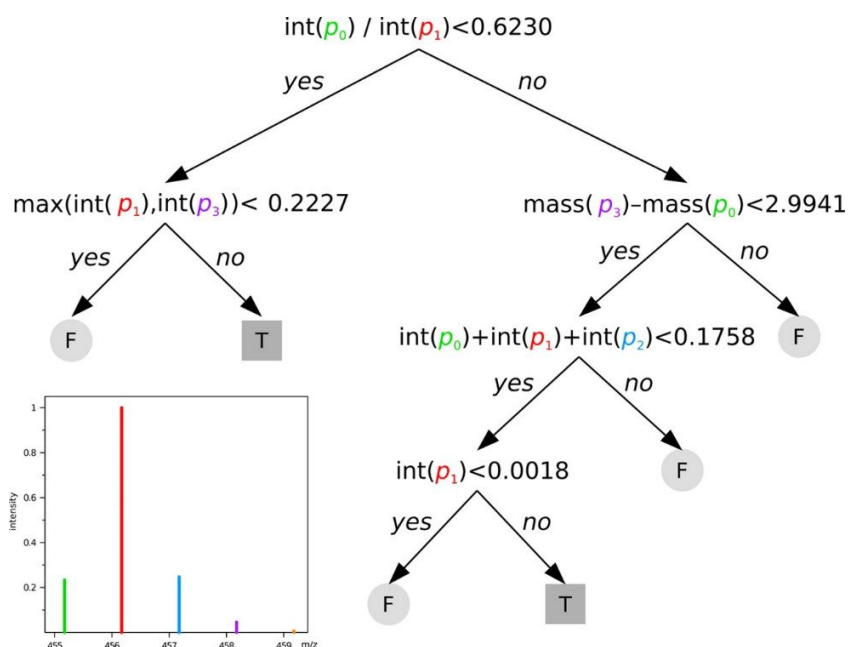
Figure 2. *Example decision tree from the random forest of the boron classifier for isotope patterns with five peaks. This tree is combined with 99 other decision trees to reach the final decision whether or not boron is present in the unknown compound.*

7.5.1 Prediction Quality on Simulated Data

There exist two types of wrong classifications: On the one hand, a classifier may return false ("element not present") although the element is part of the compound (false negative); on the other hand, a classifier may return true ("element present") when an element is not part of the compound (false positive).

In Machine Learning, one is generally interested in classifiers with high sensitivity (true positive rate)

$$\text{sensitivity} = \text{true positive rate} = 1 - \text{false negative rate}$$
$$= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$
$$\text{specificity} = \text{true negative rate} = 1 - \text{false positive rate}$$
$$= \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

and high specificity (1 – false positive rate), where   .

It must be understood that reaching both high sensitivity and specificity simultaneously is usually not possible and that we might have to trade in sensitivity for specificity or vice versa. For the downstream analysis we have in mind (identification of the molecular formula), false negative predictions are much worse than false positive predictions. Once an element that is actually part of the compound is not considered for molecular formula identification, it is difficult or impossible to counteract this failure in the subsequent analysis steps. In contrast, a false positive prediction will "only" increase the running time and the probability of identifying a wrong molecular formula. Thus, sensitivity needs to be very high, even at the cost of specificity. For the CHNOPS classifier, which is included to evaluate the overall power of the approach, we treat sensitivity and specificity as equally important.

Receiver Operating Characteristic (ROC) and the area under the ROC curve (AUC) can be used to evaluate the ability of a classifier to distinguish between true and false examples. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at all possible threshold settings of the classifier. When decreasing the threshold, both the true positive rate and the false positive rate increase. A perfect classifier would result in an area under curve of 1, that is, there is a threshold where the classifier reaches TPR of 1 without predicting any false positives. Random classifiers have an area under the curve of 0.5.

ROC curves for the classifiers with three peaks are shown in Figure 3. Results for four and five peaks can be found in Figure S.1 in the Supporting Information. For all classifiers, we achieve area under curve above 0.97 even for the high noise profile, see Table 2. The lowest area under curve was achieved for the S classifier on three peaks. Compared to the other elements to be detected, sulfur has the least distinctive isotopic distribution and is hard to distinguish from isotope patterns containing only CHNOP. The more unique the isotopic distribution of the elements get, the higher the area under curve of the corresponding classifier. For Se and B we get an area under the curve of 1 in all cases.
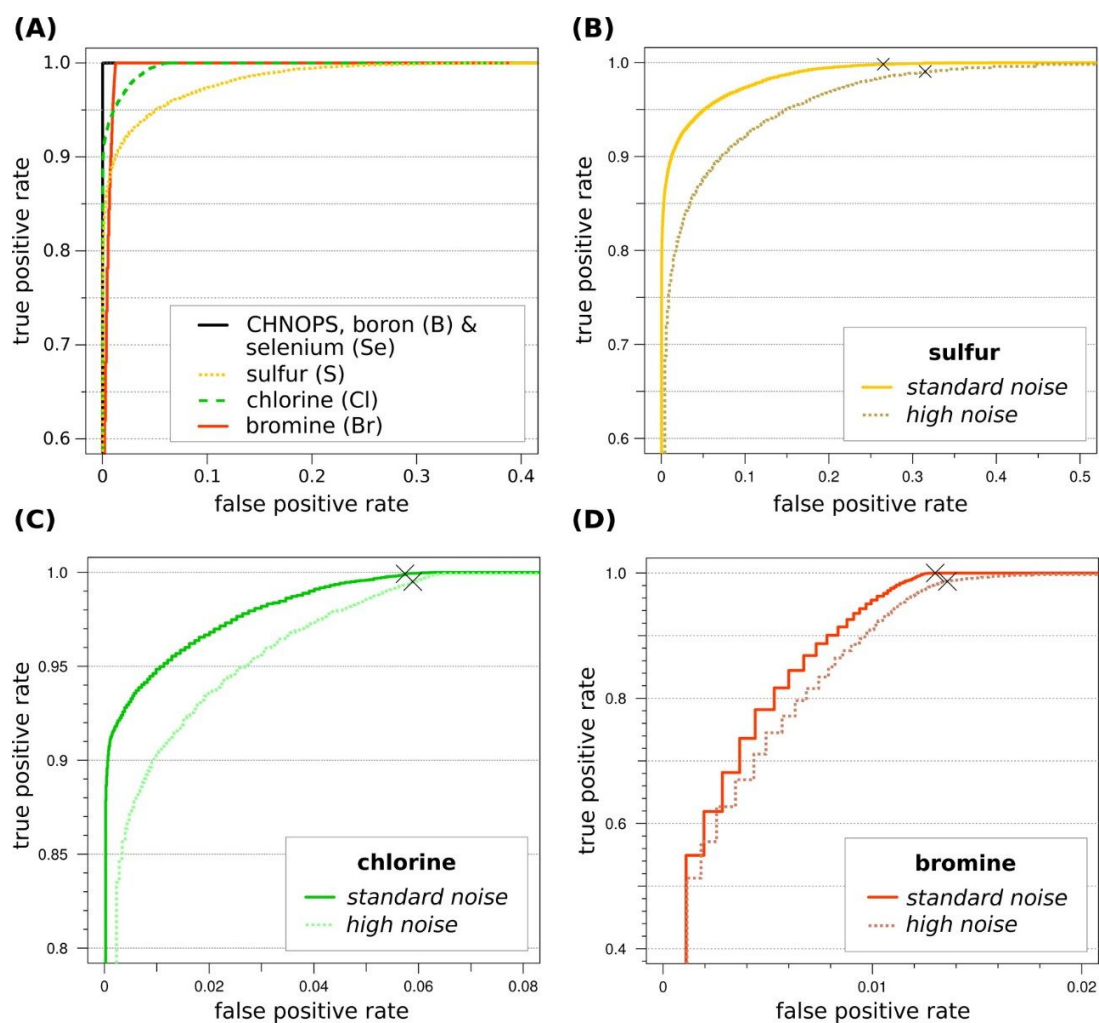
Figure 3. *(A) ROC curves for all classifiers using three peaks at the standard noise profile. (B–D) ROC curves for S (yellow), Cl (green), and Br (red) using three peaks at the standard noise profile and the high noise profile. In comparison to part A, ROC curves parts C and D are zoomed in to the top left corner. AUCs for all ROC curves can be found in Table 2. The chosen trade-offs between FPR and TPR for the analysis on the measured data sets are marked with ×.*

Table 2. *AUCs for the Standard Noise Profile and the High Noise Profile*

noise profile:	standard			high		
no. of peaks:	three	four	five	three	four	five
CHNOPS	1.0	1.0	1.0	1.0	1.0	1.0
S	0.992	0.996	0.999	0.974	0.985	0.996
Cl	0.998	0.999	1.0	0.994	0.997	0.999
Br	0.997	0.997	1.0	0.994	0.997	1.0
B	1	1	1	1	1	1
Se	1	1	1	1	1	1

If decimals are given, the given values are rounded.

We find that for all elements, area under curve increases for classifiers using isotope patterns with more peaks. This is not surprising, since isotope patterns with more peaks contain more information. In particular for the high noise profile, using more peaks improves the prediction.

To further evaluate the trade-off between sensitivity and specificity, we analyze the FPRs for all 18 classifiers at fixed TPRs for the standard noise profile (see Table 3; results for the high noise profile can be found in Table S.5). For Cl, FPR is below 6.5% in all cases, and for Br, FPR is below 1.5% in all cases. As expected, FPR improves when using more isotope peaks. For S, we get higher FPRs than for the other classifiers. For three peaks, we do not reach a TPR of 100%.

Table 3. *How Many False Positive Predictions Do We Have to Accept in Order to Reach a Certain Sensitivity?*

		three peaks			four peaks			five peaks		
	desired TPR	**0.998**	**0.999**	**1**	**0.998**	**0.999**	**1**	**0.998**	**0.999**	**1**
	CHNOPS	0.0	**0.0**	0.005	0.0	**0.0**	0.001	0.0	**0.0**	0.001
	S	**0.265**	0.313	N/A	**0.162**	0.173	0.332	**0.055**	0.072	0.160
	Cl	0.056	**0.057**	0.063	0.054	**0.055**	0.060	0.004	**0.008**	0.027
resulting FPR	Br	0.012	0.013	**0.013**	0.012	**0.012**	0.013	0.0	0.0	**0.0**
	B	0.0	0.0	**0.0**	0.0	0.0	**0.0**	0.0	0.0	**0.0**
	Se	0.0	0.0	**0.0**	0.0	0.0	**0.0**	0.0	0.0	**0.0**

We report false positive rates (FPR) for all 18 classifiers at fixed true positive rates (TPR). TPR 0.998 corresponds to two missed positive predictions in 1000 examples; TPR 1 means missing not a single positive example. Results for the standard noise profile. FPR and TPR used for the analysis on the measured datasets in **bold**. "N/A", classifier cannot reach the desired TPR.

We additionally evaluated the B classifiers against isotope patterns with an extremely high noise level (σ_{IA} = 0.008, σ_{IR} = 0.1, and σ_M = 0.002 Da) to account for the range of natural variation in the isotopic distribution of B. The prediction quality remains the same. For CHNOPS, we get a FPR 0.0 up to a TPR of 99.9%. Even for 100% true positives, the FPR is below 1%.

We manually chose the trade-off between FPR and TPR used for the analysis on the measured data sets, see Table 3. If several voting thresholds fall into the particular FPR/TPR trade-off, we choose a reasonable one (see Table S.6).

We have also tested the training of different classifiers for different mass ranges, since the mass of a compound influences the shape of its isotope pattern. However, using two or three classifiers for different mass ranges did not improve prediction quality (data not shown).

7.5.2   Prediction Quality on Measured Data

In the second part of our evaluation, we apply our classifiers to three measured data sets. The measured isotope patterns have different numbers of peaks; for each pattern, we choose the appropriate classifier. For each classifier, we use the voting threshold with the best trade-off between FPR and TPR (see Table 3 and Table S.6). Note that there is only one positive example each for B and Se.

Table 4. *Classification Results on the Three Measured Datasets*

| | | real | | predicted | | | |
		P	N	TP	TN	FP	FN
myxo	CHNOPS	78	10	78	10	0	0
	S	24	64	24	46	18	0
	Cl	8	80	8	79	1	0
	Br	0	88	0	88	0	0
	B	1	87	1	87	0	0
	Se	1	87	1	87	0	0
pesticide	CHNOPS	15	28	15	28	0	0
	S	14	29	14	27	2	0
	Cl	27	16	27	15	1	0
	Br	1	42	1	42	0	0
	B	0	43	0	43	0	0
	Se	0	43	0	43	0	0
CASMI	CHNOPS	368	164	368	164	0	0
	S	146	386	146	366	20	0
	Cl	152	380	152	375	5	0
	Br	13	519	13	516	3	0
	B	0	532	0	532	0	0
	Se	0	532	0	532	0	0

Number of real positives (P) and negatives (N) as well as number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. Recall that false positive predictions (orange) can be corrected in the downstream analysis of molecular formula identification; in contrast, false negative prediction (red) cannot.

See Table 4 for prediction results. In no case do we observe a false negative prediction for elements S, Cl, Br, B, and Se. To this end, our classifiers have perfect sensitivity for the measured data. A false negative prediction would result in an element being excluded from further analysis, despite being present in the compound. Recall that false positive prediction, in contrast, can be corrected in the downstream analysis of identifying the exact molecular formula. For Cl, we get seven false positive predictions: See Figure S.2 for the isotope patterns of bromazil ($C_9H_{13}BrN_2O_2$) and pinensin A ($C_{96}H_{139}N_{27}O_{30}S_2$). We note that pinensin A with mass 2213.962 Da is much larger than the mass range used for training our classifiers. For both compounds, the CHNOPS classifier reports a conflicting prediction. For Br, we have three false positive predictions in the *CASMI* data set. For B and Se we do not get any false positive predictions.

For S, we get 40 false positive predictions and, thus, would use an unnecessary element for the molecular formula computation. Of these, 18 compounds are from the *myxo* data set. Compounds in the *myxo* data set are much larger than compounds in the other data sets; for larger compounds, the presence of sulfur is more veiled, in particular if only a single sulfur atom is contained. See Figure S.3 for an example.

7.5.3   Molecular Formula Identification

To show the use of our predictors in a pipeline for molecular formula identification, we integrated it into the SIRIUS pipeline for molecular formula identification using isotope patterns,[17] see refs [24] and [44] for details. Since biomolecules in the three data sets may potentially contain elements CHNOPSClBrBSe, we use this set of elements when running SIRIUS to compare against. For our predictors, we use elements CHNOP plus those from SClBrBSe that were predicted (including false positive predictions) to be present in the biomolecule. We use a mass deviation of 10 ppm.

For the *myxo* data set, we excluded pinensin A ($C_{96}H_{139}N_{27}O_{30}S_2$, 2213.962 Da). Processing this compound using the alphabet CHNOPSClBrBSe, we ran into an out-of-memory exception. Processing it with the predictor-based version, 19 746 670 candidate molecular formulas were evaluated in 15 min, and the correct molecular formula was at rank 10414. For three compounds in the *myxo* data set, we had to extend the mass deviation to 30 ppm to include the correct molecular formula.

For the *CASMI* data set, we excluded four compounds and mass spectra, as these were solvents and contain an exceptionally high number of fluorines. For the remaining compounds, we used an upper bound of 6 fluorines (the remaining compound with the highest number of fluorines was flufenoxuron, $C_{21}H_{11}C_lF_6N_2O_3$, an insecticide). We assume for both approaches that we know whether iodine is present in the compound; as discussed below, the presence of iodine can be determined using the fragmentation

pattern of the compound. All running time measurements were performed on an Intel Sandy Bridge processor at 2.3 GHz with 128 GB RAM.

We observe a massive decrease in the number of candidate molecular formulas we have to consider (see Table 5): We observe a 32-fold to 261-fold decrease in candidate molecular formulas filtered with SENIOR rules,[28] compared to the default method. As expected, this results in a similar decrease in running times: The total running time for processing all three data sets decreased from 18.6 min to 9.5 s. We also observe a slight improvement in identification performance.

Table 5. *Results of the Molecular Formula Identification Using SIRIUS*

	data set	*myxo*	*pesticide*	*CASMI*
	no. of spectra	87	43	528
	median mass (Da)	592.315	260.016	281.484
	correct identifications	12	13	287
	top 3 identifications	20	36	360
	top 10 identifications	30	42	433
without element prediction	avg no. decompositions			
	...before SENIOR filtering	2737085	2242.0	44007.4
	...after SENIOR filtering	571072.4	836.2	13052.5
	avg running time (ms)	11277.2	20.1	256.9
	correct identifications	13	13	287
	top 3 identifications	21	37	364
	top 10 identifications	32	42	435
with element prediction	avg no. decompositions			
	...before SENIOR filtering	12708.8	73.4	597.6
	...after SENIOR filtering	2181.0	26.1	174.9
	avg running time (ms)	64.5	1.0	7.2

We use CHNOPSClBrBSe when running the identification without element prediction (top). For our predictors, we use CHNOP plus those elements from SClBrBSe that were predicted to be present (bottom). For *CASMI*, we also added I to the set of elements for all compounds that contain iodine and allow up to six fluorines for all compounds. We report correct identifications (top 1, top 3, top 10), average number of decompositions (before and after SENIOR filtering) and average running time for all three datasets.

7.6 Conclusion

We have presented classifiers to predict the presence of uncommon elements, based on the isotope pattern of the biomolecule. The thresholds have been chosen with regards to the setting discussed in the introduction, where an extremely high sensitivity is more important than a low false positive rate. Evaluating the classifiers on a real-world data set, we found no false negative predictions and would, thus,

never miss an element for molecular formula identification. Depending on the application, it is also possible to select thresholds based on other criteria, e.g., keeping the false positive rate low.

We have shown how our classifiers improve the subsequent steps of analyzing isotope patterns, in particular by significantly reducing the number of candidate molecular formulas we have to consider. Molecular formula identification based on isotope patterns alone is not adequate in practice (see Table 5), but results can be much improved by accompanying the isotope pattern data by fragmentation data: For example, for 56 of 60 compounds in the complete pesticide data set, Böcker and Dührkop[23] inferred the correct molecular formula using both data types combined; in all cases, the correct answer was ranked in the top 5. Running time differences will be much larger if we also analyze tandem mass spectra, as processing a single candidate may require several seconds: Böcker and Dührkop[23] report an instance where 3106 candidates were considered, requiring a total of 12.5 h for this compound.

We have used random forests for the prediction as they deliver high-quality predictions and are not susceptible to noise, as has been shown in the noise parameter evaluation. Classifiers have been trained on simulated data with relatively high noise and have not been tailored toward any particular instrumental platform. Training on instrument-specific data may further increase prediction quality.

From the uncommon elements we have considered in this paper, sulfur is the most abundant one (Table 1). It is also the element which is hardest to predict, since the effect of a single sulfur atom on the isotope pattern is less pronounced than for other uncommon elements. We have not taken into account the presence of the characteristic peak at + 1.995796 Da from the monoisotopic peak in the isotopic fine structure, as detection of this peak is dependent on the resolution of the MS instrument, which is not covered in this paper. Assuming adequate resolution, combining the classifiers presented here with a classifier based on the + 1.995796 Da peak will result in improved classification performance for sulfur.

Using the isotope pattern alone, it is not possible to predict the presence of iodine and fluorine. To a certain degree, it is possible to predict iodine using tandem mass spectral data, based on characteristic common losses and fragments.[23] In contrast, the presence of fluorine is hard to predict using tandem MS data, too; hence, it might be advisable to allow for a few fluorine atoms in the molecular formula candidates, if there is any reason to believe that fluorine might be present in the sample compounds.[4,8] Predicting the presence of other elements that have characteristic isotope distributions, such as silicon, is also possible by the methods presented here if required by the application.

Selenium-containing compounds result in a monoisotopic peak with small intensity, which may not be detected in the mass spectrum. In this case, we cannot proceed by decomposing the monoisotopic peak. Instead, we can decompose the average mass (that is, the molecular mass) of the compound;[17] the

average mass can be estimated from the experimental spectrum by taking the weighted average of isotope peak masses.

Until now, the selection of elements to include for the calculation of molecular formulas has been a manual decision. This has not been seen as a critical bottleneck in the past, as molecular formula determination was mainly performed as a manual "low throughput" procedure in the course of structural elucidation. Today, methods for small molecule analysis are being applied at a large scale in numerous research fields, including diverse applications such as the investigation of industrially relevant plants and microbes, clinical pharmacology studies, and natural products discovery. The underlying mass spectrometry applications share a tendency to produce data sets from which hundreds to thousands of unknown molecules have to be analyzed. Generation of molecular formulas is among the most commonly performed first-pass analyses in unknown identification workflows, and several advanced methods rely on this upfront evaluation step.[23,45] Whereas efforts have been made to evaluate the fit of a measured isotope pattern against the theoretical isotope pattern of a molecular formula as well as evaluation of the corresponding statistical significance and robustness of the method,[46] the issue of predicting the presence or absence of elements has been widely disregarded. To this end, we believe that our method is an important step in the automated annotation of novel biomolecules from mass spectrometry data.

7.7 References

[1] T. Hoffmann, D. Krug, S. Hüttel, R. Müller, *Anal. Chem.* **2014**, *86*, 10780.

[2] N. Psychogios, D. D. Hau, J. Peng, A. C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam et al., *PLoS ONE* **2011**, *6*, e16957.

[3] H. J. Issaq, Q. N. Van, T. J. Waybright, G. M. Muschik, T. D. Veenstra, *Journal of separation science* **2009**, *32*, 2183.

[4] D. G. Fujimori, C. T. Walsh, *Curr. Opin. Chem. Biol.* **2007**, *11*, 553.

[5] P. M. Pauletti, L. S. Cintra, C. G. Braguine, A. A. da Silva Filho, M. L. A. E. Silva, W. R. Cunha, A. H. Januário, *Marine drugs* **2010**, *8*, 1526.

[6] S. I. Elshahawi, A. E. Trindade-Silva, A. Hanora, A. W. Han, M. S. Flores, V. Vizzoni, C. G. Schrago, C. A. Soares, G. P. Concepcion, D. L. Distel et al., *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 304.

[7] a) L. Yuan, Y. Zhu, Z.-Q. Lin, G. Banuelos, W. Li, X. Yin, *PLoS ONE* **2013**, *8*, e65615; b) Ö. Çakır, N. Turgut-Kara, Ş. Arı, *InTech*, **2012**.

[8] D. O'Hagan, C. Schaffrath, S. L. Cobb, J. T. Hamilton, C. D. Murphy, *Nature* **2002**, *416*, 279.

[9] V. M. Dembitsky, *Russian Journal of Bioorganic Chemistry* **2002**, *28*, 170.

[10] a) A. L. Harvey, *Drug Discovery Today* **2008**, *13*, 894; b) W. H. Gerwick, B. S. Moore, *Chem. Biol.* **2012**, *19*, 85; c) G. M. Cragg, D. J. Newman, *Biochim. Biophys. Acta* **2013**, *1830*, 3670.

[11] T. Kind, O. Fiehn, *Bmc Bioinform.* **2006**, *7*.

[12] K. Scheubert, F. Hufsky, S. Bocker, *J.Cheminform.* **2013**, *5*, 12.

[13] F. Matsuda, Y. Shinbo, A. Oikawa, M. Y. Hirai, O. Fiehn, S. Kanaya, K. Saito, *PLoS ONE* **2009**, *4*, e7490.

[14] a) S. Böcker, Z. Lipták, *In Proceedings of ACM symposium on Applied Computing (ACM SAC 2005))*, *2005*, 151; b) S. Bocker, Z. Liptak, *Algorithmica* **2007**, *48*, 413; c) K. Dührkop, M. Ludwig, M. Meusel, S. Böcker, *Proceedings of Workshop on Algorithms in Bioinformatics (WABI 2013)* **2013**, 45.

[15] H. Kubinyi, *Anal. Chim. Acta* **1991**, *247*, 107.

[16] A. L. Rockwood, P. Haimi, *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 415.

[17] S. Böcker, M. C. Letzel, Z. Liptak, A. Pervukhin, *Bioinformatics* **2009**, *25*, 218.

[18] J. Claesen, P. Dittwald, T. Burzykowski, D. Valkenborg, *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 753.

[19] D. Valkenborg, I. Mertens, F. Lemière, E. Witters, T. Burzykowski, *Mass spectrometry reviews* **2012**, *31*, 96.

[20] M. Loos, C. Gerber, F. Corona, J. Hollender, H. Singer, *Analytical chemistry* **2015**, *87*, 5738.

[21] a) F. Rasche, A. Svatos, R. K. Maddula, C. Böttcher, S. Böcker, *Anal. Chem.* **2010**, *83*, 1243; b) T. Pluskal, T. Uehara, M. Yanagida, *Analytical chemistry* **2012**, *84*, 4396.

[22] M. A. Stravs, E. L. Schymanski, H. P. Singer, J. Hollender, *Journal of mass spectrometry : JMS* **2013**, *48*, 89.

[23] S. Böcker, K. Duhrkop, *Journal of cheminformatics* **2016**, *8*, 5.

[24] K. Dührkop, F. Hufsky, S. Böcker, *Mass spectrometry (Tokyo, Japan)* **2014**, *3*, S0037.

[25] T. Kind, O. Fiehn, *Bmc Bioinform.* **2007**, *8*.

[26] Mallard W. G., Sparkman O.D., *U.S. Department of Commerce, National Institute of Standards and Technology Standard Reference Data Program: Gaithersburg, MD*, *2014*.

[27] L. Breiman, *Machine Learning* **2001**, *45*, 5.

[28] J. K. Senior, *American Journal of Mathematics* **1951**, *73*, 663.

[29] a) J. R. de Laeter, J. K. Böhlke, P. de Bièvre, H. Hidaka, H. S. Peiser, K. J. R. Rosman, P. D. P. Taylor, *Pure and Applied Chemistry* **2003**, *75*, 683; b) M. E. Wieser, *Pure and Applied Chemistry* **2006**, *78*, 2051; c) A. H. Wapstra, G. Audi, C. Thibault, *Nuclear Physics A* **2003**, *729*, 129.

[30] P. Dittwald, D. Valkenborg, J. Claesen, A. L. Rockwood, A. Gambin, *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1732.

[31] L. Breiman, *Machine Learning* **1996**, *24*, 123.

[32] T. Hastie, R. Tibshirani, J. Friedman, *Springer Series in Statistics; Springer-Verlag: New York*, *2009*.

[33] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams et al., *Nucleic acids research* **2013**, *41*, D456-63.

[34] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey et al., *Nucleic acids research* **2014**, *42*, D1083-90.

[35] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, *Nucleic acids research* **2006**, *34*, D668-72.

[36] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong et al., *Nucleic Acids Res.* **2013**, *41*, D801-D807.

[37] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, *Nucleic acids research* **2016**, *44*, D457-62.

[38] F. M. Afendi, T. Okada, M. Yamazaki, A. Hirai-Morita, Y. Nakamura, K. Nakamura, S. Ikeda, H. Takahashi, M. Altaf-Ul-Amin, L. K. Darusman et al., *Plant & cell physiology* **2012**, *53*, e1.

[39] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker et al., *Nucleic acids research* **2016**, *44*, D1202-13.

[40] J. Gu, Y. Gui, L. Chen, G. Yuan, H.-Z. Lu, X. Xu, *PLoS ONE* **2013**, *8*, e62839.

[41] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, R. G. Coleman, *Journal of chemical information and modeling* **2012**, *52*, 1757.

[42] C. Chen, A. Liaw, L. Breiman, *Technical Report* **2004**.

[43] M. Bekkar, T. A. Alitouche, *IJDKP* **2013**, *3*, 15.

[44] K. Dührkop, K. Scheubert, S. Böcker, *Metabolites* **2013**, *3*, 506.

[45] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12580.

[46] A. Ipsen, E. J. Want, T. M. D. Ebbels, *Analytical chemistry* **2010**, *82*, 7319.

Supporting Information

Predicting the Presence of Uncommon Elements in Unknown Biomolecules from Isotope Patterns

Marvin Meusel[a], Franziska Hufsky[a,c], Fabian Panter[b], Daniel Krug[b], Rolf

Müller[b] and Sebastian Böcker*[a]

Affiliations

[a] Chair for Bioinformatics, Friedrich Schiller University Jena, 07743 Jena, Germany

[b] Department of Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland, Helmholtz Centre for Infection Research and Pharmaceutical Biotechnology, Saarland University, 66123 Saarbrücken, Germany

[c] RNA Bioinformatics and High Throughput Analysis, Friedrich Schiller University Jena, 07743 Jena, Germany

The printed version of the Supporting Information does not contain data that cannot be visualized satisfactorily on paper. To access this data please refer to the enclosed storage medium or the on-line version of this research paper.

S 7.1   Isotope patterns

The isotope distribution of elements influences the isotope pattern of the molecule in both, mass differences and intensities of peaks. Small molecules containing only CHNOP have an intense monoisotopic peak; the +1 peak has much lower intensity, and intensity of subsequent peaks is further decreasing. The more "distinctive" isotope distributions of S, Cl, Br, B, and Se are reflected in the shape of the isotope pattern of a compound as follows:

- Sulfur-containing compounds have slightly higher +2 peaks compared to CHNOP compounds. With an increasing ratio of S atoms to CHNOP atoms, the +2 peak in the isotope pattern of the molecule has an intensity higher than the +1 peak. We will deliberately ignore that sulfur containing compounds may be identified by an additional +2 peak at the characteristic mass difference of 1.995796Da from the monoisotopic peak.[1]

- Chlorine has two stable isotopes with a mass difference of 1.99705Da.[1] Isotope patterns of molecules containing Cl have a +2 peak of higher intensity than CHNOP compounds. Often the intensity of the +2 peak is higher than the intensity of the +1 peak. For a large number of Cl atoms in a molecule, the isotopic peaks with even indices have higher intensity than their preceding peaks. The most intense peak may not be the monoisotopic peak.

- Bromine also has two stable isotopes with a mass difference of 1.997954Da.[1] The intensity of the second isotope is higher than for Cl. Thus, the effects described for Cl are even stronger for Br.

- Boron has two stable isotopes in nature, ^{10}B and ^{11}B. Different from the other elements, the abundance of boron isotopes varies between habitats. The presence of B produces an isotope pattern that resembles a bell curve. The effect gets stronger with increasing B atom number. The most intense peak is often the +1 peak.

- Selenium has six stable isotopes and a very characteristic "zigzag shaped" isotope pattern, which is also reflected in the molecules containing Se. Selenium-containing compounds are often not accessible to methods decomposing the monoisotopic peak.

---

[1] http://www.sisweb.com/referenc/source/exactmas.htm

S 7.2   Features

To learn and predict the presence of rare elements from isotope patterns, the isotope patterns need to be transformed to a set of numerical features characterizing the data. The number of features varies according to the number of peaks. The full list of mass and intensity features can be found in Supplementary Table S.1.

S 7.3   Compound databases

For ChEBI (Chemical Entities of Biological Interest) we downloaded the Chebi complete.sdf on 19/03/15 and extracted the molecular formulas from the Formulae field. ChEMBL version 19 was also downloaded as SDF file on 26/01/15. The SDF file was converted to InChI files using Open Babel 2.3.2 to extract the molecular formulas. DrugBank version 4.2 was downloaded as SDF on 19/03/15. Molecular formulas were extracted from the JCHEM FORMULA field. HMDB version 3.6 was downloaded on 19/03/15 as SDF. Molecular formulas were extracted from the CHEMICAL FORMULA field. Indofine[2] was downloaded on 20/03/2015 and converted to InChI files. From this database, we use Chalcones, Chromatrography-Standards, Chromones, Coumarins, Flavanones, Flavones, Flavonoids-Coumarins, Herbal Standards, Isoflavones, Life-Science, Natural-Products, Rare-Organics- Biochemicals and Specialty-Chemicals. For Kegg, we downloaded the mol file from KEGG MEDICUS on 20/01/15. KNApSAcK was downloaded on 02/01/15 using the KNApSAcK Keyword Search. Molmall[3] was downloaded on 14/03/15 using the XLS table. PubChem was downloaded between 23/03/15 and 25/03/15 as SDF files using the PubChem FTP. The PUBCHEM MOLECULAR FORMULA field was used for extracting the molecular formulas. UNPD was downloaded as SDF on 23/03/15 and the Molecular Formula field was used. ZINC version 12 was downloaded on 20/03/15. We use the Standard-filtered Lead Like, Fragment Like, Drug Like and Shards subsets. We downloaded the property files and extracted the molecular formulas by converting the SMILEs to INCHIs using Open Babel 2.3.2.

S 7.4   Learning phase

We train the classifiers on two Intel Sandy Bridge processors at 2.3 Ghz with six cores and 64 GB RAM. Training of all 18 classifiers required 3,5 h (see Supplementary Table S.3). We find that training requires

---

[2] http://indofinechemical.com/resources/sdf/sdf.aspx

[3] http://www.molmall.net/download.html

more time the less unique the isotope pattern of the element. This is also reflected in the size of the trees in the random forests (see Supplementary Table S.4).

S 7.5   Prediction quality of the classifiers

For evaluating the prediction quality of the classifiers, we compute Receiver Operating Characteristic (ROC) curves (see Figure 3 in the main document and Supplementary Figure S.1). To evaluate the tradeoff between sensitivity and specificity, we analyze the false positive rates (FPRs) for all 18 classifiers at fixed true positive rates (TPRs) (see Table 3 in the main document and Supplementary Table S.5). Voting thresholds for the random forests result from the best tradeoff between FPR and TPR for each classifier (see Supplementary Table S.6).
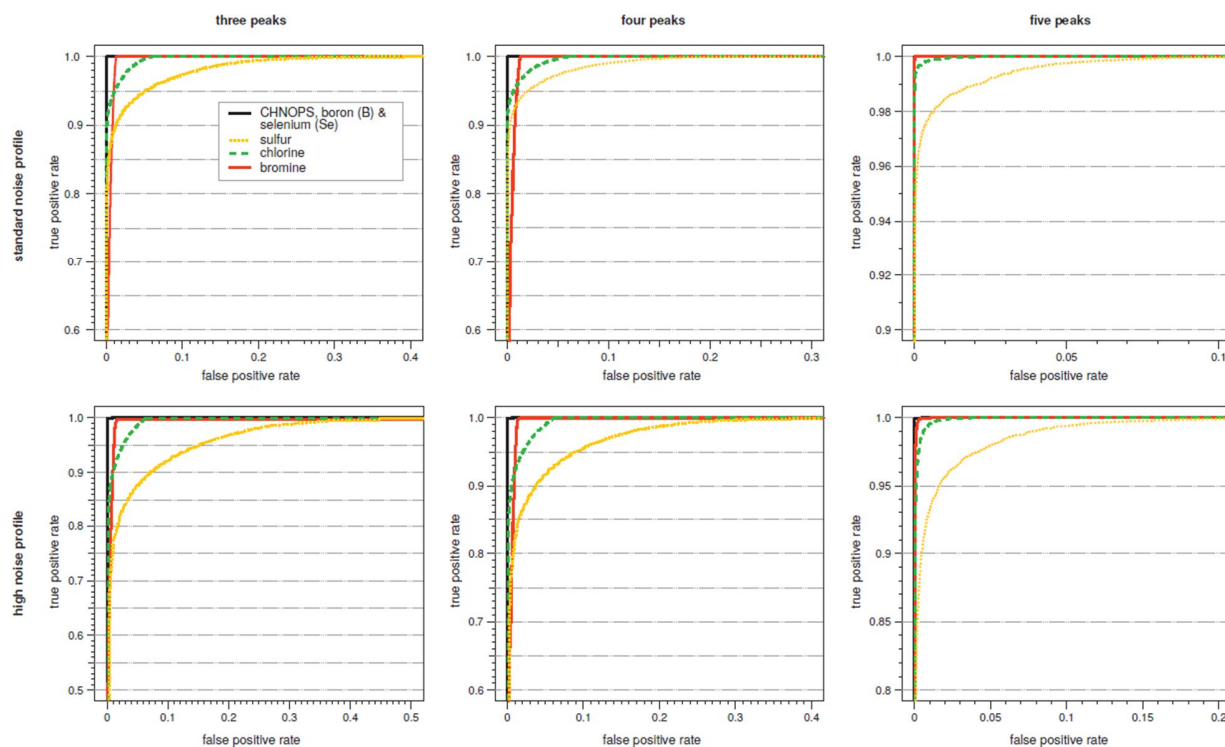
Figure S.1. *ROC curves for all classifiers using three peaks (left), four peaks (middle) and five peaks (right) at standard noise profile (top) and high noise profile (bottom). Note the different scales for both axes in the single plots.*
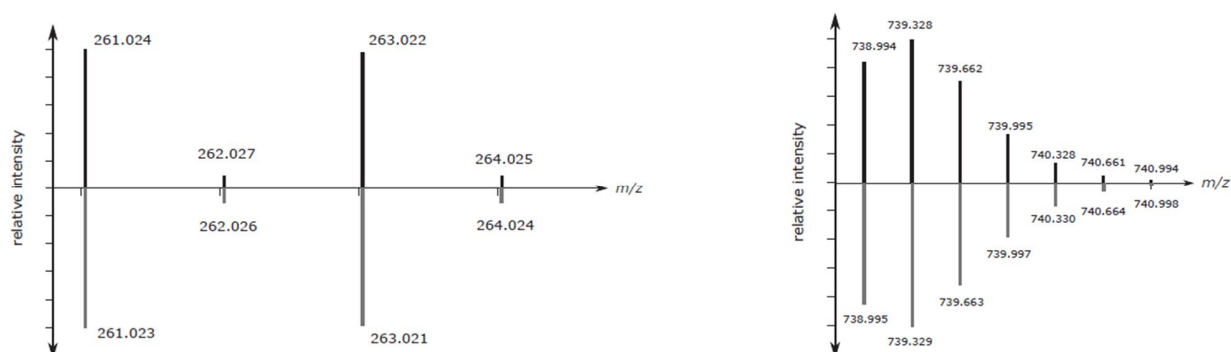
Figure S.2. *Two false positive examples for the chlorine classifier. Left: Bromazil with molecular formula $C_9H_{13}BrN_2O_2$ and exact mass 260.016 Da. Right: Pinnensin A with molecular formula $C_{96}H_{139}N_{27}O_{30}S_2$ and exact mass 2213.962 Da, triple charge. We observe no significant differences between the measured (top) and the theoretical (bottom) mass spectrum.*
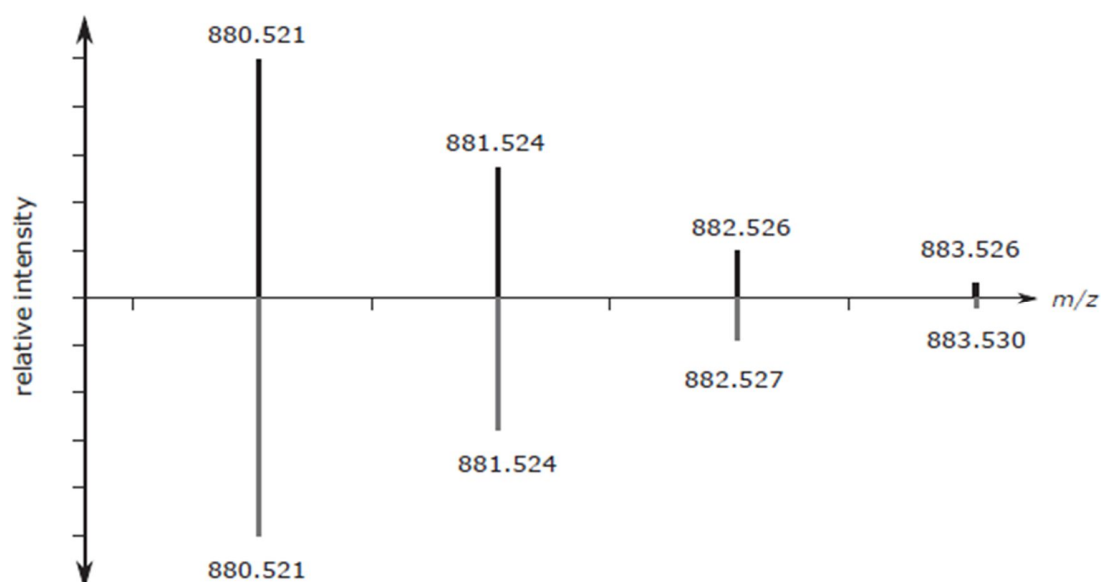
Figure S.3. *False positive example for the sulfur classifier, chivosazol A methyl ester with molecular formula $C_{50}H_{73}NO_{12}$ and theoretical mass 879.513 Da. We observe no significant differences between the measured (top) and the theoretical (bottom) mass spectrum.*

Table S.1. *Full list of mass and intensity features used for the training of the classifiers for isotope pattern with three peaks, four peaks or five peaks. p_x is the peak at position x, where x = 0 is the monoisotopic peak, x = 1 is the +1 peak and so on. int(p_x) is the intensity of peak p_x, mass(p_x) is the mass of peak p_x.*

| feature	three peaks	four peaks	five peaks
$\text{int}(p_0)$	x	x	x
$\text{int}(p_1)$	x	x	x
$\text{int}(p_2)$	x	x	x
$\text{int}(p_3)$		x	x
$\text{int}(p_4)$			x
minimal peak intensity	x	x	x
maximal peak intensity	x	x	x
median of peak intensities	x	x	x
$\text{int}(p_0) + \text{int}(p_2) + \text{int}(p_4)$	x	x	x
$\text{int}(p_1) + \text{int}(p_3)$		x	x
$\min(\text{int}(p_0),\text{int}(p_2),\text{int}(p_4))$	x	x	x
$\min(\text{int}(p_1),\text{int}(p_3))$		x	x
$\max(\text{int}(p_0),\text{int}(p_2),\text{int}(p_4))$	x	x	x
$\max(\text{int}(p_1),\text{int}(p_3))$		x	x
index of most intensive peak	x	x	x
index of second most intensive peak	x	x	x
index of third most intensive peak	x	x	x
$\text{int}(p_0) - \text{int}(p_1)$	x	x	x
$\text{int}(p_0) - \text{int}(p_2)$	x	x	x
$\text{int}(p_0) - \text{int}(p_3)$		x	x
$\text{int}(p_1) - \text{int}(p_2)$	x	x	x
$\text{int}(p_1) - \text{int}(p_3)$		x	x
$\text{int}(p_2) - \text{int}(p_3)$		x	x
$\text{int}(p_0) / \text{int}(p_1)$	x	x	x
$\text{int}(p_0) / \text{int}(p_2)$	x	x	x
$\text{int}(p_0) / \text{int}(p_3)$		x	x
$\text{int}(p_1) / \text{int}(p_2)$	x	x	x
$\text{int}(p_1) / \text{int}(p_3)$		x	x
$\text{int}(p_2) / \text{int}(p_3)$		x	x
$(\text{int}(p_0) / \text{int}(p_1)) - (\text{int}(p_1) / \text{int}(p_2))$	x	x	x
$(\text{int}(p_1) / \text{int}(p_2)) - (\text{int}(p_2) / \text{int}(p_3))$		x	x
$(\text{int}(p_0) / \text{int}(p_1)) / (\text{int}(p_1) / \text{int}(p_2))$	x	x	x
$(\text{int}(p_1) / \text{int}(p_2)) / (\text{int}(p_2) / \text{int}(p_3))$		x	x
$\text{int}(p_0) + \text{int}(p_1)$	x	x	x
$\text{int}(p_0) + \text{int}(p_1) + \text{int}(p_2)$	x	x	x
$\text{int}(p_0) + \text{int}(p_1) + \text{int}(p_2) + \text{int}(p_3)$		x	x
$\text{int}(p_1) + \text{int}(p_2) + \text{int}(p_3)$		x	x
$\text{int}(p_1) + \text{int}(p_2)$	x	x	x
$\text{int}(p_2) + \text{int}(p_3)$		x	x
$\text{mass}(p_0)$	x	x	x
$\text{mass}(p_1) - \text{mass}(p_0)$	x	x	x
$\text{mass}(p_2) - \text{mass}(p_0)$	x	x	x
$\text{mass}(p_3) - \text{mass}(p_0)$		x	x
$\text{mass}(p_4) - \text{mass}(p_0)$			x
$\text{mass}(p_2) - \text{mass}(p_1)$	x	x	x
$\text{mass}(p_3) - \text{mass}(p_1)$		x	x
$\text{mass}(p_4) - \text{mass}(p_1)$			x
$\text{mass}(p_3) - \text{mass}(p_2)$		x	x
$\text{mass}(p_4) - \text{mass}(p_2)$			x
$\text{mass}(p_4) - \text{mass}(p_3)$			x

Table S.2. Compound databases used for simulating isotope patterns for the training set and the evaluation set. Number of molecular formulas after filtering.

database	total #compounds	only CHNOPS	sulfur (S)	chlorine (Cl)	bromine (Br)	boron (B)	selenium (Se)
training set:							
PubChem	1 156 671	634 151	500 276	344 731	171 370	56 873	16 727
ChEMBL	187 840	109 919	75 968	57 613	23 470	1 268	845
ZINC	128 247	59 003	71 240	49 143	26 754	0	0
total*	1 128 059	604 506	502 529	345 799	171 596	56 808	16 680
evaluation set:							
ChEBI	17 284	15 442	3 279	1 473	349	22	47
DrugBank	4 714	3 927	1 195	581	163	47	12
HMDB	9 103	8 619	1 406	425	43	1	20
Indofine	1 571	1 201	199	221	135	52	0
KEGG MEDICUS	3 179	2 548	798	563	66	7	1
KNAPSAcK	13 584	12 515	1 143	657	485	5	10
molMall	7 290	5 309	1 685	1 297	706	12	22
UNPD	32 090	28 424	4 104	2 102	1 761	9	19
total*	51 097	42 713	9 892	5 502	3 015	141	98

*Note that the total numbers are after removing double occurrences of molecular formulas, removing molecular formulas from the measured datasets, and removing formulas in the evaluation set from the training set.

Table S.3. Running times for training and prediction. Running times for training were measured on a processor with twelve cores, whereas running times for prediction were measured on a single core. Prediction is measured including loading the classifier from hard disk, and running times are reported for the complete evaluation set (standard noise profile) with 51,097 molecular formulas. We report minimum running time over five independent runs.

classifier	training	prediction	load classifier
CHNOPS - three peaks	6 m 26 s	2.0 s	0.1 s
CHNOPS - four peaks	10 m 39 s	2.1 s	0.1 s
CHNOPS - five peaks	9 m 56 s	2.5 s	0.1 s
S - three peaks	8 m 25 s	8.4 s	4.5 s
S - four peaks	22 m 49 s	6.8 s	2.5 s
S - five peaks	24 m 08 s	5.2 s	1.5 s
Cl - three peaks	10 m 25 s	3.7 s	1.6 s
Cl - four peaks	20 m 37 s	3.7 s	1.3 s
Cl - five peaks	15 m 26 s	3.2 s	0.4 s
Br - three peaks	14 m 07 s	3.1 s	1.3 s
Br - four peaks	24 m 55 s	3.5 s	1.1 s
Br - five peaks	14 m 11 s	2.4 s	0.3 s
B - three peaks	4 m 35 s	1.5 s	0.01 s
B - four peaks	6 m 54 s	1.7 s	0.02 s
B - five peaks	8 m 09 s	2.1 s	0.02 s
Se - three peaks	4 m 19 s	1.5 s	0.02 s
Se - four peaks	6 m 52 s	1.8 s	0.02 s
Se - five peaks	7 m 58 s	1.8 s	0.02 s

Table S.4. *Tree depth and number of nodes for all 18 classifiers. Values are medians over all trees in the forest.*

	three peaks		four peaks		five peaks	
	tree depth	nodes	tree depth	nodes	tree depth	nodes
CHNOPS	14	161	12	113	12	105
S	46	12321	47	9431	39	5351
Cl	45	5515	42	4201	26	1105
Br	45	4259	43	3747	21	653
B	4	9	4	11	6	19
Se	8	19	6	11	4	9

Table S.5. *How many false positive prediction do we have to accept, in order to reach a certain sensitivity? We report false positive rates (FPR) for all 18 classifiers at fixed true positive rates (TPR). TPR 0.998 corresponds to two missed positive prediction in 1000 examples; TPR 1 means missing not a single positive example. Results for the high noise profile. 'N/A', classifier cannot reach the desired TPR.*

	three peaks				four peaks				five peaks			
desired TPR	0.997	0.998	0.999	1	0.997	0.998	0.999	1	0.997	0.998	0.999	1
CHNOPS	0.0	0.0	0.001	N/A	0.0	0.0	0.0	N/A	0.0	0.0	0.0	N/A
S	0.447	0.447	N/A	N/A	0.315	0.372	0.439	N/A	0.131	0.156	0.185	N/A
Cl	0.061	0.062	0.064	N/A	0.059	0.060	0.061	N/A	0.013	0.016	0.022	0.043
Br	0.018	N/A	N/A	N/A	0.014	0.015	0.015	N/A	0.002	0.003	0.004	0.012
B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Se	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

(row label: resulting FPR)

Table S.6. *Resulting voting thresholds for all 18 classifiers. For the respective false positive and true positive rates, see Table 3 in the main document.*

# peaks:	three	four	five
CHNOPS	0.5	0.5	0.5
S	0.06	0.14	0.19
Cl	0.14	0.16	0.20
Br	0.3	0.3	0.3
B	0.3	0.3	0.3
Se	0.3	0.3	0.3

8 Discussion and Outlook

8.1 Challenges to access bacterial secondary metabolite diversity

Natural products continue to represent a viable source for novel chemical entities exhibiting various biological activities.[1] In former times, most bioactive natural products have been found by bioactivity guided isolation approaches.[2] However, as the 'low hanging fruits' of bacterial secondary metabolism such as strongly active secondary metabolites produced in high amounts have mostly been found (a notion that appears especially evident for producer groups such as the actinomycetes, being screened already over decades), new approaches increasingly focus on accessing the 'hidden' bacterial secondary metabolome. 'Hidden' metabolites in this case comprise compounds whose production levels are too low to detect the corresponding activity in primary assays, or the – oftentimes rather limited – panel of bioactivity assays was unable to detect their bioactivity because the matching assay was simply not available. Furthermore, secondary metabolites that evade detection by the instrumental analytics set-up, either due to their low abundance or as the applied analysis system is suboptimal for detection of the specific secondary metabolite, are oftentimes also referred to as parts of the 'hidden' metabolome. Conclusive estimates about the proportion of metabolome constituents that could be regarded as 'hidden' are generally unavailable, but some basic insights can be derived from genome analysis of secondary metabolite-producing bacteria. For myxobacteria, an early notion of the fraction of secondary metabolites belonging to the 'hidden' metabolome came from an analysis of *Myxococcus xanthus* DK1622, where the majority of biosynthetic gene clusters could not be correlated to any known secondary metabolite, although transcriptomic and proteomic analysis suggested that most BGCs were expressed under the conditions tested.[3,4] This is a finding that is reflected among other secondary metabolite producing bacteria as well although under debate.[5] The number of BGCs in bacteria altogether clearly exceeds the number of known secondary metabolites to date.[3] Another myxobacterial study that focused on *Chondromyces crocatus* Cmc5 – a relatively well investigated myxobacterial strain – reflects this finding: although 5 compound families from this strain were already known before it was genome-sequenced, additional compounds were readily discovered and connected to their BGCs by genome-mining after its whole genome sequence became available.[6]
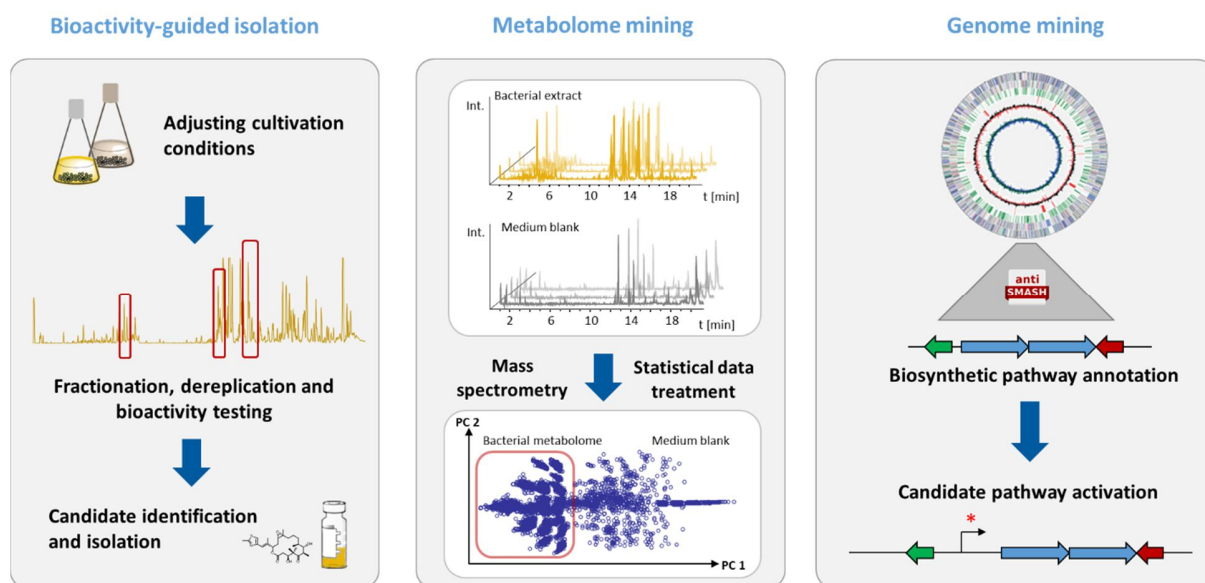
Figure 1. *The three pillars of contemporary natural products discovery, representing the fundamentally different approaches for secondary metabolite identification.*

It is thus of vital importance to further develop innovative ways of secondary metabolite searching and detection, which includes the application of improved analytical setups as well as pursuing genomics-driven approaches, both of which are exemplified in this work. Improved access to myxobacterial secondary metabolites by lowering detection limits and streamlined mass spectrometry data processing to filter the complex data according to discovery-rationales will improve chances to 'harvest' from the vast chemical space of myxobacterial secondary metabolites. Access to this structural diversity in turn is key to finding novel biologically active molecules with innovative targets, instead of rediscovering the 'low hanging fruits' of the secondary metabolome that come with the risk of previous description. Throughout the studies presented here, the use of recently emerging techniques has been highlighted to uncover novel secondary metabolites of myxobacterial origin, reaching beyond the 'classical' activity guided secondary metabolite identification, isolation and structure elucidation. Careful evaluation of novel developments in instrumental analytics, statistical data treatment of MS and MS^n data, DNA sequencing techniques as well as *in silico* analysis of genomic big data function together in these studies to reveal both prospects and limitations of each technique in terms of their utility for the identification of novel secondary metabolite scaffolds.[7]

8.1.1 Prospects of novel analysis techniques for secondary metabolite detection

Depending on used extraction and analysis methods, different secondary metabolites may be easily detectable or readily fall below their individual limit of detection. In recent years, most metabolomics

labs have switched to exclusively apply UPLC-ESI-MS hyphenated analytics – a versatile method that is able to detect a wide variety of secondary metabolites – albeit disregarding all molecular features detectable using other ionization techniques like APCI, or such solely visible in direct infusion MS. While solely relying on UPLC-ESI-MS analytics has been very successful in the past, it appears reasonable to assume that a significant proportion of the so-called 'hidden' secondary metabolome is probably unseen, because corresponding molecules are in the blind spot of UPLC-ESI-MS. Application of state of the art UPLC-ESI-MS analytics should be therefore in the future amended with additional techniques to venture into the unknown chemical space in bacterial secondary metabolomics that could be – according to genome analyses – even larger than the already known chemical space.[3] Some of these techniques, used in this work to a varying degree – will be discussed in the following.

8.1.2 Improvements in UPLC-MS based instrumental analytics

Since the development of ultra-performance liquid chromatography (UPLC) that operates at LC pressures above 400 bars on sub 3 µm particles, liquid chromatography separation power received a tremendous boost. Combined with electrospray ionization (ESI) and a fast scanning high resolution mass spectrometer (HRMS, mostly a QqToF device in this work) as a second dimension, the detection of thousands of molecular features in a short UPLC-MS analysis run is feasible.[8] The accuracy of the high resolution mass spectrometers additionally allows for prediction of plausible molecular formulas for every detected feature, thus unlocking additional information early during the discovery process.[9] However, chemical screening efforts in bacterial secondary metabolomics are still complicated by matrix effects, as it is impossible to discern a priori molecular features that are constituents of the bacterial secondary metabolome from molecular features originating from the medium ingredients (see section 8.1.6). Moreover, the abundant matrix of biological samples can lead to severe ion suppression during the ESI process.[10,11] Among the main agents causing ion suppression effects in LC-ESI-MS are salts, ion paring agents e.g. phospholipids, primary metabolites such as amino acids as well as residual proteins, all of which occur in bacterial extracts.[10] Furthermore, bacterial secondary metabolites occupy an extremely wide range of polarities. Nucleoside antibiotics such as pseudouridimycin or aminoglycosides such as kanamycin are problematic to detect by UHPLC-ESI-MS, as they elute very close to void volume leading to their MS signal being suppressed by co-eluting salt.[12] Very non-polar components of the bacterial secondary metabolome such as slightly oxygenated polyketide or terpene scaffolds might be too hydrophobic for elution from a RP-HPLC column, thus evading detection in UPLC-MS. Furthermore, most standard detection and quantification protocols for terpenes such as steroids rely on atmospheric pressure chemical ionization (APCI)-MS, as they are hardly accessible to protonation in liquids, making

them invisible in ESI.[13] While ESI based UPLC-HRMS is without question currently the most versatile technique to detect bacterial secondary metabolites and therefore widely employed by most natural products analysis labs, the aforementioned limitations should serve as a word of caution that exclusive usage of this technique is likely to miss secondary metabolite scaffolds that simply evade detection using this technique.

8.1.3   Emergence of DI-FTICR based secondary metabolomics

Recently, Fourier transform ion cyclotron resonance spectrometers have become more accessible to natural products research laboratories. These spectrometers exceed the resolution of QqToF or Orbitrap spectrometers, reaching up to 1000000 in resolving power whilst having at the same time very low limits of detection.[14] Still, contrary to ToF or Orbitrap spectrometers, the FTICR needs significantly longer acquisition times to reach these resolving power values, as mass accuracy rises with the residence time of ions in the acquisition cell. Thus, LC-FTICR applications are bound to create coverage problems as long cycle times would mean high resolving power but also miss some molecular features. If one reduces the cycle time to lower values to fix this problem, resolving power would drop to values one can reach with Orbitrap or ToF systems, removing the resolving power advantage of the FTICR over said systems. Therefore, FTICR systems are currently mainly used in direct infusion (DI) mode. Due to the impressively low limit of detection, biological samples can be measured as highly diluted samples to limit ion suppression effects. Thereby, a full scan of a biological sample with ultra-high resolution can be measured in less than 4 minutes giving this technique a competitive edge to LC-HRMS applications that often take around 20 minutes per sample. Furthermore, resolving power of the FTICR allows dissection of the isotopic fine structure allowing much higher precision in determination of molecular formulas than using ToF or Orbitrap machines.[15] Still, DI-FTMS measurements have significant drawbacks compared to LC-HRMS analysis. First of all, because chromatography as an orthogonal analysis dimension is missing, it is impossible to characterize compounds based on chromatographic retention time that provides insights into compound polarity and provides a second comparison parameter in addition to the high resolution mass. Furthermore, as natural products occupy only a limited chemical space it is likely to encounter isobaric compounds in a sample. By design, DI-FTICR would be unable to distinguish these compounds while LC based application can discern them via retention time differences. Last but not least, DI-FTMS applications suffer to a higher extend from ion intensity modulation by suppression and enhancement meaning that absolute feature intensity in DI-FTICR experiments does not necessarily relate to the amounts of the corresponding compound in the biological sample.[11] One can thus conclude that some of the features seen in these experiments account for compound amounts that

are insufficient for structure elucidation even if large scale fermentations are applied. Published results, as for example a comprehensive study on ginseng metabolites, as well as preliminary in house data (Bader, Haack, Panter; unpublished results) suggest DI-FTICR to provide insights into the bacterial secondary metabolome that are complementary to widely used LC-HRMS applications, indicating this technique to be a valid asset to tap the biosynthetic dark matter of bacterial secondary metabolism in the future.[16]

8.1.4 Mass spectrometry imaging (MSI) in bacterial secondary metabolomics

One core premise in secondary metabolism is that secondary metabolites are often produced under specific conditions only, e.g. when certain cultivation parameters are met, or at distinct time points during the bacterial growth curve. In order to study secondary metabolism under very specialized conditions such as in bacterial interaction zones on an agar plate or during fruiting body formation, spatial resolution of mass spectrometry data is necessary.[17] Mass spectrometry imaging realizes the acquisition of mass spectrometric data in a spatially resolved manner in order to identify compounds associated with a topological range. The main ionization techniques applied in MSI of biological samples are matrix associated laser desorption/ionization (MALDI) and desorption electrospray ionization (DESI). Both techniques rely on spatially resolved desorption of analytes from a surface, in case of MALDI using a laser beam, in case of DESI using an electrospray source.[18] The big advantage of this technique for investigation of bacterial secondary metabolites is spatial resolution of the results. As demonstrated in this work (chapter 6), MALDI MSI has been successfully used to uncover the homospermidine lipids, compounds that exclusively form during myxobacterial fruiting body formation, by differential analysis of vegetative myxobacterial cells compared to the fruiting body.[17] In principle, this technique could be used to monitor myxobacterial predation and one might be able to find compounds belonging to the bacterial chemical warfare arsenal, which might become candidates as novel antibiotic drug leads. On the other hand, as matrix embedded biological samples are very densely concentrated, both MALDI and DESI experiments suffer from strong ion suppression effects. It is thus close to impossible to directly detect weakly ionizable compounds directly from the solid surface, which severely limits MALDI and DESI techniques to easily ionizable secondary metabolites.

8.1.5 Chromatography applications using supercritical CO_2 as mobile phase

A technique that has long existed, but has only recently been developed to become a reliable chromatography technique is supercritical fluid chromatography or SFC. While being underused so far in bacterial secondary metabolomics, the technique has proven its applicability and usefulness in

phytochemistry.[19] SFC applications use most commonly supercritical CO_2 as a mobile phase to which a gradient of an organic co solvent is added during separation. The fluid used for separations combines desirable qualities of gas chromatography such as the high diffusivity with desirable qualities of a liquid, i.e. high solubility. Polarity wise, SFC applications are more reminiscent of normal phase applications as polarity of supercritical CO_2 is approximately the polarity of hexane, which also explains why mainly normal phase column material is used in these kinds of separations. The method has proven to be orthogonal to liquid chromatography allowing the separation of many mixtures that are inseparable using standard UPLC.[19,20] Furthermore, with the development and improvement of SFC-MS hyphenation the technique develops more and more towards an additional analytical asset for screening approaches of biological samples. The real strength of SFC in natural products research turns out to be its applicability in domains where UPLC-MS is notoriously difficult to use due to the non-polarity and insolubility of the analytes, such as for terpenes, flavonoids and in fatty acid analysis.[21] Some of these non-polar analytes can be analyzed by GC-MS, but this requires derivatization and thermostability, which is often a major issue with molecules of biological origin. Additionally, due to the high diffusivity of analytes in supercritical CO_2, SFC separations are faster than their UPLC counterparts.[20,22] Although SFC is mainly used in phytochemical applications, the preparation of Myxothiazol Z on SFC, as it has been done in this work, demonstrates that application utility of SFC for natural products purification. Still, SFC applications have drawbacks as for example the incompatibility with water which is immiscible with supercritical CO_2. Thus polar compounds, especially larger hydrophilic peptides are mainly insoluble in SFC solvents meaning that bacterial secondary metabolomics with regard to polar non-ribosomal peptides cannot be transferred to SFC. This highlights once again the technique to be a complementary one in cases that are difficult to treat by reversed-phase LC, but it will likely not replace conventional UPLC.[7]

8.1.6   Statistical means for big data processing

High end UPLC-HRToF as well as DI-FTMS equipment nowadays easily records more than 10,000 molecular features in one single bacterial extracts, comprising both medium borne matrix features as well as features representing constituents of the bacterial secondary metabolome.[8,16] It is thus of utmost importance to automate workflows separating the 'unimportant' matrix features from the bacterial metabolome. A crucial point in this analysis is the dissection of LC-MS or DI-MS data into molecular features. With the improvement of feature finding and adduct detection algorithms from earlier simplistic forms to the contemporary algorithms for DI-MS and LC-MS data, the number of detectable and discernable molecular features rises significantly.[23] Combined with well-controlled

chromatography conditions that lead to reproducible retention times, buckets can be confined more precisely, leading to increasing numbers of features that can be used as input for statistical tools like principle component analysis, as exemplified in the fulvuthiacene study (chapter 3). Therefore, in line with increases in mass spectrometric performance, the data processing capacity needs to evolve as well. Recent developments in the timeframe of this thesis indicates continuous improvement in this respect, since the vendor-specific analysis framework ("Metaboscape", Bruker Daltonics, Bremen) became significantly better in feature finding compared to previous software products. As most metabolomics experiments rely on 'big data' in some form that cannot be treated manually, it is of utmost importance that software development keeps up with developments in mass spectrometric instrumentation in order to fully exploit the potential of state of the art mass spectrometry. The ever increasing amounts of data created in state of the art metabolomics approaches urgently need intelligent post processing and data mining approaches to extract a limited amount of valuable data that can be treated manually out of the 'big data' that is created in these approaches. To achieve this, other considerations have to be taken into account, like for example to look at the bacterial secondary metabolome in a taxonomy resolved manner.[24] Another approach is the extraction of a maximum of structural information at an early time point in the analysis as it was exemplified by the discovery of the fulvuthiacenes in chapter 3 of this work.

8.1.7 Developments in extraction of structural information from MS^n data

A key limiting factor in natural products science remains full structure elucidation of molecules as on the one hand, bacterial secondary metabolites tend to be complex often exceeding 600 Daltons, while on the other hand production and separation is often not trivial. This finding is well exemplified for several molecules purified as part of this thesis.[3,7,9,25] Thus, especially since secondary metabolite detection performance received a large boost in recent times, mainly due to advances in mass spectrometry instrumentation, the most costly and time consuming steps in secondary metabolite research are still the isolation and NMR based structure elucidation efforts. It is therefore of vital importance to make well-informed decisions about candidate compounds for isolation - considering presumed structural novelty - in order to avoid isolation of 'less interesting' molecules. While NMR data used for full structure elucidation require rather large amounts of pure substances, mass spectrometry data can be acquired much faster and directly from crude mixtures. Recent advances in computational mass spectrometry such as fragmentation trees and molecular networking boosted the utility of MS^n fragment spectra.[26,27] In this work, both the assignment of the fulvuthiacenes' as well as the chloromyxamides' minor derivatives relied on extracting early structural information out of mass

spectrometry data.[28] With the above-mentioned enhancements of the detection stage fueled by both advancements in analytical systems as well as advancements in 'big data' treatment software, the application of pre-filters to judge, which MS signals represent compounds worth focusing on becomes imperative. An element to support these efforts is realized by community curated projects such as GNPS as large scale data repositories for MS² data in order to be able to extract as much CID MS² based information as possible from an open source platform.[27] Spectral comparison, whether it is done by spectral networking or fragmentation tree computation can thus help place an uncommon MS² spectrum in context using publicly available CID fragmentation information.[26,27] While such database-powered identification and association tools have long become standard in Protein MS, e.g. the MASCOT database and GC-MS using NIST database identification for decades, LC-ESI-CID based MS² experiments were always presented with additional challenges.[29,30] GC-MS applications capitalize on the fact that they almost exclusively use electron impact ionization which is very reliable and reproducible across a wide variety of instruments. Protein MS on the other hand is simplified by peptide bond fragmentation that follows established rules, which makes protein fragmentation patterns predictable and comparable to theoretical fragments.[30] Small molecule ESI-CID-MS always suffered from limited comparability of MS² spectra across different MS platforms, which was the main obstacle of developing these community projects so far. Advances in software solutions to tackle this problem as well as improving open access data repositories to store natural product MS² data will be key to make better decisions on structural novelty of a natural product in the future. This can guide isolation and structure elucidation efforts towards unexplored subspaces in the chemical universe of natural products diversity, which translates into chances to discover the much-desired new targets and biological activities.

8.1.8   Improvements in sequencing technology and biosynthetic pathway prediction

Since the development of fast and cheap genome sequencing techniques, bacterial secondary metabolism is increasingly tackled from an *in silico* point of view. DNA sequencing techniques such as illumina sequencing and the development of PacBio (Pacific Biosciences) sequencing enabled elucidation of bacterial genomes starting from minimal amounts of DNA.[31] In recent years prices per sequencing run dropped significantly leading to an explosion of genome information available to the public domain. Furthermore, automated predictions of coding sequences combined with easy access to computing power made high throughput genome annotations possible.[32] Community curated gene repositories such as GenBank that provide an ever increasing source of annotated genes improve comparative annotations as more reference genes of known function exist that the various algorithms can compare to.[33,34] This is particularly helpful as proteins of unknown function in a biosynthetic gene cluster that can

be linked to an enzymatic activity – even though this enzymatic activity might have been found in a completely different organism and field of interest – can help clarify unprecedented biosynthetic steps in a secondary metabolite biosynthesis. The predictive power of comparative tools is dependent on a database of assigned genes which will increase with every gene with an assigned function. In combination with one of the main advances in the field of genomics-driven natural products discovery, the automated procedures to predict and annotate secondary metabolite biosynthetic gene clusters, nowadays routinely available through the online software platforms antiSMASH and Prism, one can not only predict the core biosynthetic machinery but also explain tailoring enzyme steps.[35,36,37] These tools allow users to assess *in silico* the biosynthetic potential of large numbers of secondary metabolite producing organisms in a very short time, streamlining both the search for biosynthetic gene clusters corresponding to known secondary metabolites as well as getting a glimpse on the predicted biosynthetic potential of an organism. To efficiently prioritize bacterial – in our case myxobacterial – strains, this is of vital importance, as the biosynthetic potential varies from strain to strain. Analyses across genomes allow natural products scientists to focus on strains showing high numbers of architecturally interesting biosynthetic gene clusters, disregarding those strains with a limited number yet unassigned of BGCs. Additionally, self-resistance predictor tools such as ARTS come with the promise to automate the finding of self-resistance determinants in biosynthetic gene clusters that provide additional guidance for gene cluster prioritization.[38] In this work, the presence of a topoisomerase inhibitor self-resistance protein – a so-called pentapeptide repeat protein – in a cryptic gene cluster has been explicitly shown to confer self-resistance to the corresponding secondary metabolite, which in the end turned out to be a strong topoisomerase IV poison.[39] Further advances in secondary metabolite biosynthetic gene cluster community curation have been made by the MiBiG repository that allows storage and comparison of biosynthetic gene cluster information as well as easy access free of charge.[40,41] Like in the case of mass spectrometry data, community sharing and data curation of biosynthetic gene cluster data (sometimes termed the 'clustome') will be a key aspect to further secondary metabolite research as it provides scientists with tools to judge biosynthetic potential of an organism prior to compound isolation, as well as enabling the better understanding of complex biocatalysis involved in secondary metabolite biosynthesis.

8.1.9   Conclusions for bridging the genome to metabolome gap

As laid out in this work there are several important aspects to improving the prospects of finding novel biological entities from natural sources. Natural products research is a interdisciplinary field where high performance chemical analytics meets experimental biological and bioinformatics-driven

evaluation, e.g. by varying cultivation conditions, genetics experiments such as promotor exchanges and *in silico* analysis.[17,35,36,39] As discussed before, analysis of high resolution mass spectrometric 'big data' as well as *in silico* secondary metabolite gene cluster mining increasingly rely on large community curated databases. Examples of such are GenBank or MiBiG for *in silico* genetic analysis as well as GNPS, MetFrag, SIRIUS and – in this work – the in-house Myxobase project concerning treatment of mass spectrometric data.[27,33,40,42] Still, with all these bioinformatics tools available, the genome to metabolome link is not a trivial endeavor. While there are clear translation rules to transform genes into proteins – which are readily exploited in MS and MS^n based proteomics – the biosynthetic gene cluster to secondary metabolite connection generally lacks concise *in silico* prediction methods.[30] Some advances in this field were made for non-ribosomal peptides, a compound class that often abides by the collinearity rule during biosynthesis and shows rather 'predictable' amino acid fragments in multidimensional mass spectrometry analysis, as exemplified by the software Pep2Path.[43] Still, these methods rely on biosynthetic collinearity as well as somewhat 'regular' MS fragmentation, thus limiting their utility for many other chemical classes of natural products. Progress in artificial intelligence based machine learning might become helpful for narrowing the gap between the bacterial metabolome and its genome or 'clustome' in the future, but automated annotation of secondary metabolites to the respective clusters is neither generally feasible nor is it on the horizon. Currently, it seems even likely that secondary metabolite structure prediction will never be fully automatable as it appears impossible to predict enzyme reactions that trigger complex reaction cascades for, example. In the foreseeable future, as exemplified also in this work, correctly linking secondary metabolites to their respective gene clusters and understanding their biosynthesis will continue to comprise a lot of manual and experimental labor, although the mentioned software tools do indeed facilitate the rate and depth of insights gained.

8.2   Biochemistry lessons from myxobacteria

In addition to the instrumental analytics and data mining part of this work, a large part of this thesis was dedicated to investigating biosynthesis. Along the lines of this work, insights into three different, very peculiar structural elements of myxobacterial natural products as well as their underlying biosynthesis were obtained.

8.2.1   A dual function protein for heterocycle formation using amino acids

During pyxidicycline formation (chapter 2), the enzyme PcyJ is able to catalyze a cascade of enzymatic reactions that amend the polyketide precursor originating from PKS type II biosynthesis with a nitrogen containing heteroaromatic system via incorporation of an amino acid.[39]
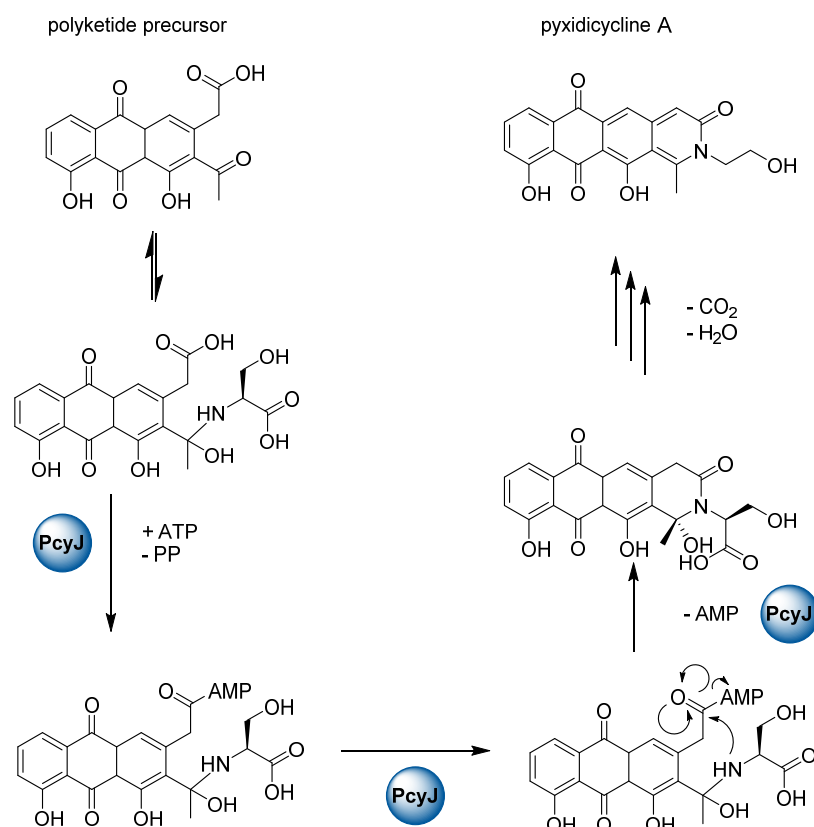
Scheme 1. *Enzymatic reaction cascade leading to the formation of the heterocycle in pyxidicycline biosynthesis*

The reaction starts with the polyketide precursor from the type II PKS assembly line depicted in scheme 1. This polyketide precursor is in an equilibrium with amino acid hemiaminal structures (Scheme 1). If a hemiaminal is formed with serine, this adduct should fit into the binding pocket of PcyJ, a protein that by design of its binding pocket probably already encourages the formation of exactly this hemiaminal structure. Once this moiety is formed, ATP is used to activate the carboxylic acid present on the precursor to form an activated carboxyl-AMP ester that subsequently undergoes intramolecular cyclisation with the hemiaminal nitrogen. This reaction is clearly enzyme catalyzed as it relies on the presence of PcyJ and no pyxidicyclines can be detected in PcyJ deletion mutants, while the polyketide precursor accumulates.[39] Furthermore, and contrarily to the case of jadomycin, PcyJ uniquely incorporates serine into the pyxidicycline scaffold.[44] Therefore, pyxidicycline biosynthesis reveals a new

method to synthesize nitrogen containing six membered heteroaromatic systems, which might be relevant in other type II PKS biosynthesis pathways that will be characterized in the future.

8.2.2 Biosynthesis of β-methoxymethacrylates in myxobacteria

There is a wide variety of compounds that are currently used as antifungal agents having a β-methoxymethacrylate moiety as part of their pharmacophore.[45] Biosynthesis of this peculiar unit has been of special interest in order to understand enzymatic processes behind biogenesis of this moiety to finally attempt to create derivatives usable as fungicides. One peculiarity of this biosynthetic transformation is that it is partly formed on the assembly line and partly in post assembly line tailoring steps.[46] The fulvuthiacenes that have been treated extensively in this work also contain a β-methoxymethacrylate moiety which has been subject of intense study (chapter 3). Already on the assembly line in fulvuthiacene biosynthesis we observe peculiar biosynthesis steps, for example the direct methylation of a keto-function by an O-methyl transferase (OMT).
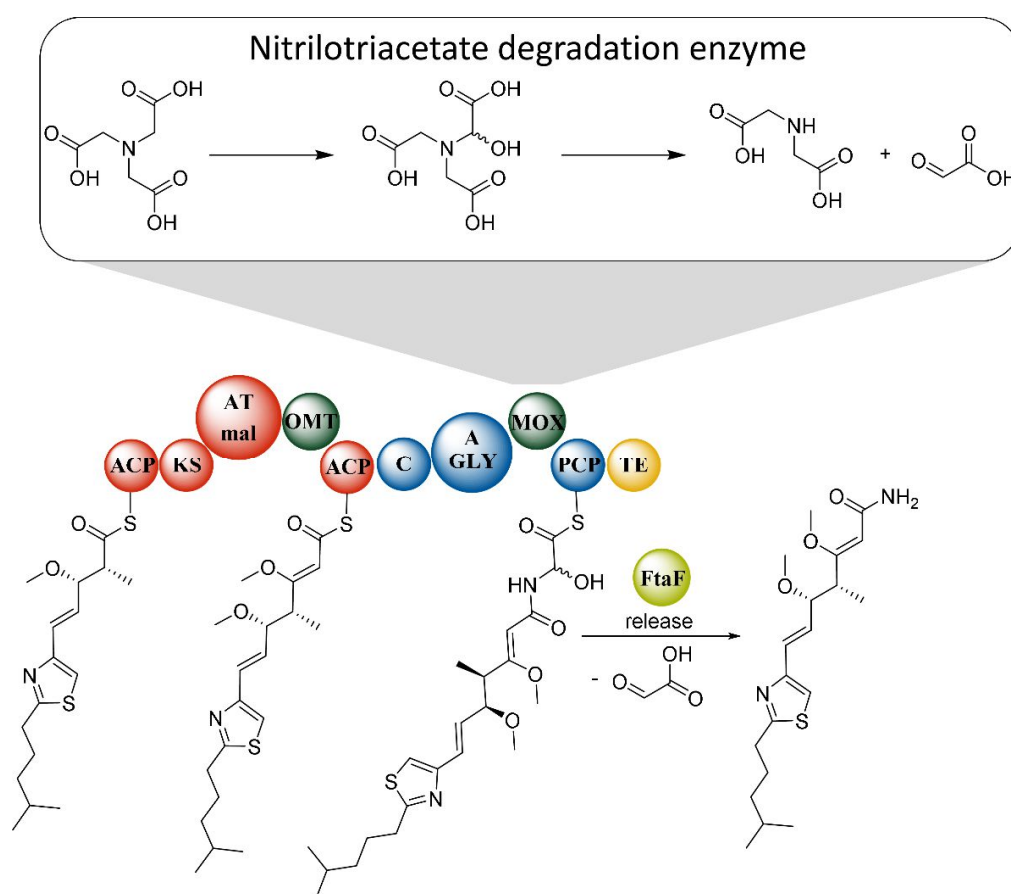

Figure 2. Last part of the fulvuthiacene assembly line including direct methylation of a keto-function and glycine degradation as known from nitriloacetate degrading enzymes

This reaction creates a methyl vinyl ether – a rather uncommon structural feature in polyketide biosynthesis – that will later be part of the terminal β-methoxymethacrylate. The terminal module in fulvuthiacene biosynthesis catalyzes an even more uncommon step in fulvuthiacene biosynthesis that is addition and break down of a glycine residue to the terminal carboxylic acid of the fulvuthiacene precursor. Adenylation and condensation modules that attach glycine to the fulvuthiacene precursor are not particularly different from most glycine specific domains, but there is an unusual monooxygenase (MOX) inserted into this module. Blast comparison of the sequence of this monooxygenase reveals high similarity to nitriloacetate degradation enzymes that catalyze the α-oxidation and breakdown reaction depicted in Figure 2. These enzymes are an excellent example of nature reutilizing similar enzymology to fulfill completely different tasks, which highlights again the necessity to characterize as many biochemical reactions as possible as they might become important in completely different contexts. After release combined with glycine breakdown with the assembly line releases a fulvuthiacene acrylamide structure which is then converted into fulvuthiacene using two post assembly line tailoring steps, catalyzed by an amidase called FtaH and another O-methyltransferase called FtaG.
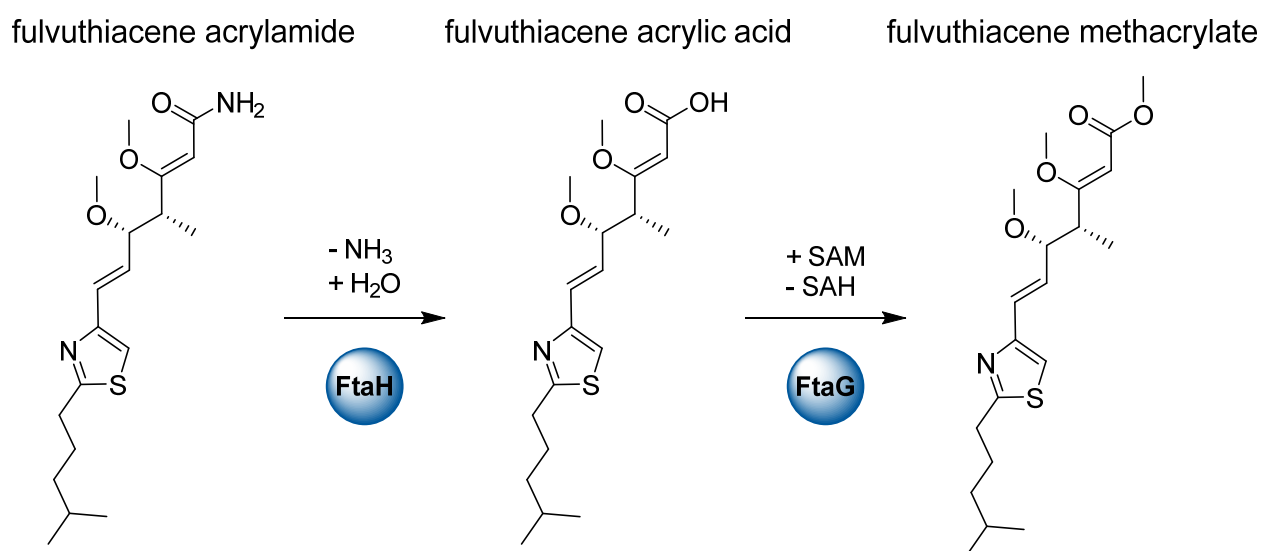

Scheme 2. *Post assembly line modification steps in the fulvuthiacenene biosynthesis*

The fulvuthiacene cluster contains the genes FtaH and FtaG which encode an amidase and a methyl transferase respectively. While the corresponding reactions are readily expected from these enzymes, as was discussed in the fulvuthiacene study (chapter 2), this reaction sequence shows biosynthesis to sometimes undertake seemingly 'unnecessary' detours. If one would remove the precursor from the penultimate module of the assembly line as depicted in Figure 2, one would directly obtain fulvuthiacene

acrylic acid without having to add and completely degrade a glycine molecule. This finding suggests that evolution of secondary metabolite biosynthesis pathways is to some degree genetic trial and error; in a wider sense Nature, although efficient at optimizing biosynthesis through evolution, is apparently not error proof in its quest to find the easiest pathway to a desired product. More importantly, the discovery of this biosynthetic 'detour' can be seen as direct evidence that in studying living organisms we are not looking at 'completed evolution', which by design does not exist, but we are rather looking at a snapshot of evolution in progress.

8.2.3   Biogenesis of CMPA as a new amino acid in NRPS biosynthesis

As mentioned in this work, the most peculiar part of chloromyxamide biosynthesis is the biosynthesis of a new natural amino acid building block called 6-chloromethyl-5-methoxy-pipecolic acid or CMPA. Feeding experiments followed by targeted tandem MS experiments have proven the corresponding moiety not to originate from lysine - like pipecolic acid in tubulysin biosynthesis - but to be derived from glutamate.[47] A key element in CMPA biosynthesis is its carrier protein, a LysW analog that, like in lysine biosynthesis, performs condensations at the γ-Glu position of the terminal Glu residue in CmxD.[48] The big difference between CmxD and LysW is, that while LysW accepts α-aminoadipate as a substrate for lysine biosynthesis, CmxD utilizes a glutamate as a starting point for CMPA biosynthesis.[28,48] The next peculiarity in CMPA biosynthesis is the final chlorination induced cyclization step. Here, although the present study could not obtain ultimate proof for the reaction sequencing on the molecular level, several hypothetical alternative deviations from the presented scheme can be excluded from a chemical reactivity point of view. In scheme 3 the reduction of the δ-carbonyl function is presented before addition of the generated chloronium ion to the terminal double bond. This is highly likely as on the one hand, the α,β-insaturated carbonyl function is probably too electron depleted to allow electrophilic attack of the cloronium and on the other hand, as seen in Scheme 4, the δ-carbonyl function would direct the intramolecular cyclisation reaction toward seven membered ring formation instead of formation of the pipecolic acid ring.
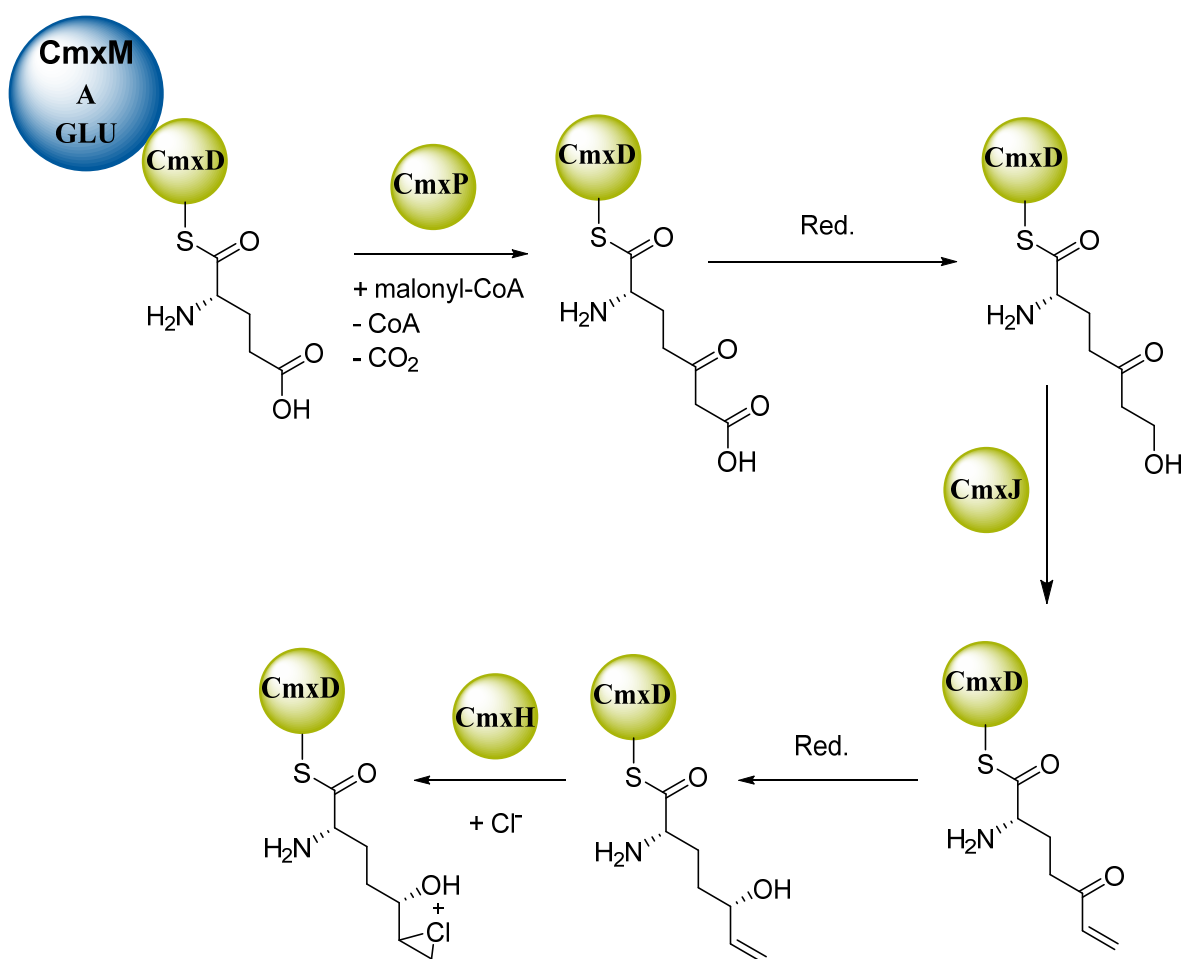
Scheme 3. *Presented reaction scheme for the first part of CMPA biosynthesis*
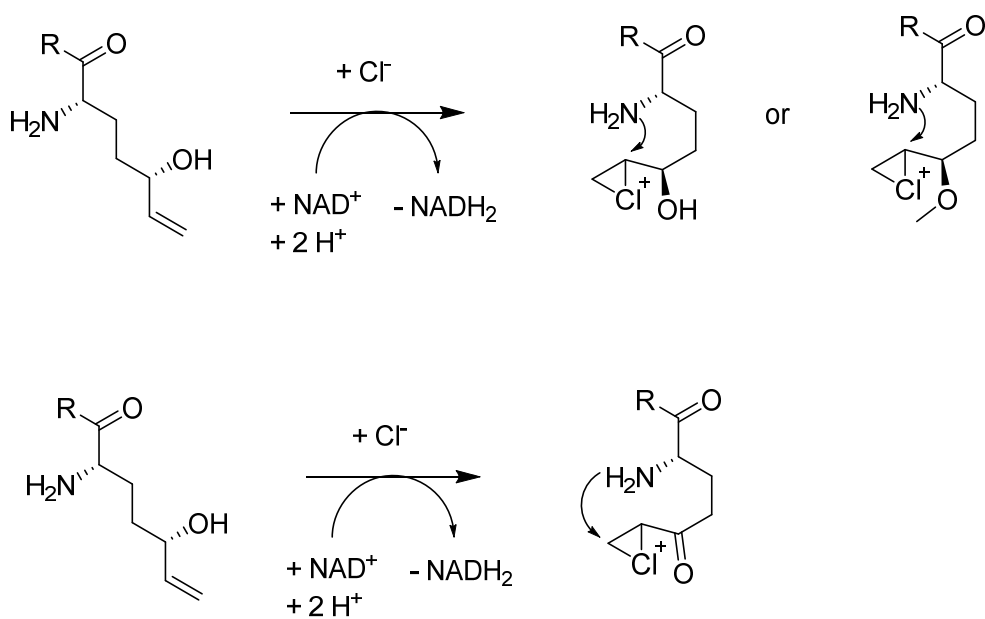
Scheme 4. *Differential chlorination schemes regarding whether chlorination occurs before or after δ-ketoreduction*

It could not yet be clarified whether methyl transfer to complete the CMPA unit occurs as a last step as it is described in the chloromyxamide study (chapter 3), or whether an intermediate as seen on the top right side of Scheme 4 exists. The current working hypothesis consists of late stage methylation as methyl transfer is often the last step in the production of non-canonical building blocks, e.g. in tomaymycin and vioprolide biosynthesis.[49] Insights into CMPA biosynthesis might have additional value as chloromyxamides are unlikely to be the only NRPS systems to contain CMPA as a building block. Other CMPA containing NRPS systems will likely use the same machinery to produce CMPA unit, as reproducing this machinery via heterologous gene transfer is more likely during evolution in contrast to development of an alternative CMPA pathway from scratch.

8.2.4   Conclusions regarding the importance of understanding biosynthesis

During biochemical evaluation of the three biosynthetic pathways for pyxidicycline, fulvuthiacene and chloromyxamide biosynthesis, light was shed on key biosynthetic steps that might become important in completely different biochemical contexts. Biochemical evaluation of natural products in conjunction with their gene clusters will continue to reveal intriguing biotransformation steps to build up complex structural moieties at an accelerating pace. During evolution the underlying enzymology has been shared among producing organisms, meaning that similar structure elements in natural products are often constructed using similar enzymatic steps by homologous enzymes. Thus, in order to better understand secondary metabolite biosynthesis at large, it is essential to build a better understanding of the biosynthetic steps bacteria use to synthesize specific substructures. This has implications not only for the formation of the specific compound at hand but also allows logical deductions for different assembly lines responsible for the productions of yet to-be-discovered molecules featuring the same structural element. This concept highlights once again the importance of in-detail investigation of biochemical processes in secondary metabolite research that extends beyond the traditional "grind and find" approach to natural products discovery.

8.3　References in Discussion and Outlook

[1]　G. M. Cragg, D. J. Newman, *Biochimica et biophysica acta* **2013**, *1830*, 3670.

[2]　H. Becker, J. M. Scher, J.-B. Speakman, J. Zapp, *Fitoterapia* **2005**, *76*, 580.

[3]　S. C. Wenzel, R. Müller, *Mol. Biosyst.* **2009**, *5*, 567.

[4]　C. Schley, Naturwiss.-Technische Fakultät III der Universität des Saarlandes, **2006**.

[5]　G. C. A. Amos, T. Awakawa, R. N. Tuttle, A.-C. Letzel, M. C. Kim, Y. Kudo, W. Fenical, B. S. Moore, P. R. Jensen, *Proceedings of the National Academy of Sciences* **2017**, *4*, 201714381.

[6]　a) N. Zaburannyi, B. Bunk, J. Maier, J. Overmann, R. Müller, *Appl. Environ. Microbiol.* **2016**, *82*, 1945; b) K. Viehrig, F. Surup, K. Harmrolfs, R. Jansen, B. Kunze, R. Müller, *J. Am. Chem. Soc.* **2013**, *135*, 16885; c) K. Viehrig, F. Surup, C. Volz, J. Herrmann, A. Abou Fayad, S. Adam, J. Kohnke, D. Trauner, R. Müller, *Angew. Chem. Int. Ed.* **2017**.

[7]　J. J. Hug, C. D. Bader, M. Remškar, K. Cirnski, R. Müller, *Antibiotics* **2018**, *7*, 44.

[8]　T. Hoffmann, D. Krug, S. Hüttel, R. Müller, *Anal. Chem.* **2014**, *86*, 10780.

[9]　D. Krug, G. Zurek, B. Schneider, R. Garcia, R. Müller, *Anal. Chim. Acta* **2008**, *624*, 97.

[10] T. M. Annesley, *Clin. Chem.* **2003**, *49*, 1041.

[11] D. Remane, M. R. Meyer, D. K. Wissenbach, H. H. Maurer, *Rapid Commun. Mass Spectrom.* **2010**, *24*, 3103.

[12] a) H. Umezawa, *Ann. N. Y. Acad. Sci.* **1958**, *76*, 20; b) M. F. Chellat, R. Riedl, *Angewandte Chemie (International ed. in English)* **2017**, *56*, 13184.

[13] Y.-C. Ma, H.-Y. Kim, *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 1010.

[14] a) Balogh M. P., *Spectroscopy* **2004**; b) X. Feng, M. M. Siegel, *Anal. Bioanal. Chem* **2007**, *389*, 1341.

[15] R. J. M. Weber, A. D. Southam, U. Sommer, M. R. Viant, *Anal. Chem.* **2011**, *83*, 3737.

[16] K. H. Park, M. S. Kim, S. J. Baek, I. H. Bae, S.-W. Seo, J. Kim, Y. K. Shin, Y.-M. Lee, H. S. Kim, *Plant methods* **2013**, *9*, 15.

[17] M. Hoffmann, D. Auerbach, F. Panter, T. Hoffmann, P. C. Dorrestein, R. Müller, *ACS Chem. Biol.* **2018**, *13*, 273.

[18] a) L. Nyadong, E. G. Hohenstein, A. Galhena, A. L. Lane, J. Kubanek, C. D. Sherrill, F. M. Fernandez, *Anal. Bioanal. Chem* **2009**, *394*, 245; b) R. Cramer, *Methods in molecular biology (Clifton, N.J.)* **2009**, *564*, 85.

[19] N. Gibitz Eisath, S. Sturm, H. Stuppner, *Planta Med.* **2017**.

[20] L. Novakova, A. G.-G. Perrenoud, I. Francois, C. West, E. Lesellier, D. Guillarme, *Anal. Chim. Acta* **2014**, *824*, 18.

[21] a) H. Kallio, P. Laakso, R. Huopalahti, R. R. Linko, P. Oksman, *Anal. Chem.* **1989**, *61*, 698; b) M. Wang, E. J. Carrell, A. G. Chittiboyina, B. Avula, Y.-H. Wang, J. Zhao, J. F. Parcher, I. A. Khan, *Anal. Bioanal. Chem* **2016**, *408*, 4649; c) Y. Huang, Y. Feng, G. Tang, M. Li, T. Zhang, M. Fillet, J. Crommen, Z. Jiang, *J. Pharm. Biomed. Anal.* **2017**, *140*, 384.

[22] S. Khater, M.-A. Lozac'h, I. Adam, E. Francotte, C. West, *J. Chromatogr. A* **2016**, *1467*, 463.

[23] a) K. M. Aberg, R. J. O. Torgrip, J. Kolmert, I. Schuppe-Koistinen, J. Lindberg, *J. Chromatogr.* **2008**, *1192*, 139; b) R. Tautenhahn, C. Bottcher, S. Neumann, *Bmc Bioinform.* **2008**, *9*, 504.

[24] T. Hoffmann, D. Krug, N. Bozkurt, S. Duddela, R. Jansen, R. Garcia, K. Gerth, H. Steinmetz, R. Müller, *Nat. Commun.* **2018**, *9*, 803.

[25] a) S. C. Wenzel, R. Müller, *Nat. Prod. Rep.* **2007**, *24*, 1211; b) J. Herrmann, A. A. Fayad, R. Müller, *Nat. Prod. Rep.* **2017**, *34*, 135.

[26] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12580.

[27] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan et al., *Nat. Biotechnol.* **2016**, *34*, 828.

[28] J. Gorges, F. Panter, L. Kjaerulff, T. Hoffmann, U. Kazmaier, R. Müller, *Angew. Chem. Int. Ed. Engl.* **2018**.

[29] Mallard W. G., Sparkman O.D., *U.S. Department of Commerce, National Institute of Standards and Technology Standard Reference Data Program: Gaithersburg, MD*, *2014*.

[30] T. Koenig, B. H. Menze, M. Kirchner, F. Monigatti, K. C. Parker, T. Patterson, J. J. Steen, F. A. Hamprecht, H. Steen, *J. Proteome Res. (Journal of Proteome Research)* **2008**, *7*, 3708.

[31] A. Rhoads, K. F. Au, *Genomics Proteomics Bioinformatics* **2015**, *13*, 278.

[32] A. L. Delcher, K. A. Bratke, E. C. Powers, S. L. Salzberg, *Bioinformatics* **2007**, *23*, 673.

[33] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers, *Nucleic acids research* **2013**, *41*, D36-42.

[34] a) S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389; b) S. McGinnis, T. L. Madden, *Nucleic acids research* **2004**, *32*, W20-5.

[35] K. Blin, M. H. Medema, R. Kottmann, S. Y. Lee, T. Weber, *Nucleic Acids Res.* **2017**, *45*, D555-D559.

[36] T. Weber, K. Blin, S. Duddela, D. Krug, H. U. Kim, R. Bruccoleri, S. Y. Lee, M. A. Fischbach, R. Müller, W. Wohlleben et al., *Nucleic Acids Res.* **2015**, *43*, W237-W243.

[37] a) K. Blin, H. U. Kim, M. H. Medema, T. Weber, *Briefings in bioinformatics* **2017**; b) M. A. Skinnider, N. J. Merwin, C. W. Johnston, N. A. Magarvey, *Nucleic Acids Res* **2017**, *45*, W49-W54.

[38] M. Alanjary, B. Kronmiller, M. Adamek, K. Blin, T. Weber, D. Huson, B. Philmus, N. Ziemert, *Nucleic Acids Res* **2017**, *45*, W42-W48.

[39] F. Panter, D. Krug, S. Baumann, R. Müller, *Chem. Sci.* **2018**, *9*, 4898.

[40] M. H. Medema, R. Kottmann, P. Yilmaz, M. Cummings, J. B. Biggins, K. Blin, I. de Bruijn, Y. H. Chooi, J. Claesen, R. C. Coates et al., *Nat. Chem. Biol.* **2015**, *11*, 625.

[41] S. C. Epstein, L. K. Charkoudian, M. H. Medema, *Standards in Genomic Sciences* **2018**, *13*, 16.

[42] a) C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, S. Neumann, *Journal of cheminformatics* **2016**, *8*, 3; b) S. Böcker, M. C. Letzel, Z. Liptak, A. Pervukhin, *Bioinformatics* **2009**, *25*, 218; c) K. Dührkop, K. Scheubert, S. Böcker, *Metabolites* **2013**, *3*, 506.

[43] a) M. Strieker, A. Tanović, M. A. Marahiel, *Current opinion in structural biology* **2010**, *20*, 234; b) M. H. Medema, Y. Paalvast, D. D. Nguyen, A. Melnik, P. C. Dorrestein, E. Takano, R. Breitling, *PLoS Comput. Biol.* **2014**, *10*, e1003822.

[44] N. Tibrewal, P. Pahari, G. Wang, M. K. Kharel, C. Morris, T. Downey, Y. Hou, T. S. Bugni, J. Rohr, *J. Am. Chem. Soc.* **2012**, *134*, 18181.

[45] a) W. Trowitzsch, G. Reifenstahl, V. Wray, G. Höfle, *J. Antibiot.* **1980**, *33*, 1480; b) F. Sasse, B. Böhlendorf, M. Herrmann, B. Kunze, E. Forche, H. Steinmetz, G. Höfle, H. Reichenbach, M. Hermann, *J. Antibiot.* **1999**, *52*, 721.

[46] S. Weinig, H.-J. Hecht, T. Mahmud, R. Müller, *Chem. Biol.* **2003**, *10*, 939.

[47] A. Sandmann, F. Sasse, R. Müller, *Chem. Biol.* **2004**, *11*, 1071.

[48] S. Fujita, S.-H. Cho, A. Yoshida, F. Hasebe, T. Tomita, T. Kuzuyama, M. Nishiyama, *Biochem. Biophys. Res. Commun.* **2017**, *491*, 409.

[49] a) A. von Tesmar, M. Hoffmann, J. Pippel, A. A. Fayad, S. Dausend-Werner, A. Bauer, W. Blankenfeldt, R. Müller, *Cell chemical biology* **2017**, *247*, 1; b) F. Yan, D. Auerbach, Y. Chai, L. Keller, Q. Tu, S. Hüttel, A. Glemser, H. A. Grab, T. Bach, Y. Zhang et al., *Angew. Chem. Int. Ed. Engl.* **2018**.

9 Appendix

9.1   Curriculum vitae

Personal information

Born on 04.07.1989 in Homburg

1999–2008   Saarpfalzgymnasium in Homburg/Saar, A level *(1,4)*

Education

Nov. 2014 - Today   PhD studies at Helmholtz-Institute for Pharmaceutical Research Saarland,

Department of microbial natural products (MINS)

- Identification and structure elucidation of novel biologically active substances from myxobacteria
- Application of high resolution LC-QqTOF/MS and LC-QqTOF/MS² methods as well as 1D and 2D NMR spectroscopy
- Genetic manipulation of the corresponding bacterial strains to elucidate the corresponding biosynthesis pathways

Apr.-Oct. 2014   Master's thesis at Helmholtz-Institute for Pharmaceutical Research Saarland,

Department of microbial natural products (MINS) *(1.0)*

- LC-MS und LC-MS² profiling of myxobacterial extracts via high resolution LC-QqTOF/MS and LC-QqTOF/MS²
- Statistical data treatment with principal component analysis to find novel metabolite scaffolds
- Purification and structure elucidation of identified interesting metabolites via NMR spectroscopy

2009 - 2014   Integrated bilingual master at Saarland University and Ecole européenne

de chimie polymères et matériaux (ECPM) in Strasburg, France

Diplôme d'Ingénieur *(17.126/20)*

Saarland University, Saarbrücken Master of Science *(1.0)*

- Work on micro projects with industrial partners during the program ECPM innov', Development of methods to remove vinyl chloride from waste water
- Law, economics and toxicology courses
- Trilingual lecture program consisting of German, English and French courses
- Choice of analytical chemistry major 2012-2014

Experience

May - Sept. 2013   Internship at Bosch in Gerlingen / Stuttgart in the elemental analysis team of the central laboratory

- Elemental analyses by ICP-OES of different metals and ceramic materials
- Planning and execution of polymer permeation experiments
- Gas heat extraction experiments to determine Sulfur and Oxygen contents in metals
- Validating a method to determine copper content in coal brushes via ICP-OES

January 2013   Scientific internship at LIMAA institute (Laboratoire d'ingénerie moléculaire appliquée à l'analyse) in Strasburg

'Evaluation of the influence of fluorine ions on the fluorescence of Lanthanide complexes'

- Evaluation of different ligands for lanthanide ions to optimize quantum yield
- Evaluation of the influence of fluorine binding on fluorescence of the complex

July - August 2012   USP Hospital Costa Adeje, San Eugenio, Spain

Internship in the clinical laboratory "General lab" of the Hospital in San Eugenio

- Application of antigen tests and microbiological cultivation methods for diagnostic purposes
- Collaboration in an international team of scientist